Measurement and Analysis of Cloud User Interest: A Glance From BitTorrent

Lei Ding
University of Alberta
Edmonton, Alberta, Canada
Email: lding1@ualberta.ca

Yang Li, Haiyang Wang
University of Minnesota Duluth
Duluth, Minnesota, USA
Email: {yangli, haiyang}@d.umn.edu

Ke Xu
Tsinghua University, Beijing, China
Email: xuke@tsinghua.edu.cn

***Abstract*—Cloud computing has recently emerged as a compelling method for deploying and delivering services over the Internet. In this paper, we aim to shed new light on the learning of cloud user interest. Our study for the first time shows the existence of cloud users in such real-world content distribution systems as BitTorrent. Based on this observation, we further explore the similarity of content preferences between cloud and non-cloud users. Surprisingly, our statistical model analysis indicates that the users in the cloud AS have significantly different interests from all the observed non-cloud ASes. More dedicated researches are therefore required to better manage this elevating yet unique cloud traffic in the future.**

1. Introduction

Cloud computing has rapidly emerged as the driving trend in global Internet services. The existing cloud-based measurement studies are, however, limited to the resource usage behavior as well as system scaling issues [1], [2]. To mitigate such a challenge, our study for the first time explores the existence of cloud users in such real-world content distribution systems as BitTorrent. We successfully captured the existence of cloud peers in BitTorrent (BT) networks. This observation naturally bridges Internet P2P systems to cloud computing. In particular, our measurement indicates that 17 percent of BT torrents has over 10% peers from cloud. The ratio of cloud peers can even exceed 50% for some very popular torrents. Moreover, the existence of cloud user in BT also provides an initial yet important step to understand the similarity of cloud and non-cloud autonomous systems (ASes) and the cloud users are distributed inhomogeneously over the Internet torrents.

It is worth noting that we can hardly compare the cloud AS to all other ASes on the Internet. It is therefore hard to say if the cloud users are indeed having different interests from all the other users. To this end, we further developed a novel clustering approach to measure the similarity of user interests across all the observed ASes. This approach successfully distinguishes Amazon cloud with other non-cloud ASes. This implies that the cloud users/ASes have significantly different interests from all other non-cloud ASes. Therefore, we will need to build special and more dedicated traffic management strategies to manage/optimize the cloud users in the future.

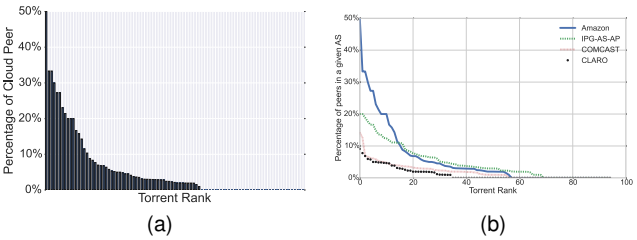

Figure 1. (a) Percentage of cloud peers in torrents (b) Peers ratio of different ASes

2. Measurement of Cloud User Interest From BitTorrent

In this experiment, we have applied a PlanetLab-based experiment to obtain the peer information of Internet BitTorrent swarms. It is known that the most popular torrents generate the majority of BT traffic [3]. To simplify our later statistical analysis, we selected the top 100 most popular torrents for discussion. The selection was done in a completely random way so the selection bias could be neglected in this study.

Different from the existing P2P measurement studies, our dataset indicates that a considerable number of BT peers are from cloud. Our investigation indicates that these IP addresses are assigned to Amazon's EC2 [4] virtual machines (VMs). As a popular cloud service provider and an autonomous system[1], Amazon is ranked at the second most popular AS in our dataset. More cloud peers will be observed if we further extend the scale of our analysis. In Figure 1a, we can see the popularity of cloud peers in different torrents. In detail, many torrents have over 10% peers from cloud. The ratio of cloud peers can even exceed 50% for some very popular torrents. This means the existence of cloud peers is not a special case in BT torrents. Therefore, we can use this information to understand the cloud user interest in BitTorrent.

To explore cloud user's preference, Figure 1b compares Amazon to three very popular ASes in our dataset. We

1. It is known that Amazon consists of many autonomous systems. For the sake of simplicity, we use one AS (ASN:16509) to refer Amazon in this paper.

can see that the users in typical non-cloud ASes, such as *COMCAS* and *CLARO*, are more equally distributed in different torrents. The distribution of Amazon users, on the other hands, is clearly skewed. This means the cloud users are more likely to have clear preferences on certain types of contents/torrents which are movie and TV contents. Very few cloud users are willing to join torrents with music and software applications.

Based on the above measurements, it is easy to see that the cloud users are now an emerging force in such Internet applications as BitTorrent. These users also have a clear preference on movie and TV contents. However, we do not know if their detailed interests are similar/dissimilar to all other non-cloud ASes.

3. Statistical Analysis

The existence of the cloud-based AS (Amazon) brings up the question of learning its characteristics and its difference from the traditional ASes. The data set can be described by an $m \times n$ data matrix $\mathbf{A}$, where rows and columns correspond to torrent files and distinct ASes, respectively. Element A_{ij} represents the count of peers in the jth AS which were downloading the ith torrent file. Each AS can therefore be represented by a m-dimensional vector called its *profile*. The profile of the kth AS is $I_k = (A_{1k}, A_{2k}, \ldots, A_{mk})$ where $k = 1, 2, \ldots, n$. For an easy comparison, Amazon cloud is placed in the first column while all other traditional ASes are ranked in decreasing order by their total number of peers. The structure of the data set can alternatively be illustrated using a weighted bipartite graph.

We perform a projection onto the AS space by constructing an n-vertex simple graph where vertices represent ASes and two ASes are connected by an edge with a weight corresponding to their similarity. We applied the Pearson correlation coefficient which is a measure of linear correlation between two ASes profiles, defined as

$$r_{k,l} = \frac{\sum_{i=1}^{m}(A_{ik} - \bar{A}_k)(A_{il} - \bar{A}_l)}{\sqrt{\sum_{i=1}^{m}(A_{ik} - \bar{A}_k)^2}\sqrt{\sum_{i=1}^{m}(A_{il} - \bar{A}_l)^2}}, \quad (1)$$

where $\bar{A}_k$ is the mean of the kth profile. An $n \times n$ correlation matrix $\mathbf{B}$ with elements $B_{kl} = r_{k,l}$ from (1). The column-wise mean value of $\mathbf{B}$ represents the overall similarity of a given AS to all other ASes in the study. Figure 2 shows a bar plot of these mean values. In particular, the first bar, which represents Amazon cloud, is the only one with negative mean value. It shows that, on average, Amazon is dissimilar to other traditional ASes in terms of peers' downloading behaviors.

To better understand the patterns of all ASes, we perform a clustering analysis to group those ASes with similar downloading behaviors in the same clusters. Since the bigger the value B_{kl}, the more similarity between two profiles, B_{kl} can be considered as a "distance", such that I_k and I_l have a small distance if B_{kl} is large. We work on the matrix $1 - \mathbf{B}$ instead of $\mathbf{B}$ since entries in $1 - \mathbf{B}$ are positively correlated with the distance. Figure 3 shows the hierarchical clustering

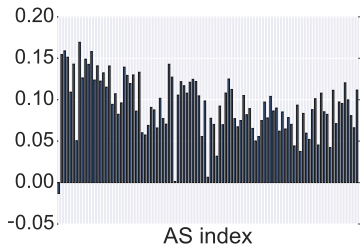

Figure 2. Mean of the Pearson correlation coefficients of an AS to all other ASes. Amazon, shown as the first bar, is the only negative one.

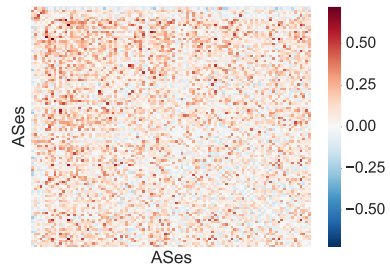

Figure 3. Heat map of the generated clusters ($c = 2$). 101 ASes are divided into two clusters. The first cluster which is shown in the upper left corner has 5 ASes including Amazon cloud. The second cluster has 96 ASes.

result heat map where the ASes are divided into $c = 2$ clusters of size 5 and 96. Amazon cloud is located in the first cluster which is in the upper left corner of Figure 3. It is separated from the majority of the data indicating a different user downloading interest from most of the traditional ASes. For larger values of c, the general picture is very similar. Amazon tends to be in a very small cluster, indicating its distinct behaviors compared to the traditional ASes.

4. Conclusion

This paper takes an initial step towards the understanding of cloud user interest. Our measurement from BitTorrent showed the existence of cloud peers in BT. The follow-up comparison further revealed that the user interest of cloud users/ASes is significantly different from the classic non-cloud users/ASes. For futher work, we are interested in the detailed reasons of why cloud users/ASes are so unique. Moreover, we also aim to explore better traffic management approaches to handle the increasing cloud traffic.

References

[1] O. A. Abdul-Rahman and K. Aida, "Towards understanding the usage behavior of google cloud users: The mice and elephants phenomenon," in *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, December 2014, pp. 272–277.

[2] 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 *Proceedings of the Third ACM Symposium on Cloud Computing*, ser. SoCC '12, 2012, pp. 7:1–7:13.

[3] S. Le Blond, A. Legout, and W. Dabbous, "Pushing bittorrent locality to the limit," *Comput. Netw.*, vol. 55, no. 3.

[4] Amazon, Inc., "Amazon Elastic Compute Cloud (Amazon EC2)," http://aws.amazon.com/ec2/.