Analysis of Server Workload and Client Interactions in a News-on-Demand Streaming System

Frank T. Johnsen, Trude Hafsøe and Carsten Griwodz
Dept. of Informatics, University of Oslo, Norway
{frankjo, truhafso, griff}@ifi.uio.no

Abstract

This paper investigates several aspects of streaming in News-on-Demand services on the Internet by analyzing log files of a news service from Norway's largest online newspaper. We investigate both short and long term effects of client uasge of the service. By comparing our results with earlier work, we found that variations in server load depend strongly on local culture. Furthermore, we found that there are slight variations between client usage of the audio and video material.

1. Introduction

News-on-Demand (NoD) is becoming increasingly popular. In Norway today we see that online newspapers are gaining more and more users and that paper editions are losing ground [1]. As a consequence, an increasing amount Internet traffic is caused by NoD. Currently Internet hosting centers use statistical multiplexing to meet the requirements of competing hosted services. An understanding of NoD long-term and short-term client behavior is needed in order to provide even better services in the future.

The contribution of this paper is an analysis of server load and user behavior in a NoD environment. We analyze access patterns, stream interactions, lifetime- and popularity of streams and traffic peaks. We have obtained logs of streaming media accesses from *Verdens Gang* (VG), Norway's largest online newspaper [1]. The logs span almost two years of streaming, with a total of 4.6 million client requests after having removed commercials and failed requests logs. These requests access 3486 different files, of which about 1000 were audio files. For the remainder of the paper, when we use the term "access" in our analysis discussions, *one access corresponds to one successful request line in the log*.

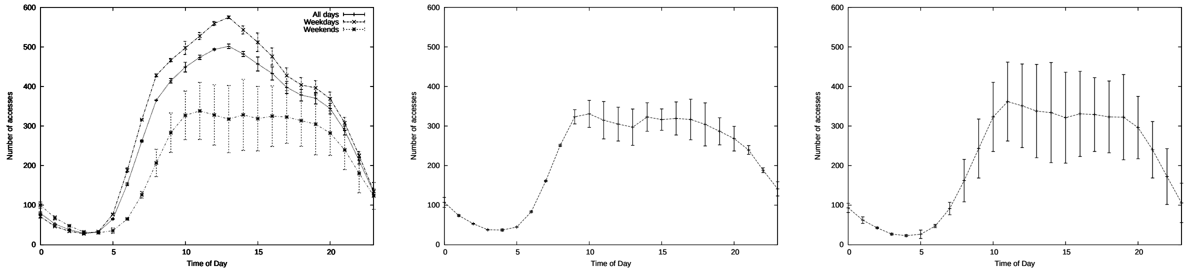
We divide the results of our analysis into long- and short term effects. Long term effects are aggregated results showing properties that are representative for the entire timespan of our log material. Short term effects describe localized anomalies, i.e. traffic peaks.

2. Long term effects

2.1. Access Intensity

The load experienced by a news streaming server follows regular patterns which can be detected by a study of how the number of accesses to the media changes over time. News traffic, unlike for instance peer-to-peer traffic, requires interactivity from the users. This means that access patterns, and thereby server load, depend largely on the daily habits of users, and will vary from one culture to another. User access patterns experienced by a Spanish news service [2] show how users have a long midday lunch break, and that users are most active during the evening. Norwegian social habits are very different from those experienced in Spain; Norwegians work from 8 am to 4 pm, and these working hours are strictly adhered to by the majority of employees. This pattern is clearly visible in Figure 1(a), which shows the average number and standard deviation of accesses for all days, weekdays and weekends, respectively. During workdays the main bulk of accesses occurs during working hours, with the highest point being reached around noon. During weekends the total number of accesses is significantly lower, and the standard deviation is higher. In addition, users start using the news service later in the day during weekends.

Another interesting aspect of Norwegian social habits can be illustrated by a closer study of access intensity on Saturdays and Sundays. Both of these days are non-working days for office workers, but there is a significant difference between the access patterns experienced on these days. Saturdays, as shown in Figure 1(b), follows a fairly regular pattern, but it never reaches the same amounts of accesses as workdays. Figure 1(c) shows that news service usage on Sundays follows the same average access curve as Saturdays, but the variation is much higher. This is likely to



(a) Hour-to-hour access intensity

(b) Access intensity, Saturdays

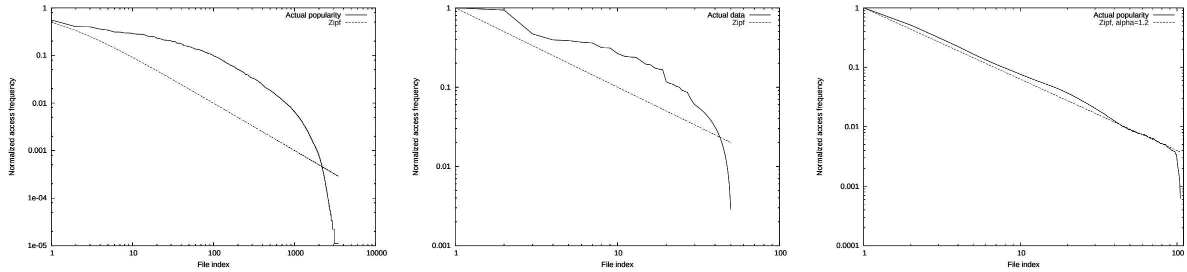
(c) Access intensity, Sundays

Figure 1. Access intensities by time of day

be because most Norwegians have a fairly regular Saturday pattern which includes shopping. On Sundays shops are closed, and outdoor activities are predominant, which means that weather fluctations may be a factor contributing to the high variance.

2.2. Popularity

The Zipf distribution is used to describe the popularity distribution of multimedia content in various applications. It unclear if this assumption holds for NoD. Cherkasova and Gupta [3] note for an enterprise media server workload that workloads do not necessarily follow the Zipf distribution on a longer time scale, while they do on short time scales. Like the authors of many other papers, they observe a common problem with the Zipf distribution. It is hyperbolic function and not additive. When you cut off part of a Zipf function, for example by handling highly popular items through caching and observe only the remaining tail, that tail can not follow a Zipf distribution. In fact, if the original data followed a Zipf distribution, then the tail can not be fitted correctly to a Zipf distribution for another skew factor. Furthermore, if a popularity distribution can be matched by a Zipf distribution for one skew factor, then this approximation may hold for as long as the popularity of all observed items remains unchanged. If the popularity of items develops over time, then the relative popularity of an item differs between its short term relative popularity and its long term relative popularity; if a day's top news item is replaced by another one every day, then the long term relative popularity of these news items can not be Zipf-distributed. This is not considered in [2], where the authors rather try to find a limit value for the skew factor when the time (and the number of archived news items) goes toward infinity. Kim et al. [4] propose a "new popularity model called the Multi-Selection Zipf distribution". The semantic problem of identifying model and distribution nonwithstanding, they use the one-month average popularity of news and try to match this with a Zipf distribution althoughy the relative popularity of news items is known to change rapidly. Predictably, the outcome is that popularity differences at the top are less pronounced than a Zipf distribution would express. The authors observe correctly that popularity spreads over several articles (in the course of a month) but the conclusion drawn is not that the granularity of the observation is too coarse but that articles should be grouped together. Critically, this was done without considering the temporal correlation of those articles. Such problems can be avoided by considering only phases that are so short that the popularities are stable, as in [5]. To exploit the data set from a duration spanning several stable phases, the items for each phase can be sorted by popularity and the values averaged for every index among all of the stable phases. If every single stable phase can be approximated by a Zipf distribution, then it is also possible to approximate the averaged short term period to this distribution function. However, Almeida et al. [5] do not observe a Zipf distribution in spite of considering only stable phases. They observe a different popularity distribution that they describe as a concatenation of two Zipf-like distributions. Obviously, the popularity distribution of their content is extremely heavy-tailed. The data presented by Yu et al. [6] shows the same properties, and so does our data: The entire log period as a phase is shown in Figure 2(a). The picture is totally different when we consider the duration of the stable phase. Figure 2(c) shows the results for an average where the phase is a day. The Zipf distribution, here plotted with a skew factor of 1.2, does not quite fit; there is a heavy tail but it describes the top movies in an excellent manner. Furthermore, picking a random day, i.e. one single stable phase, and investigating the file popularity distribution also shows little adherence to the Zipf distribution, see Figure 2(b). Yu et al. attribute the heavy tail to old videos that remain in the VoD system that they investigated. In the NoD system considered here, we can attribute the tail to the co-existence of up-to-date news clips with an open-ended archive for our data.



(a) Accesses of all files

(b) Accesses of all files on one day

(c) Accesses of all files; day-to-day average

Figure 2. Normalized accesses and the Zipf distribution

Popularity development over time

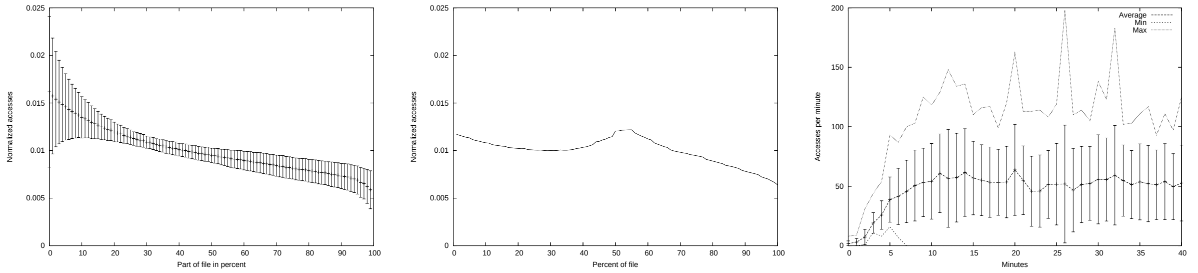
In the following analysis popular files are defined as *the collection of files which topped the daily hit statistics at least once*. The majority of popular files reaches top ranking on the same day as they are published. This is a consequence of the nature of news, where the breaking news is presented first on news websites and is of high interest to most readers. The few files that reach top ranking late in their lifespan are mostly audio files containing music. Once a file has reached the highest ranking it is likely to drop in popularity shortly after. Most files drop to below tenth place in just a few days. This indicates that more recent news has captured the attention of readers, and older news is less interesting to view. An analysis of the files' lifespan supports this. The lifespan of a file is a count of all the days that a file has been accessed at least once. For most files this is one continuous period, but some files have a few days when they are not accessed at all. Such days are not considered a part of the files' lifespan, and are thereby not counted. 50% of the popular files have a lifespan of 20 days or less, whereas a few files have relatively long lifespans; the longest being 200 days. Only one file was accessed for that long.

2.3. Stream Interaction

Out of the 4.6 million total accesses, 66.9% (over 3 million) are partial. Of the partial accesses, around 700000 accesses have an offset greater than zero, meaning that the user jumped and started playback somewhere else than at the beginning of the file. This constitutes 15.3% of all the accesses, indicating that 51.6%, just over half of all accesses which start at the beginning of the media, are aborted before playing to completion. Figure 3(a) shows the normalized access distribution averaged for all files, including the standard deviation. Accesses for all files are concentrated at the start of the file, and client interest is uniformly declining throughout the duration of the file. This illustrates the fact that most users will start playback at the beginning of the file, and view it either to completion or until they lose interest and stop. For video files that start playback at the beginning of the file, 34.8% of the users stop during the first third of the stream, while most other users do not stop until the last third. Few users, less than 3%, stop viewing video files in the middle of the file. For audio files this is somewhat skewed, as only 6.2% stop playback in the first two thirds combinded, while the rest listen to the file until the last third. So far we have seen that on average NoD exhibits low interactivity and accesses are weighted towards the beginning of the files. There are, however, exceptions to this when looking at single files. One example, shown in Figure 3(b) illustrates the access distribution of the single most popular file in the system. In this file the middle part is more popular than the beginning. It is easy to explain this anomaly; the file is a video of a car accident, and the most popular part shows the cars crashing. Users who have seen the file return to watch this part again, sometimes more than once.

3. Short term effects

Some days have an access intensity far surpassing the average. We have investigated the cause of such peak days, and looked into how the files causing such peaks increase in popularity. Typically, we have limited our investigations to files with over 10000 accesses on a given day, and defined the start of a peak to be *the first time the amount of accesses per minute is greater than 10*. Some files have a low but stable access pattern, so we do not consider files with a max access per minute less than 20 overall. Files vary a lot when it comes to the start of the peak. From the first access to a file, the shortest time until a peak started was 2 minutes, whereas the longest was 31 hours. The average time from the first access to a file till the peak started was 2.6 hours. In order to obtain knowledge of how peaks form, we had to compare all the peaks from the different



(a) All accesses to parts of files

(b) Distribution of all accesses for one example file

(c) Traffic peak formation

Figure 3. File access patterns

files. Since different files start peaking at different times, we had to identify the peak starting point and employ a cutoff 2 minutes before the start of the peak activity to be able to compare the different files. We calculated the average accesses per minute during each peak, and also the standard deviation. Figure 3(c) illustrates this, along with the minimum and maximum number of accesses per minute. We see that a peak builds fairly rapidly with a massive increase in accesses during the first 5-10 minutes, after which it remains fairly stable for an extended period of time. These peaks come without warning, put high loads on the server and demand a lot of bandwidth from the network.

4. Conclusions and future work

In this paper we have investigated usage patterns for streamed NoD, and focused on access intensity, file popularity, stream interactions and traffic peaks. Access intensity follows a daily periodic pattern based on the time of day, but these patterns are strongly dependent on local culture. In Norway most accesses occur during working hours. Weekend access patterns differ from the weekday pattern by having fewer hits and a larger variation in access counts. The Zipf distribution can not be used to accurately describe the NoD system we investigated here because the co-existence of new files and old files leads to a heavy-tailed distribution of the content. However, it can accurately describe the popularity of videos in the most recent articles. Our study of stream interaction patters revealed that almost most accesses are partial, meaning that the clients do not view the entire file. On average accesses are biased towards the start of the file, with uniformly decreasing interest through the duration of the file. We have found that access peaks often start shortly after a file is first published, and that once they start forming, they grow from low interest to a large peak within 10 minutes on average.

We plan to perform further analysis of server load and NoD traffic patterns in order to develop a model of user interactivity. Such a model can be a valuable tool when evaluating new server solutions, and may aid in providing more efficient content distribution services in the future.

Acknowledgments

We would like to acknowledge Espen Nilsen who put in a lot of effort on our behalf by contacting all the major news providers in Norway. Furthermore we would like to thank Anders Berg at VG who decided to share log material with us. We could not have done this work without their support.

References

[1] Computerworld. Nettaviser i toppen (in Norwegian). `http://www.computerworld.no/index.cfm/bunn/artikkel/id/50177`. Accessed 2006-06-05.

[2] M. Vilas, X.G. Pañeda, R. García, D. Melendi, and V.G. García. User behaviour analysis of a video-on-demand service with a wide variety of subjects and lengths. In *The 31st EUROMICRO Conference on Software Engineering and Advanced Applications*. IEEE Computer Society, 2005.

[3] Ludmila Cherkasova and Minaxi Gupta. Characterizing locality, evolution, and life span of accesses in enterprise media server workloads. In *NOSSDAV '02*, pages 33–42, New York, NY, USA, 2002. ACM Press.

[4] Y.-J. Kim, T. U. Choi, K. O. Jung, Y. K. Kang, S. H. Park, and Ki-Dong Chung. Clustered multi-media NOD: Popularity-based article prefetching and placement. In *IEEE Symposium on Mass Storage Systems*, pages 194–202, 1999.

[5] Jussara M. Almeida, Jeffrey Krueger, Derek L. Eager, and Mary K. Vernon. Analysis of educational media server workloads. In *NOSSDAV '01*, pages 21–30, New York, NY, USA, 2001. ACM Press.

[6] Hongliang Yu, Dongdong Zheng, Ben Y. Zhao, and Weimin Zheng. Understanding user behavior in large scale video-on-demand systems. In *Proceedings of EuroSys*, Leuven, Belgium, April 2006.