Does Internet Media Traffic Really Follow Zipf-like Distribution?

Lei Guo[1], Enhua Tan[1], Songqing Chen[2], Zhen Xiao[3], and Xiaodong Zhang[1]

[1]The Ohio State University
{lguo, etan, zhang}@cse.ohio-state.edu

[2]George Mason University
sqchen@cs.gmu.edu

[3]IBM Research
xiaozhen@us.ibm.com

Categories and Subject Descriptors: C.2.4 [Computer-Communication Networks]: Distributed Systems

General Terms: Measurement

Keywords: Media, Zipf-like, stretched exponential

1. INTRODUCTION

It is commonly agreed that Web traffic follows the Zipf-like distribution, which is an analytical foundation for improving Web access performance by client-server based proxy caching systems on the Internet. However, some recent studies have observed non-Zipf-like distributions of Internet media traffic in different content delivery systems. Due to the variety of media delivery systems and the diversity of media content, existing studies on media traffic are largely workload specific, and the observed access patterns are often different from or even conflict with each other. For *Web media systems*, study [3] reports that the access pattern of streaming media is Zipf-like in a university campus network, while study [2] finds that it is not Zipf-like in an enterprise media server. For *VoD media systems*, study [1] finds that it is not Zipf-like in a multicast-based Media-on-Demand server of a campus network, while study [9] reports it is Zipf-like in a large VoD streaming system of an ISP. For *P2P media systems*, study [4] reports that the access pattern of media workload in KaZaa system collected in a campus network is not Zipf-like, while study [5] reports that it is Zipf-like in another campus network. For *live streaming media systems*, study [8] reports it is Zipf-like while study [6] reports it is not Zipf-like. A number of models have been proposed to explain the observed media access patterns, such as the generalized Zipf-like model [7], "fetch-at-most-once" model [4], and two-mode Zipf model [6]. However, each of these models can only explain a very limited scope of measurement results. A general model of Internet media access patterns is highly desirable for traffic engineering on the Internet and is critical to design, benchmark, and evaluate Internet media delivery systems.

In this study, we have analyzed a wide variety of media workloads on the Internet. The workloads were collected from both the client side and the server side in Web, VoD, P2P, and live streaming environments between 1998 and 2006. The duration of these workloads ranges from a few days to more than two years and the user population ranges from several thousands to more than one hundred thousand. The number of client requests ranges from tens of thousands to hundreds of million, and the number of objects in each workload ranges from several hundreds to several million. Through extensive analysis, we find that the reference ranks of media objects in all sixteen workloads follow the **stretched exponential (SE) distribution**, and a biased measurement may lead to a Zipf-like observation on media access patterns. With such a request pattern, the temporal locality in media systems is hard to exploit by client-server based caching systems. The stretched exponential model implies that peer-to-peer collaborative caching systems can effectively deliver Internet media content. Current technology advancements such as PPLive and BitTorrent have demonstrated the strong advantages of P2P collaboration on the delivery of Internet media content.

2. THE STRETCHED EXPONENTIAL DISTRIBUTION OF MEDIA TRAFFIC

Figures 1(a), 1(b), 1(c), and 1(d) show the reference rank distributions of media objects in typical Web, VoD, P2P, and Live media systems, respectively. In each figure, the x coordinate represents the reference rank of each object, plotted in log scale, while the y coordinate represents the number of references to this object, plotted in both log scale (marked on the right of y-axis) and a powered scale (by a constant c, as marked on the left of y-axis). These figures show that the reference rank distributions of all these workloads cannot be fitted with a straight line in a log-log scale, meaning they are not Zipf-like. Instead, by selecting a proper constant c, all these workloads can be well fitted with a straight line in log-y^c scale. Such a distribution is called a *stretched exponential distribution*. As marked in the figures, the coefficient of determination of the stretched exponential fitting result, R^2, is very close to 1 for all workloads.

The cumulative probability function of a stretched exponential distribution can be expressed as

$$P(X < x) = 1 - e^{-(\frac{x}{x_0})^c}, \tag{1}$$

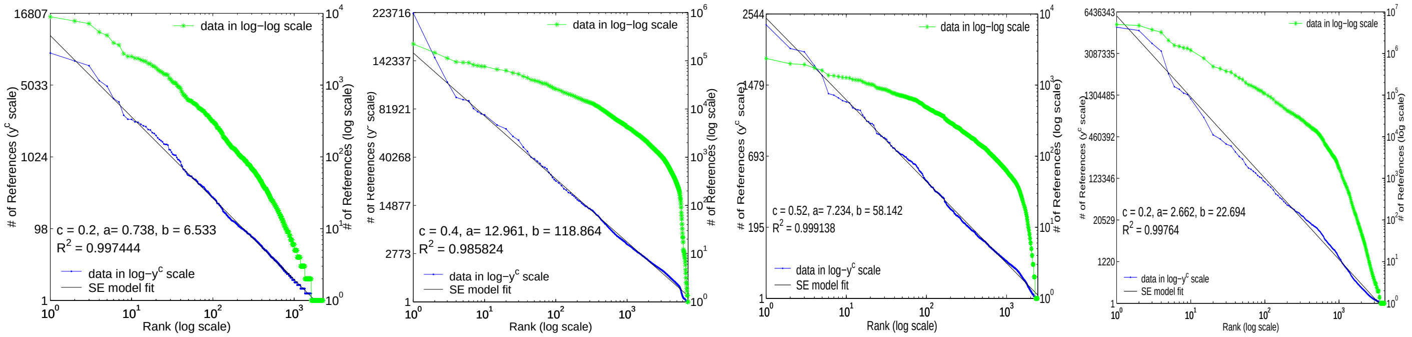
where c and x_0 are constants. If we rank the N objects in the workload in descending order of their reference numbers y_i $(1 \le i \le N)$, we have $P(y_n > y_i) = i/N$. So the reference rank distribution can be expressed as follows

$$y_i^c = -a \log i + b \ \ (1 \le i \le N), \tag{2}$$

where $a = x_0^c$ and $b = y_1^c$. Since b is a normalization parameter, the shape of an SE distribution is determined by c, the *stretch factor* of y coordinate, and a, the slope of the straight line in log-y^c scale.

For on-demand media systems, the stretch factor c of the object reference rank distribution is highly related with the

SIGMETRICS'07, June 12–16, 2007, San Diego, California, USA.
ACM 978-1-59593-639-4/07/0006.



(a) Web media system (b) VoD media system (c) P2P media system (d) Live media system

Figure 1: Reference rank distributions of different kinds of media systems

sizes of files in the system. In general, for media workloads delivered by similar kinds of systems or techniques, the stretch factors of their corresponding object reference rank distributions increase with their median file sizes; for workloads with similar median file sizes, the stretch factors of their corresponding object reference rank distributions are similar regardless of their underlying media systems and delivery techniques. Furthermore, for objects accessed in different time periods in a media system with roughly constant object birth rate, the stretch factor c of corresponding reference rank distributions is a time-invariant constant.

For media systems with roughly constant request rates and object birth rates over time, the parameter a (the slope of the SE line in log-y^c scale) of the object reference rank distribution increases with its stretch factor c and the average number of requests per object in the workload. Furthermore, due to the increase of the average number of requests per object over time, parameter a increases with the length of the workload duration gradually but converges to a constant, which is determined by the ratio of the media request rate to the object birth rate and the stretch factor c.

For a stretched exponential reference rank distribution with slope a in log-y^c scale and total N objects, the difference between this distribution and its corresponding Zipf-like model in log-log scale increases with $a \log N$. For a workload with large media files, both the average number of requests per object and the stretch factor c are large. Thus a is large, and the difference between its reference rank distribution and the corresponding Zipf-like model is large. For a workload with small media files, the difference between its reference rank distribution and the corresponding Zipf-like model is also large when the workload duration is long enough (at least months to years).

3. IMPLICATIONS ON MEDIA CACHING

Internet media objects commonly have long lifespans because they are seldom updated and have low production rates compared to Web objects. Most requested media objects are created long time ago, and most media requests are for objects created long time ago. For example, for a media workload collected at a large residential cable network in 2005, more than 50% requested objects are created at least 250 days ago, and more than 50% requests are for objects older than 150 days.

The temporal locality in a computer system comes from the concentration and correlation of requests to the content in the system. During a short period such as one week, the popularity of media objects is almost *stationary*, thus the temporal locality mainly comes from the request concentration. We have modeled the optimal hit ratios of typical short term media workloads and Web workloads, where request concentration dominates the temporal locality. In such cases, caching of media (SE) workload is far less efficient than that of Web (Zipf) workload. For example, assuming all objects are cachable and have the same file size, caching 1% Web content can achieve about 40% hit radio, while caching 1% media content can only achieve 18% hit ratio, even though they have the same hit ratio with an unlimited cache.

Nevertheless, the request concentration in a media workload (parameter a) increases with time. Furthermore, due to the long lifespan of media objects, the request correlation becomes important with time. With a much higher temporal locality, long-term caching can have a high hit ratio greater than 85% with caching 10% content. However, it may take months to years and a huge amount of storage to achieve such an improvement, for which peer-to-peer techniques can be much effective.

4. CONCLUSION

Our study shows that Internet media access patterns follow the stretched exponential distribution. Thus, the performance of media caching with a client-server model is far less effective than that of Web content caching. The stretched exponential distribution lays out an analytical foundation to establish peer-to-peer caching systems for delivering the rapidly increasing Internet media content.

5. REFERENCES

[1] S. Acharya, et al. Characterizing user access to videos on the world wide web. In *Proc. of MMCN*, 2000.

[2] L. Cherkasova, et al. Characterizing locality, evolution, and life span of accesses in enterprise media server workloads. In *Proc. of NOSSDAV*, May 2002.

[3] M. Chesire, et al. Measurement and analysis of a streaming media workload. In *Proc. of USENIX USITS*, March 2001.

[4] K. P. Gummadi, et al. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proc. of ACM SOSP*, October 2003.

[5] A. Iamnitchi, et al. Small-world file-sharing communities. In *Proc. of IEEE INFOCOM*, March 2004.

[6] K. Sripanidkulchai, et al. An analysis of live streaming workloads on the Internet. In *Proc. of ACM SIGCOMM IMC*, October 2004.

[7] W. Tang, et al. Medisyn: A synthetic streaming media service workload generator. In *Proc. of ACM NOSSDAV*, June 2003.

[8] E. Veloso, et al. A hierarchical characterization of a live streaming media workload. In *Proc. of the ACM SIGCOMM IMW*, November 2002.

[9] H. Yu, et al. Understanding user behavior in large scale video-on-demand systems. In *Proc. of EuroSys*, April 2006.