Analysis and characterization of a video-on-demand service workload

Ali-Eldin, Ahmed; Kihl, Maria; Tordsson, Johan; Elmroth, Erik

Published in:
[Host publication title missing]

DOI:
10.1145/2713168.2713183

2015

Analysis and Characterization of a Video-on-Demand Service Workload

Ahmed Ali-Eldin
Dept. of Computing Science
Umea University, Sweden
ahmeda@cs.umu.se

Maria Kihl
Dept. of Electrical and Information Technology,
Lund University
Maria.Kihl@eit.lth.se

Johan Tordsson
Dept. of Computing Science
Umea University, Sweden
tordsson@cs.umu.se

Erik Elmroth
Dept. of Computing Science
Umea University, Sweden
elmroth@cs.umu.se

ABSTRACT

Video-on-Demand (VoD) and video sharing services account for a large percentage of the total downstream Internet traffic. In order to provide a better understanding of the load on these services, we analyze and model a workload trace from a VoD service provided by a major Swedish TV broadcaster. The trace contains over half a million requests generated by more than 20000 unique users. Among other things, we study the request arrival rate, the inter-arrival time, the spikes in the workload and their cause, the video popularity distribution, the streaming bit-rate distribution and the video duration distribution. Contrary to some previously analyzed workloads in the literature, our results show that the user and the session arrival rates for the TV4 workload does not follow a Poisson process. The arrival rate distribution is modeled using a lognormal distribution while the inter-arrival time distribution is modeled using a stretched exponential distribution. We observe the "impatient user" behavior where users abandon streaming sessions after minutes or even seconds of starting them. Both very popular videos and non-popular videos are specially affected by impatient users. We also show that this behavior is an invariant in VoD workloads and is neither affected by the average bit-rate nor by the number of videos a user watch.

1. INTRODUCTION

Over the past decade, Video on Demand (VoD) and Video sharing online services have been on the rise. A recent report estimated that more than 50% of the total downstream traffic during peak periods in North America originate from Netflix and YouTube [15]. It is thus required to analyze and characterize VoD workloads in order to understand how to improve and optimize the network usage and the perceived Quality-of-Service (QoS) by the service users.

Many VoD service providers utilize the power of cloud computing to host their services [3, 19, 32]. Since a typical cloud hosts multitudes of applications with differing workload profiles [1], Cloud service providers need to understand the workload characteristics of the running applications including the VoD workload dynamics. This understanding is crucial as application co-hosting can result in performance interference between collocated workloads [9, 34]. Furthermore, resource management problems such as service admission control [42], Virtual Machine (VM) placement [10], VM migration [41] and elasticity [35] can be further complicated depending on the workload characteristics [29]. To better understand VoD workloads, we obtained recent workload traces from TV4, a major Swedish VoD service provider, detailing the requests issued by the premium service subscribers to TV4's VoD service. The VoD service is hosted on a number of cloud platforms. We provide an extensive analysis and characterization study of the traces.

Table 1: Example of one entry in the trace.

Video title	Farmen del 1_2255111
viewer ID (hashed)	a257d2e7788db3238f
Streaming start time	2013-01-13 17:00:00
Streaming end time	2013-01-13 17:08:00
Number of minutes viewed	8
average Bitrate (Mbps)	0.8
categories	None
category Tree	Nöje/Farmen
video Category	Farmen

Table 1 shows an example entry in the trace. Each entry contains nine fields, out of which the following seven are used in the analysis, the title of the video requests, the hashed ID of the premium user who requested the video, the time the streaming of the video started and ended, the number of minutes the video was streamed, and the average bitrate of the stream. The remaining two fields are not used in the analysis since they are missing for some videos.

The traces contain logged data between the 31^{st} of De-

cember 2012 until the 18^{th} of March, 2013 from two cities. [1] City A is one of the largest cities in the Nordic countries with a population over half a million inhabitants while City B is a much smaller city with less than fifty thousand inhabitants. The number of premium users who used the service during that period is 23102 users. The users viewed 17131 unique videos and started 532421 streaming sessions.

There are two main contributions of this work. First, in Section 2, we provide an extensive study of the workload properties related to the session arrivals. We show that the session arrivals and the user arrivals can not be modeled using a Poisson process, contrary to what has been assumed [25] or reported [45] in some prior work in the literature. We identify the main events that resulted in spikes in the number of arrivals. Since the spikes cause non-stationarity in the workload, the Hilbert-Huang transform is used to perform spectral analysis on the workload.

Second, Section 3 contains analysis for the properties of the workload related to the streamed sessions. Our analysis shows that a large percentage of the sessions started are terminated within the first few minutes before the video ends. Based on this behavior and the popularity distribution of the videos, we suggest a new caching strategy to help reduce the wasted resources by the VoD service provider. We conclude in Section 5.

2. WORKLOAD ANALYSIS: ARRIVALS

2.1 Arrival rate

The Probability Distribution Function (PDF) and the Cumulative distribution Function (CDF) of the hourly session arrival rate is shown in Figure 1 (in blue) on a Log-Log plot. An almost identical plot was also obtained for the user arrival rate since one user almost always does not start more than one session per hour. The PDF suggests that the arrival rate process can be modeled using a heavy tailed distribution. we have fitted the arrival rate data to different distributions and compared the goodness of fits in order to find a good fit. The data was fitted to lognormal, exponential, truncated power law, stretched exponential, gamma and power law distributions, three of which are shown in Figure 1.

The plots show that either a lognormal distribution, an exponential distribution or a stretched exponential distribution is a good fit. To choose the best fit, we used the Kolmogorov-Smirnov (KS) test [13]. The p-value for both the lognormal distribution and the stretched-exponential distribution [31] was greater than 0.05, the least significance level required to validate the null hypothesis that the empirical data does not follow the distribution. To be precise, the KS distance for the lognormal distribution is 0.077 with a p-value of 0.09, and the KS distance for the stretched exponential distribution is 0.059 with a p-value of 0.31.

[1]The non-disclosure agreement prohibits us from revealing the city names

Both the lognormal and the stretched exponential distributions are possible fits given the p-values of the KS test. To identify the better fit, we use the log-likelihood ratio between the distributions [13]. The log-likelihood ratio of the lognormal distribution was higher with a p-value of 0.01. We thus conclude that the lognormal distribution is the best distribution to fit our data from the distributions tested. The fitted lognormal distribution is different from the arrival rate distribution of the VoD service provided by China Telecom discussed by Yu et al. [45] where the arrival rates follows a modified Poisson distribution.

2.2 Inter-Arrival time

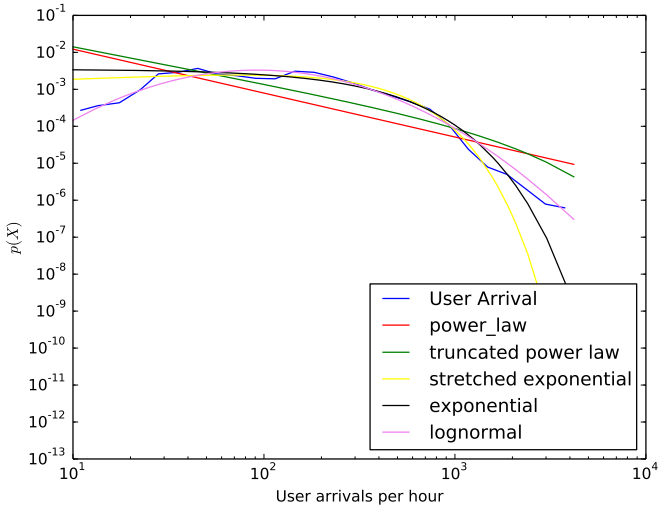
Figure 2 shows the PDF of the video sessions' inter-arrival times (seconds) on a log-log scale. More than 50% of the sessions start after one or less than one second from the arrival of the previous session and around 90% of the sessions start within a minute from a previous session. The maximum inter-arrival time is around 24 minutes. We have again tried fitting a distribution to the Inter-Arrival time following the same steps described above. Again, the KS test showed that both the lognormal distribution (p-value=$0.08 > 0.05$) and the stretched exponential distribution (p-value=$0.08 > 0.05$) to be two viable fits. Testing using the log-likelihood method described by Clauset et al. [13], the stretched-exponential distribution has a higher likelihood than the lognormal distribution with a p-value=0.

While it is popular to model session and user arrival rates as Poisson processes in workload generators [25], our results suggest that for different VoD services, different models of arrival might occur. Poisson processes require the inter-arrival time distribution to be exponential. Figure 2 shows also the best exponential distribution fit we could achieve for the inter-arrival data. the deviation clearly shows that the inter-arrival time distribution is not exponential, and thus the arrivals do not follow a Poisson process. Poisson processes were considered the defacto processes to model network arrivals until the seminal work by Paxson and Floyd [36]. It is thus worth investigating if Poisson processes fail also to model arrival processes for VoD systems. We unfortunately do not have sufficient data from enough VoD providers to come to such a conclusion.

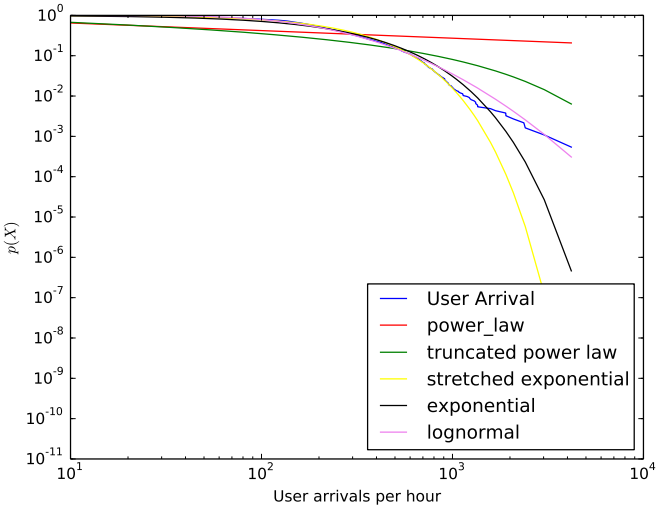
Since at least for the TV4 workload, the user and request arrival processes can not be modeled using Poisson processes, many of the previously developed theories and models based on the assumption of requests/users generated from a Poisson process will either be inaccurate or will be wrong for systems like TV4 [17, 20, 25]. Since VoD workloads are scarce, we can not compare our results with systems other than the very few available in the literature.

2.3 Workload Spikes

Figure 3 shows the number of video sessions started per hour in the trace. The workload has a very clear daily pattern and a weaker weekly pattern. The pattern is violated due to



(a) PDF of the distribution of the arrival rate and different fits.



(b) CCDF of the distribution of the arrival rate and different fits.

Figure 1: The hourly request arrival rate distribution can be modeled as a lognormal distribution seen in the Log-Log plots of the PDFs and CDFs.

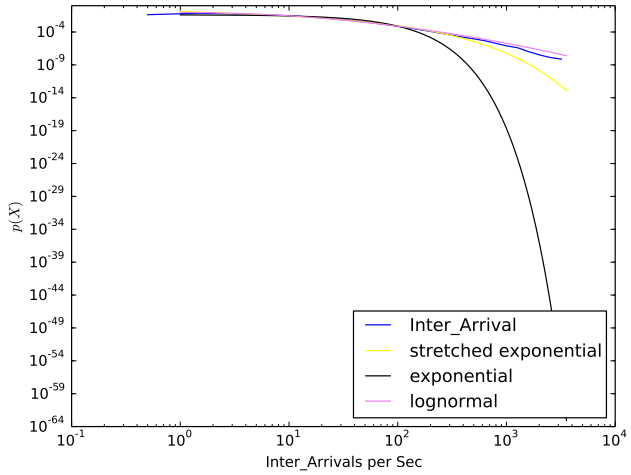


Figure 2: CDF of inter-arrival time (Log-Log plot).

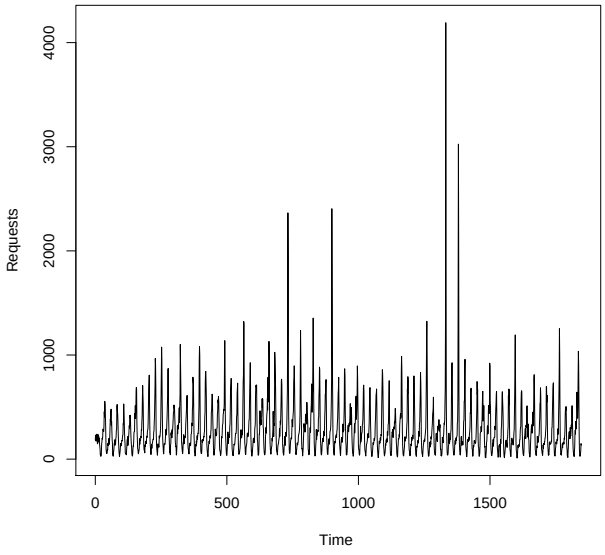


Figure 3: Number of video sessions per hour starting from 09:00 on the 31^{st} of December, 2012, to 09:00 on the 18^{th} of March, 2013. Some sessions run for a few seconds while others run for hours.

four main significant spikes. The most significant spike occurred on Sunday, the 24^{th} of February, 2013 when the load increased from 687 video sessions at 18:00, to 4670 video sessions at 20:00. We investigated the main cause of this spike. To our surprise, the most viewed video was a live stream of a football game between two French teams in the French soccer league, Paris Saint-Germain and Olympique Marseille. Paris Saint-Germain is the team where, Zlatan Ibrahimovic, one of Sweden's favorite soccer players play [40]. The spike caused a workload increase by roughly four to six folds from normal behavior seen in the previous weeks.

The second most significant spike occurred two days later on the 26^{th} of February, when the load increased from 582 sessions at 19:00 to 3025 session at 21:00. Again, the main cause of the spike was a semi-final match in the Spanish cup

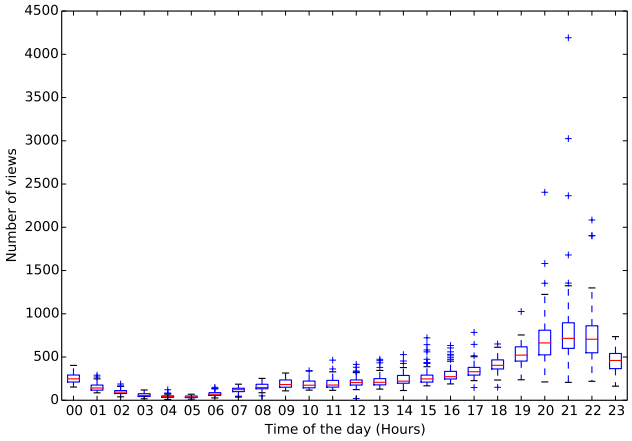


Figure 4: A Box and Whiskers graph showing the effect of the time of the day on the number of sessions.

between Barcelona and Real-Madrid. The third largest peak occurred on the 6^{th} of February, 2013 at 20:00 when an international football match was played between the Swedish and the Argentinian national teams. The load increased to 2405 sessions in that hour. The fourth largest peak occurred on the 30^{th} of February, 2013 at 21:00 when the load reached 2365 sessions. Again, the main cause of the spike was a football match in the Spanish cup between Barcelona and Real Madrid, the first leg prior to the match that caused the second largest spike.

Three of the four major spikes in the workload are generated by events that are not directly related to Sweden. This is a clear example showing the complexity of workload spike and burstiness management. A service provider should be able to cope with such events with no reduction in the QoS perceived by the service customers. The easy but expensive way to provide high QoS guarantees in the presence of spikes, is to over-provision, buy or rent enough server resources to handle the largest future spike well in advance before such a spike occurs. Another solution would be to utilize the power of cloud computing where new resources can be provisioned whenever needed and released when not used anymore [4]. The problem with the second approach is the difficulty of detecting spikes as they occur to be able to provision resources with no QoS degradation.

2.4 Daily patterns

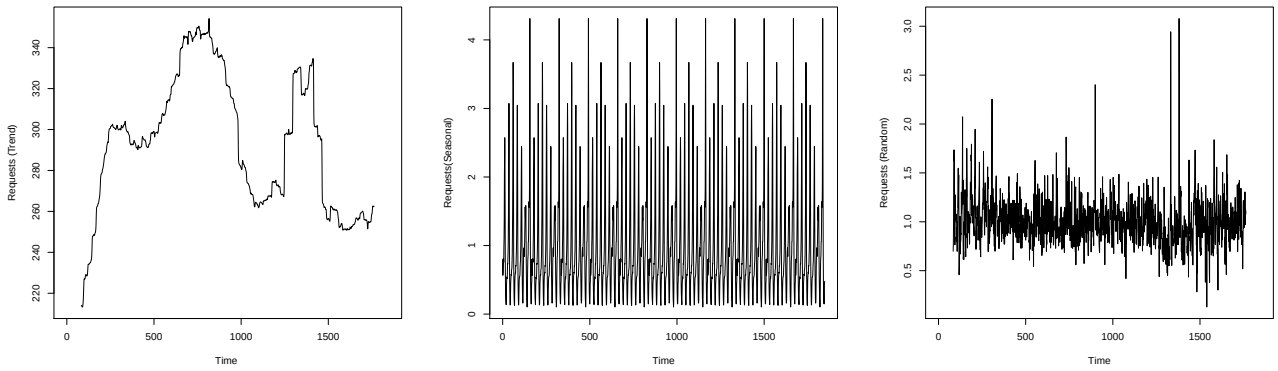
Figure 4 shows a box and whiskers plot showing the effect of the time of the day on the number of video sessions. A box-plot is a way to visualize the quartiles and the dispersion of the distributions of data [33]. A box is plotted for all 24 hours of the day where the lower edge of the box represents the first quartile of the number of sessions arriving on any particular hour. The third quartile is represented by the top edge of the box. From figure 4, the diurnal pattern of the trace is evident. the hour with the least arrivals is at 05:00 every morning with almost no variability. The hours with the highest arrival are at 20:00, 21:00 and 22:00 at night. Since there is very little variability in the load between mid-night and 15:00, and, the arrivals at these hours are low, a VoD service provider can use the available unused resources for business analytics [30] or can release them to save costs. Before hours with higher variability, the provider can provision more resources.

Similar patterns can be seen in other workloads, e.g., the load on Wikipedia [5], where the load decreases significantly between mid-night and noon. A cloud service provider can thus benefit from having services from different time-zones running in the datacenter. The multiplexing between the different services from different time-zones should provide higher revenues with a much lower risk of service performance interference.

2.5 Frequency representation

The request arrival rate represents a time-series. Any time series, X, can be decomposed into three components, the trend, the seasonality and the random components [27]. The trend, T is a slowly changing component which captures the change in the mean of the time series with time. The seasonality, S, represents the periodic components in the load. The random component, r is the remaining signal. The decomposition can be performed such that $X = T + S + r$ or such that $X = T \times S \times r$ [27]. Time-series can be classified into either stationary or non-stationary time-series. A stationary time-series has a non-changing mean and variance. A non-stationary time-series has one or both of mean and variance changing. To use traditional time-series analysis models such as ARIMA models, the studied time-series needs to be stationary [14]. The request arrivals time series for the TV4 data is non-stationary due to the presence of large spikes.

In their seminal work to model non-stationary time-series, Huang et al. introduced a novel empirical method to characterize the frequency variations in non linear and non-stationary time-series [22], recently, known as the Hilbert-Huang Transform (HHT) [21]. At the the core of the HHT is the *Empirical Mode decomposition (EMD)* method and its different variations [22, 43, 16]. The EMD (and all its other variations) are methods with which any complicated data set can be decomposed into a finite and often small number of *Intrinsic Mode Functions (IMF)* that admit well-behaved Hilbert transforms. The Hilbert spectrum can then be used to visualize the produced IMFs and frequency variations in the original signal. Since the number of IMFs produced is low, it is a more efficient way of spectral analysis compared to, for example, the Fourier transform which typically requires an infinite number of sinusoidal frequencies to represent any time-series. Huang et al. and others have discussed the strengths and weaknesses of their proposed method and showed the superiority of the HHT compared to other available spectral analysis methods such as the wavelet transforms and Fourier transforms [16, 22, 23, 39, 43]



(a) The trend component T. (b) The seasonality component S. (c) The random component r.

Figure 5: The result from using multiplicative decomposition of the workload.

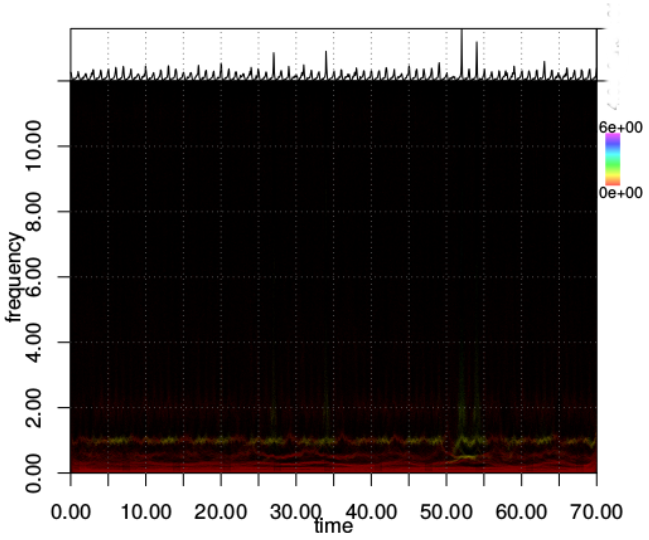


Figure 6: The Hilbert spectrum for the session arrivals time-series.

Figure 5 illustrates the decomposition of the arrivals time-series, in Figure 3, into multiplicative factors. Since the trend component has is a DC component with no frequency variations, we have used the HHT method to perform spectral analysis on $X = S \times r$. The Hilbert spectrum is shown in Figure 6. The X-axis is the time in days and the Y-axis is the frequency in weeks. The colors represent the intensity of the frequency component at any point in time. On top of the graph, the analyzed time-series is plotted. The low frequency components dominate the time-series. The strongest of these components is the weekly component.

At the times of the four major spikes discussed previously, between 25 and 35 days, and 50 and 55 days in Figure 6, the spectral pattern is distorted. The spikes cause an increase in the power and dispersion of the spectrum of the time-series. This suggests that a possible way to detect spikes as they occur would be to use spectral analysis methods to detect the beginning of the spike [6]. We leave this for future work.

3. WORKLOAD ANALYSIS: VIDEO SESSIONS

3.1 Video popularity

Videos offered by a VoD service provider differ in popularity between the users. Figure 7 shows the PDF of the popularity distribution of all videos viewed in our trace. We have fitted the popularity data to different distributions in a way similar to the way the arrival rate was fitted. Using the KS test, all the tried fits had a very low p-value and therefore were bad. Popularity in often modeled with a power law or a Zipfian distribution [25, 45, 2]. Figure 7 suggests a heavy-tailed distribution would be good. We suspect that our fits are bad because of the large gap in the tail of the empirical distribution.

3.2 Video duration

Videos provided by the VoD service are of variable length.

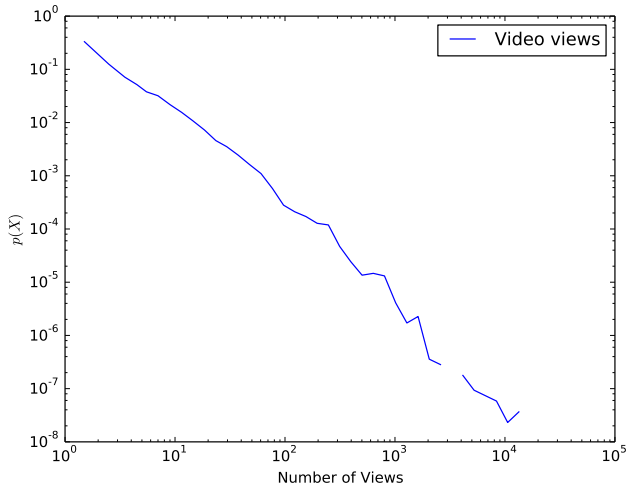


Figure 7: The distribution of the popularity of videos follow a truncated power law distribution with $\alpha = 1.6$ (Log-Log plots).

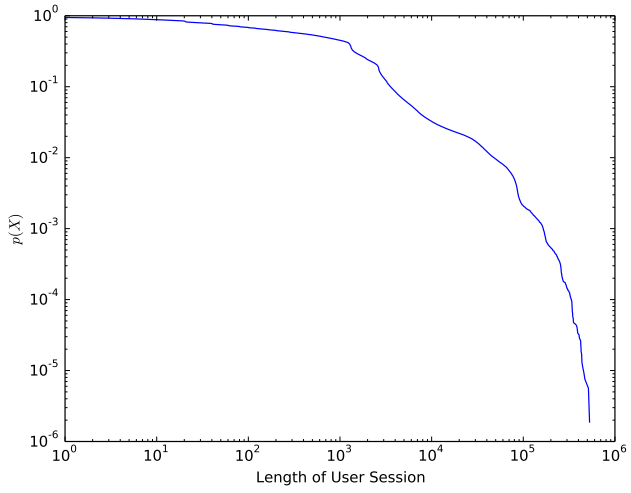


Figure 8: CCDF of the length of the sessions (Log-Log plot).

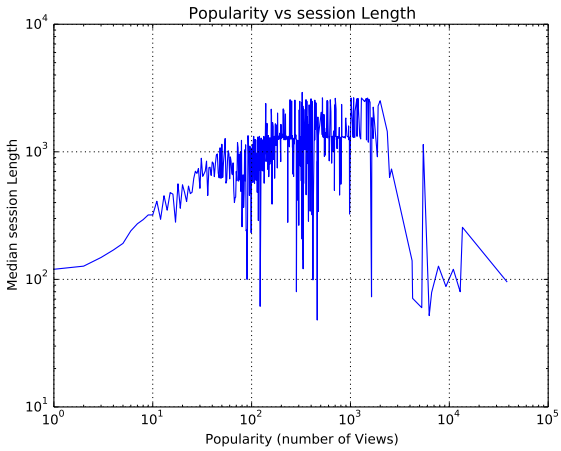


Figure 9: Video popularity versus median session length (Log-Log plot).

Not all users who start viewing a video stream continue to watch until the end. Some users keep replaying the video, going back and forth in the video and pausing the video. This leads to sessions having very different length. Figure 8 shows the complementary CDF (CCDF) for the length of the VoD sessions. More than 90% of the sessions last for less than one hour, with more than 50% of the total sessions lasting less than 12 minutes. More than 20% of the sessions gets terminated within the first 30 seconds from their start time.

These numbers confirm the "impatient user behavior" discussed in previous studies described by Yu et al. [45]. Although the difference between our study and Yu et al.'s study is around 6 years, the numbers we find here do not differ considerably from their study. For example, Yu et al. found that 50% of the users terminate a session within the first ten minutes from when they start it and that more than 90% of all sessions terminate within 60 minutes from when they start. The main difference between our study and Yu et al.'s study in this respect is that the users of the TV4 VoD are more likely to stay than the users in Yu et al.'s study if they make it past the first 10 minutes.

It is interesting to note that a session lasted for 528771 seconds, which is equivalent to over 6 days. During the session, only one video has been streamed. The length of the video on the VoD service servers is around one hour. The average bit-rate for this stream was zero Mbps. We suspect that this is a session started by a user, paused and then forgotten about for six days with the receiving device on during that whole period. Another possibility is a fault in the receiver device resulting in no proper termination for the session.

3.3 Caching

To handle the "impatient crowd", Yu et al. suggest to cache the first 10 minutes of videos to handle the load from up to 50% of the viewers. Since the popularity of the avail-

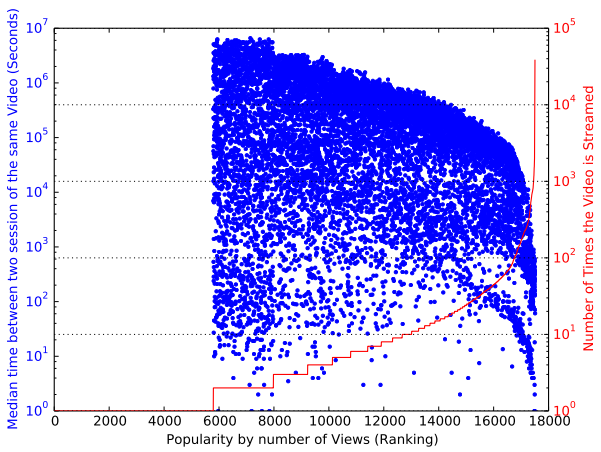


Figure 10: Video popularity versus the median time between two streams of the video.

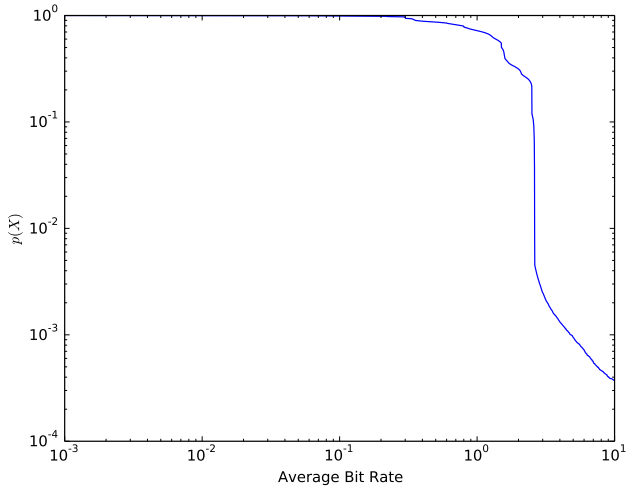


Figure 11: CCDF of the streaming bit-rate in Mbps (Log-Log plot).

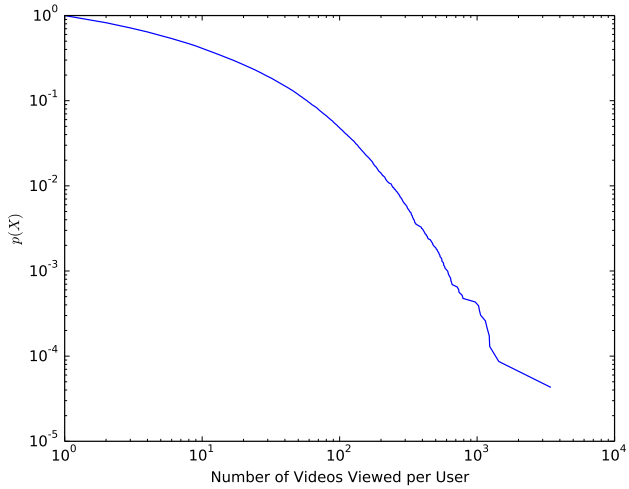


Figure 12: CCDF of the distribution of the number of videos viewed per user (Log-Log plot).

able videos are not the same, it would be a waste of resources to cache the first 10 minutes of unpopular videos or videos which get abandoned by most users before 10 minutes. Figure 9 shows how the median session length changes with the popularity of the videos (number of views). For extremely unpopular and extremely popular videos, the "impatient user behavior" is quite high with the median session length of around 100 seconds. Videos with a medium popularity seem to have longer session times. An advanced caching and prefetching policy [28] should utilize this difference to be able to improve the QoS while reducing wasted resources, e.g., by caching the first 16 to 20 minutes for videos with medium popularity, streaming the first 3 minutes for videos with low popularity and caching and prefetching the first 10 minutes for videos with high popularity.

In order to further understand better how to improve the caching strategy, we plot Figure 10. The X-axis of the figure represents the rank of the videos based on the number of times it has been streamed starting from least popular (with rank 1) to the most popular (with rank 17506). We then plot the median time between the arrivals of two consecutive streaming sessions for the same video (the blue dots). We also plot the total number of times the video has been streamed (the red line). The least popular 5791 videos were streamed only one time during the whole period. It is therefore useless to cache these videos since they are seldom streamed. Looking at Figure 9, the median time for a session for these videos is less than 200 seconds. As the video popularity increases, the median time between two streaming sessions decreases considerably. Video popularity is volatile. Many of the higher ranked videos are streamed many times for a short period and never streamed again.

3.4 Bit-rates

Figure 11 shows the CCDF of the distribution of the average bit-rate transfer speed in Mbps. More than 90% of the VoD service users were streaming at a bit rate equal to or greater than one Mbps. More than 99% of the service users were streaming at a bit rate less than 3 Mbps. The bit-rate is highly affected by the length of the session. Extremely short sessions have typically lower average bit-rates since the session ends before the network session connection is stabilized. Extremely low average bit-rate measurements should be correlated to the session length in order not to draw the wrong conclusions about the network performance.

3.5 Video views per user

Figure 12 shows the CCDF of distribution of the number of videos viewed per user. More than 90% of the service users view less than 70 videos during the period of the study, i.e., less than one video per day. Many of these sessions last for less than 10 minutes. To see how long a user uses the VoD service, Figure 13 shows the CCDF of the total time a

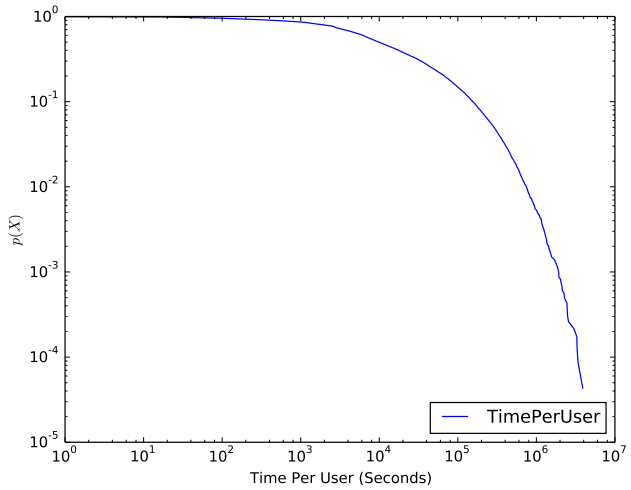


Figure 13: CCDF of the distribution of the number of seconds viewed per user (Log-Log plot).

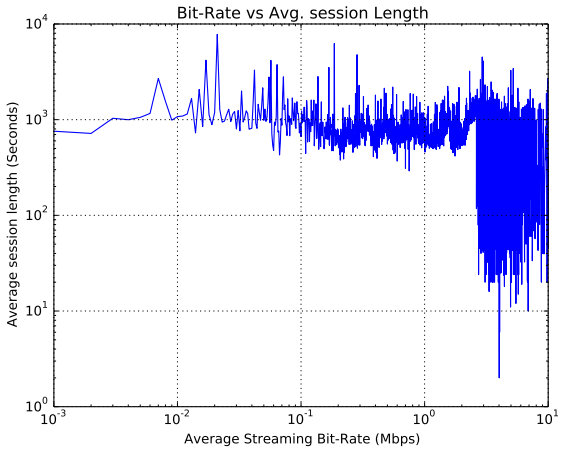


Figure 14: The bit-rate does not have any effect on the decision of a user to abandon a session early (Log-Log plot).

user used the VoD service. Some users have used the service for just a few minutes, with more than 25% of the users using the service for 45 minutes or less. Other users have used the service heavily. The longest usage was by a customer who used the service for a total of 45 days and a few hours. This can be either a user who has the service running for over 15 hours per day, like a restaurant using the service, or a customer who has multiple devices all connected to the service using the same ID.

3.6 Impatient users

To better understand why users abandon streams early, we investigated two hypotheses. The first hypothesis was that users abandon sessions due to low quality of the streaming, i.e., low bit-rate. Figure 14 shows how the average median session length of all users changes with the average streaming bit-rate. The figure shows that across most of the seen average bit-rates, the behavior of impatient users does not change. From Figure 11, average bit-rates more than 3 Mbps are rare, and thus the variation seen when the bit-rates are more than 3 Mbps in Figure 14 should not be interpreted as a change in the user-behavior but rather as outliers.

The second hypothesis was that users who use the service more will have a different average median session length. Figure 15 shows that the session length does not differ between users who use the service very often from those who do not. Thus, both hypotheses are not true. The session length distribution is an invariant in the system.

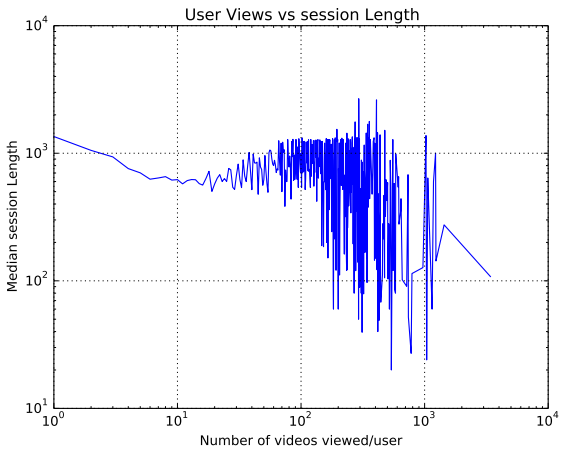


Figure 15: Users have an almost homogeneous median session lengths no matter how often do they use the service (Log-Log plot).

4. RELATED WORK

Several server workloads for different services have been analyzed in depth previously [7, 8, 24, 38, 44]. Many of these studies focused on online video services and video streaming. One of the first and largest studies was conducted by Yu et al. [45] on a VoD system deployed by China Tele-

com, covering a total of 1.5 million unique users for a period of seven months in 2004. They focus their analysis on logs from a single representative city with a total user base of 150,000 users. They study the user arrival rates, session lengths and video popularity dynamics, and how they can affect the caching strategy used.

Choi et al. analyzed service logs generated for every VoD request or VoIP call made on a day in April,2009 in a nationwide commercial IP network in Korea [12]. The number of subscribers for both services is over 1.2 million generating over 10.5 million requests in total during that day. The authors focus on workload characteristics having a direct effect on the performance of IP networks such as session arrivals and session holding times.

Multiple studies have crawled and analyzed traces from YouTube, Yahoo! videos and DailyMotion. Video streaming from an ISP perspective has been studied by Pilssonneau and Biersack by analyzing 10 packet traces from a residential ISP network [37]. They focus on video streams from YouTube and DailyMotion with a focus on analyzing the flow performance of the videos and the user behavior of the two services. The authors study the influence of the reception quality on the users ans show that videos with bad reception quality are seldom fully downloaded.

Khemmarat et al. [28] collect user browsing pattern data for YouTube. They show that video buffering affects the QoS for YouTube users due to disruptions for buffering. The authors propose a video prefetching approach for user-generated video sharing sites like YouTube based on the site's recommended videos list for any video. They show that prefetching considerably improves the QoS of YouTube.

Kang et. al. crawled Yahoo! videos website for 46 days [26]. Around 76% of the videos crawled are shorter than 5 minutes and almost 92% are shorter than 10 minutes. They discuss the predictability of the arrival rate with different time granularities. A load spike typically lasted for no more than one hour and the load spikes are dispersed widely in time making them hard to predict. Gill et al. [18] collected data on all YouTube usage at the University of Calgary network for 85 consecutive days, starting January 14, 2007. In addition, they monitored the 100 most popular videos on YouTube for the same period. They examined the usefulness of caching and content distribution networks for improving performance and scalability of similar applications. Similarly, Chang et al. crawled YouTube for four months in early 2007 collecting data for more than 3 million videos [11]. Their study did not consider the rate of request arrivals for the different videos but rather focused on some statistics such as the video category, length, size and bitrate. They also discuss some of the social networking aspects of YouTube.

Barker and Shenoy studied the effect of background workloads on the QoS of a multimedia service hosted on a cloud system [9]. They show that co-located applications can affect the QoS perceived by the multimedia service customer considerably. The degree of interference variations is most pronounced for disk-bound latency-sensitive tasks, which can degrade by nearly 75% under sustained background load. Their experiments revealed two main insights, the lack of proper disk isolation mechanisms in the hypervisor between co-located VMs can hurt performance, and that network isolation mechanisms in the hypervisor present a trade-off between mean latency and metrics such as jitter and timeouts. Having dedicated caps on the network usage yield lower average latency, while fair sharing the network between the VMs yields lower timeouts and somewhat lower jitter.

5. CONCLUSION

Video-on-Demand workloads are not well studied in the literature. Our analysis of VoD traces from a Swedish service provider aims to add some insights to better understand and design VoD systems. The results of our analysis show that the user and request arrival rates *can not* be modeled as a Poisson process in the analyzed traces. The arrival rates can be modeled using a lognormal distribution while the inter-arrival time can be modeled using an extended exponential distribution. There are four spikes in the workload caused by football matches that interested the Swedish audience. Three of these matches were played in foreign football leagues, unrelated or weakly related to Sweden, making the spikes in the load hard to plan for without extensive social analysis of what attracts the local population. Comparing the user behavior in our study to the user behavior in Yu et al.'s study [45], we can conclude that the rate of users abandoning streaming sessions a few minutes from when they start it seems to be an invariant in VoD workloads. In both studies, 50% of the sessions started were abandoned after less than 12 minutes from their beginning by the highly"impatient users" of the VoD services. That is despite of our study being conducted on a Swedish VoD service and their study being conducted on a Chinese VoD Service with almost six years between the two studies. This impatient behavior can be used to improve prefetching and caching of the videos provided by the VoD service provider.

6. ACKNOWLEDGMENT

We thank TV4 for providing us with the workload traces. Financial support has been provided in part by the Swedish Government's strategic effort eSSENCE, the European Unions Seventh Framework Programme under grant agreement 610711 and the Swedish Research Council (VR) under contract number C0590801 for the project Cloud Control.

7. REFERENCES

[1] Amazon Elastic Compute Cloud (Amazon EC2). `https://aws.amazon.com/solutions/case-studies/`. Accessed: October, 2014.

[2] L. A. Adamic. Zipf, power-laws, and pareto-a ranking tutorial. *Xerox Palo Alto Research Center, Palo Alto, CA,*

http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html, 2000.

[3] V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z.-L. Zhang. Unreeling netflix: Understanding and improving multi-cdn movie delivery. In *INFOCOM, 2012 Proceedings IEEE*, pages 1620–1628. IEEE, 2012.

[4] A. Ali-Eldin, M. Kihl, J. Tordsson, and E. Elmroth. Efficient provisioning of bursty scientific workloads on the cloud using adaptive elasticity control. In *ACM ScienceCloud*, pages 31–40. ACM, 2012.

[5] A. Ali-Eldin, A. Rezaie, A. Mehta, S. Razroev, S. Sjöstedt-de Luna, O. Seleznjev, J. Tordsson, and E. Elmroth. How will your workload look like in 6 years? analyzing wikimedia's workload. In *IEEE IC2E*, pages 349–354, 2014.

[6] A. Ali-Eldin, O. Seleznjev, S. Sjöstedt-de Luna, J. Tordsson, and E. Elmroth. Measuring cloud workload burstiness (to appear). In *UCC*. IEEE Computer Society, 2014. Preprint available at: `https://www8.cs.umu.se/~ahmeda/CloudControl6.pdf`.

[7] M. Arlitt and T. Jin. A workload characterization study of the 1998 world cup web site. *IEEE Network*, 14(3):30–37, 2000.

[8] B. Atikoglu, Y. Xu, E. Frachtenberg, S. Jiang, and M. Paleczny. Workload analysis of a large-scale key-value store. In *PER*, volume 40, pages 53–64. ACM, 2012.

[9] S. K. Barker and P. Shenoy. Empirical evaluation of latency-sensitive application performance in the cloud. In *ACM SIGMM MMSys '10*, pages 35–46. ACM, 2010.

[10] N. Bobroff, A. Kochut, and K. Beaty. Dynamic placement of virtual machines for managing sla violations. In *IEEE IM*, pages 119–128. IEEE, 2007.

[11] X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *IWQoS*, pages 229–238. IEEE, 2008.

[12] Y. Choi, J. A. Silvester, and H.-c. Kim. Analyzing and modeling workload characteristics in a multiservice ip network. *Internet Computing, IEEE*, 15(2):35–42, 2011.

[13] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[14] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990.

[15] S. N. Demographics. Global internet phenomena report: Autumn 2013, 2013.

[16] P. Flandrin, G. Rilling, and P. Goncalves. Empirical mode decomposition as a filter bank. *Signal Processing Letters, IEEE*, 11(2):112–114, 2004.

[17] L. Gao and D. Towsley. Supplying instantaneous video-on-demand services using controlled multicast. In *Multimedia Computing and Systems, 1999. IEEE International Conference on*, volume 2, pages 117–121. IEEE, 1999.

[18] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: A view from the edge. In *ACM IMC*, pages 15–28. ACM, 2007.

[19] K. HIRAI. A philosophy of kando: Cultivating curiosity to reclaim the power of wow. online: http://blog.sony.com/press/sony-ces-2014-keynote-transcript/.

[20] C. Huang, J. Li, and K. W. Ross. Can internet video-on-demand be profitable? *ACM SIGCOMM Computer Communication Review*, 37(4):133–144, 2007.

[21] N. E. Huang and S. S. Shen. *Hilbert-Huang transform and its applications*, volume 5. World Scientific, 2005.

[22] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998.

[23] N. E. Huang and Z. Wu. A review on hilbert-huang transform: Method and its applications to geophysical studies. *Reviews of Geophysics*, 46(2), 2008.

[24] M. Jeon, Y. Kim, J. Hwang, J. Lee, and E. Seo. Workload characterization and performance implications of large-scale blog servers. *ACM TWEB*, 6(4):16, 2012.

[25] S. Jin and A. Bestavros. GISMO: a generator of internet streaming media objects and workloads. *ACM SIGMETRICS PER*, 29(3):2–10, 2001.

[26] X. Kang, H. Zhang, G. Jiang, H. Chen, X. Meng, and K. Yoshihira. Understanding internet video sharing site workload: A view from data center design. *Journal of Visual Communication and Image Representation*, 21(2):129–138, 2010.

[27] M. Kendall, A. Stuart, and J. K. Ord. The advanced theory of statistics. *The advanced theory of statistics.*, (4th Ed), 1983.

[28] S. Khemmarat, R. Zhou, L. Gao, and M. Zink. Watching user generated videos with prefetching. ACM MMSys, pages 187–198. ACM, 2011.

[29] M. Kihl, E. Elmroth, J. Tordsson, K.-E. Årzén, and A. Robertsson. The challenge of cloud control. In *8th International Workshop on Feedback Computing*, 2013.

[30] M. Kutare, G. Eisenhauer, C. Wang, K. Schwan, V. Talwar, and M. Wolf. Monalytics: online monitoring and analytics for managing large scale data centers. In *ICAC*, pages 141–150. ACM, 2010.

[31] J. Laherrere and D. Sornette. Stretched exponential

distributions in nature and economy:fat tails with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.

[32] R. Lawler. Verizon taps clearleap for cloud-based vod content delivery. online: http://gigaom.com/2010/07/26/verizon-taps-clearleap-for-cloud-based-vod-content-delivery/.

[33] R. McGill, J. W. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[34] R. Nathuji, A. Kansal, and A. Ghaffarkhah. Q-clouds: managing performance interference effects for QoS-aware clouds. In *ACM EuroSys*, pages 237–250. ACM, 2010.

[35] D. Niu, H. Xu, B. Li, and S. Zhao. Quality-assured cloud bandwidth auto-scaling for video-on-demand applications. In *INFOCOM, 2012 Proceedings IEEE*, pages 460–468. IEEE, 2012.

[36] V. Paxson and S. Floyd. Wide area traffic: the failure of poisson modeling. *IEEE/ACM Transactions on Networking (ToN)*, 3(3):226–244, 1995.

[37] L. Plissonneau and E. Biersack. A longitudinal view of http video streaming performance. In *ACM MMSys*, pages 203–214. ACM, 2012.

[38] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *ACM SoCC*, pages 7:1–7:13, 2012.

[39] G. Rilling, P. Flandrin, P. Goncalves, et al. On empirical mode decomposition and its algorithms. In *IEEE-EURASIP workshop on nonlinear signal and image processing*, volume 3, pages 8–11. NSIP-03, Grado (I), 2003.

[40] B.-M. Ringfjord. Learning to become a football star : Representations of football fan culture in swedish public service television for youth. In *We love to hate each other : mediated football fan culture*, pages 285–299. 2012.

[41] P. Svärd, B. Hudzia, J. Tordsson, and E. Elmroth. Evaluation of delta compression techniques for efficient live migration of large virtual machines. *ACM Sigplan Notices*, 46(7):111–120, 2011.

[42] L. Tomas and J. Tordsson. An autonomic approach to risk-aware data center overbooking. *Cloud Computing, IEEE Transactions on*, 2014. Pre-print.

[43] Z. Wu and N. E. Huang. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in adaptive data analysis*, 1(01):1–41, 2009.

[44] H. Xi, J. Zhan, Z. Jia, X. Hong, L. Wang, L. Zhang, N. Sun, and G. Lu. Characterization of real workloads of web search engines. In *IISWC*, pages 15–25. IEEE, 2011.

[45] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. In *ACM EuroSys*, pages 333–344. ACM, 2006.