The Interplay Between Caching and Popularity

Majed Haddad, Eitan Altman

The Interplay Between Caching and Popularity

Majed Haddad and Eitan Altman
INRIA Sophia Antipolis, 10 route des Lucioles,
06902 Sophia Antipolis, France.

***Abstract*—The increased availability of meta-data in Web 2.0 (as opposed to traditional Web) can and should be exploited to make such techniques more effective. Caching should improve the performance and scalability of multimedia service streaming (e.g., YouTube). In this contribution, we introduce new directions and considerations in the analysis of caching popular content in the Web which allows us to gain insight on deriving more informative indications for quality of service development. We provide a dynamic model for the impact of popularity on the access speed due to caching policies of a service provider. More specifically, we assume that caches are spatially deployed as a Poisson distribution and that users are distributed over the geographical area in a Poissonian manner. Our model is formulated as epidemic type process of file dissemination. We then study the transient behavior of caches where information is replicated and disseminated according to an epidemic type dynamics based on the popularity of the content. Simulation results show that the proposed scheme provides significant improvement in terms of the system throughput.**

***Index Terms*—Caching, YouTube, epidemic model, mean field approximation.**

I. Introduction

The analysis of the YouTube workload [1, 2, 3] emphasizes the fact that access patterns are strongly correlated with human behaviors (e.g., time-of-day, day-of-week). Moreover, video files are much larger than files of other types, and some videos are more popular than others. These and other characteristics suggest that caching should improve the performance and scalability of YouTube videos. In particular, caching popular videos in the proxy in the client site can reduce the client access time and start up delay in watching video. However, enabling anyone to publish content means growth in content will not only be larger than for traditional Web and media, but sustainable. This will place greater strain on centralized resources, and require decentralized approaches such as caching and (Content Delivery Networks) CDNs. In addition, since much of the content is likely to be unpopular (the long tail effect), it will be important to minimize the cost for storing these contents. This has several implications for the service provider, who must plan, purchase, install, operate and maintain the central infrastructure used by the site. To do so, one of the design criteria is the popularity distribution of video clips requested by users in the access network[1]. The higher the popularity is, the fastest it will take to reach it if the access to the content is ranked according to measures such as its page rank. Despite there is considerable published literature on caching of adaptive multimedia streams [4, 5, 6], we found that few studies have focused to derive analytical models from the traces.

[1]Yet other measures are recommendation systems and publicity.

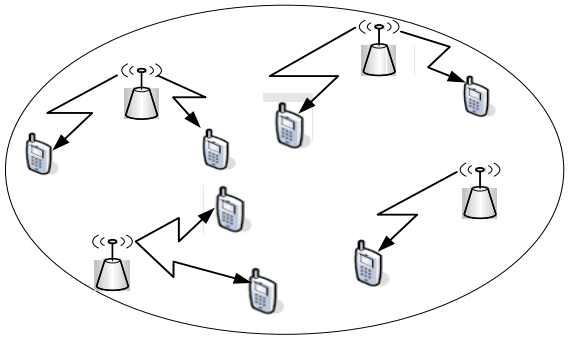

Fig. 1. Network Model for the cache.

In this work, we shall assume that some content becomes available and we wish to model how it propagates. We assume that two factors affect the rate of propagation: the access provider, through a large network of caches, and the content provider, through a ranking procedure that favors visibility of popular content. The content is accessible through downloaded from some site that specializes in such contents. The site is assumed to have some subscribers or some public of fans that connect to it regularly. A user that downloads a content has the option of recommending the content to others by clicking on a recommendation tag. The site keeps track of how many accessed the content's description, how many downloaded it and how many recommended it. These figures are collected over a given period which may be the total life time of the content or the last week or the last day. Below, we consider the case of the first configuration only. This information is used to rank the files. The higher the rank is, the larger the probability to be aware of it and then to download it. This is due to the fact that those searching for similar contents will be offered the various possibilities at the order of their *page rank*. Moreover, the most popular content may be explicitly displayed in social network like Facebook or in media services like YouTube. A caching strategy is used which enables to decrease the download time as a content becomes more popular.

II. Modeling the Caching

We provide a dynamic model for the impact of popularity on the access speed due to caching policies of a provider.

We study a network made up of M users or nodes that are potentially interested in some content. At a given time t, an amount x_t (normalized by the population size M) of the content has been already downloaded. We assume that the content provider has a cache policy according to which the probability of finding a copy of the content at a cache is proportional to its x_t which is considered as a popularity index.

We assume that the caches are of the form of small base stations (or throwboxes) scattered over the space according to a Poisson process with a rate λ (see Figure 1). We assume that caching policy is such that a content is found at a cache independently on other cases. We consider the following simple caching strategy: the probability that a content is given at a cache is proportional to the number of downloads. At time t it is given by $\alpha \cdot x_t$. We consider here mobile users that connect through a wireless channel to the closest cache. Thus, the caches containing the content have a Poisson distribution over the space, with parameter $\lambda \cdot \alpha \cdot x_t$. This means that the number of caches that contain the file in a zone of area B is a Poisson random variable with rate $\lambda \cdot \alpha \cdot x_t \cdot B$.

III. Performance Analysis

The signal to noise ratio γ, received by a mobile at a distance d from a cache is assumed to be determined by the distance d to the cache and is given according to $\gamma(d) = B_0 \cdot d^{-\beta}$, where B_0 is a proportionality constant and β is the pathloss coefficient. Consider the DBPSK modulation. Then, at low Signal to Noise Ratio (SNR) regime, the bit error rate (BER) is [7]: $BER = \frac{1}{2} \cdot \exp(-\gamma)$. Assume that K bits form one packet and that bit errors are i.i.d. The packet success rate governing the geometric successful transmission distribution is given by $(1 - BER)^K$ and the packet error rate (PER) is determined by $PER = 1 - (1 - BER)^K = 1 - (1 - \frac{1}{2}\exp(-\gamma))^K$. Approximating the PER, we obtain that $PER = \frac{K}{2}\exp(-\gamma)$. Let $z = \frac{K}{2}\exp(-\gamma)$. In order to analyze the performance of the proposed caching policy in terms of achievable throughput, we resort to compute the expected exponential SNR over the space, namely $\mathbb{E}(z)$.

Choose an arbitrary user. Assume that (i) a packet is retransmitted until received correctly, and (ii) that x_t varies slowly in the following sense: during the period of transmission and retransmission of n packets, where n is some large integer, x_t is constant. Then the time it takes to receive a packet whose transmission starts at time t is geometrically distributed with parameter $K \cdot \gamma$. We make the observation that the distance from that user to the closest cache with the content is greater than d if and only if there is no cache with a copy of the file within a circle of radius d centered at that user. It follows that the probability that the closest cache from a mobile is at a distance of at least d is

$$Prob\,(y \geq d) = \exp(-\nu(x_t) \cdot B(d)) \tag{1}$$

where $\nu(x_t) = \lambda \cdot \alpha \cdot x_t$ and $B(d)$ is the volume of a ball with radius d. It is $2 \cdot d$ in dimension 1 (line case), and is $\pi \cdot d^2$ in dimension 2 (plane case). The probability density of the distance to the closest file is then given by:

$$\phi_x(d) \ = \nu(x_t) \cdot \tfrac{\partial B(d)}{\partial d} \cdot \exp(-\nu(x_t) \cdot B(d)) \tag{2}$$

Combining with (2), we see that the expected exponential SNR of a user is given by

$$\mathbb{E}(z) \ = \tfrac{K}{2} \int_0^\infty \exp(-B_0 \cdot y^{-\beta}) \cdot \phi_x(y) dy \tag{3}$$

IV. Mean Field Approximation

Under two-hop forwarding assumption, users can obtain the content from the cache, but they are not able to infect other nodes, hence the destination can be reached at most in two hops. At some time, say $t = 0$, the content becomes available. We assume that download is done at a fixed rate of μ, and the size of the content is of one unit. The expected time to transmit the packet is $1/(K \cdot \gamma \cdot \mu)$. The average number of packets transmitted[2] in a time unit at $[t, t+\Delta)$ is approximately $K \cdot \gamma \cdot \mu \cdot \Delta$. Assume that there are ζ packets in the content that is transferred. As the packet spreads at a rate proportional to its throughput, we can use the following equations to predict the impact of the proposed caching policy on the system dynamics using the mean field limit [8]:

$$\dot{x}_t = \frac{K \cdot \mu}{\zeta} \cdot (1 - x_t) \cdot g(x_t) \tag{4}$$

where the throughput $g(x_t)$ is function of the popularity and given by $g(x_t) = \min(c/\sqrt{\mathbb{E}(PER)}, T_{max})$, c is a proportionality constant and T_{max} is an upper bound throughput allowed by the codecs.

A. Line case

Let us first consider the line case. The expression of the expected loss probability is given by:

$$\mathbb{E}(PER) = K \cdot \nu(x_t) \cdot \int_0^\infty \exp(-B_0 \cdot y^{-\beta} - 2 \cdot \nu(x_t) \cdot y)\, dy \tag{5}$$

In what follows, we derive an explicit form of $\mathbb{E}(PER)$ depending on the pathloss parameters β:

For β=2

$$\mathbb{E}(PER) = \frac{K\sqrt{B_0}}{\sqrt{2\pi}} \cdot \nu(x_t) \cdot G_{0,3}^{3,0}\left(\nu(x_t)^2 B_0 \,\middle|\, {}_{1/2,0,-1/2}\right) \tag{6}$$

where G is the Meijer G-function.

For β=3

$$\mathbb{E}(PER) = \frac{K\sqrt{3}\sqrt[3]{B_0}}{6\,\pi} \cdot \nu(x_t) \cdot G_{0,4}^{4,0}\left(\frac{8}{27}\,\nu(x_t)^3 B_0 \,\middle|\, {}_{2/3,1/3,0,-1/3}\right) \tag{7}$$

For β=4

$$\mathbb{E}(PER) \ = \frac{K\sqrt{2}\sqrt[4]{B_0}}{8\,\pi^{3/2}} \cdot \nu(x_t) \cdot \tag{8}$$

$$G_{0,5}^{5,0}\left(1/16\,\nu(x_t)^4 B_0 \,\middle|\, {}_{3/4,1/2,1/4,0,-1/4}\right)$$

[2] over some m consecutive packets with $m \leq n$.

B. Plane case

Considering the case of the plane, the expected loss probability can be expressed as:

$$\mathbb{E}(PER) = K \cdot \pi \cdot \nu(x_t) \cdot \int_0^{\infty} \exp(-B_0 \cdot y^{-\beta} - \pi \cdot \nu(x_t) \cdot y^2) \cdot y \, dy \tag{9}$$

Similarly to the line case, we obtain the following explicit expressions for different pathloss parameter β:

For β=2

$$\mathbb{E}(PER) = K \sqrt{\pi B_0 \nu(x_t)} \text{BesselK}_1\left(2\sqrt{\pi B_0 \nu(x_t)}\right) \tag{10}$$

where $\text{BesselK}_n(\cdot)$ is the modified Bessel function of the second kind.

For β=3

$$\mathbb{E}(PER) = \frac{K (B_0)^{2/3} \sqrt{3} \sqrt[3]{2}}{24\sqrt{\pi}} \nu(x_t) \cdot G_{0,5}^{5,0}\left(\frac{1}{108} \pi^3 {B_0}^2 \nu(x_t)^3 \,\middle|\, {}_{2/3,1/3,1/6,0,-1/3}\right) \tag{11}$$

For β=4

$$\mathbb{E}(PER) = \frac{K\sqrt{\pi B_0}}{4} \nu(x_t) \cdot G_{0,3}^{3,0}\left(\frac{1}{4} \pi^2 B_0 \nu(x_t)^2 \,\middle|\, {}_{1/2,0,-1/2}\right) \tag{12}$$

V. Numerical Results

To go further with the analysis, we resort to numerical results. We choose to consider the plane case where users can profit from throughputs up to $T_{max} = 1Mbps$ where $K = 16$, $\lambda = 0.1$, $\alpha = 0.1$, $\zeta = 1000$ and $\mu = 1024$. Figure 2 captures the variation of the system throughput as function of the ratio of users who already have a copy of the content x. As intuition would expect, it is shown that the system throughput increases as x increases. Obtaining $g(x)$, we can compute the solution of this ordinary differential equation and plot the dynamic of the number of infected users for increasing time t. Figure 3 depicts the amount of users that have already downloaded the content. It is shown that all users get the file at $t = 0.7$ seconds, respectively $t = 0.9$ seconds, for $\beta = 2$, respectively $\beta = 4$.

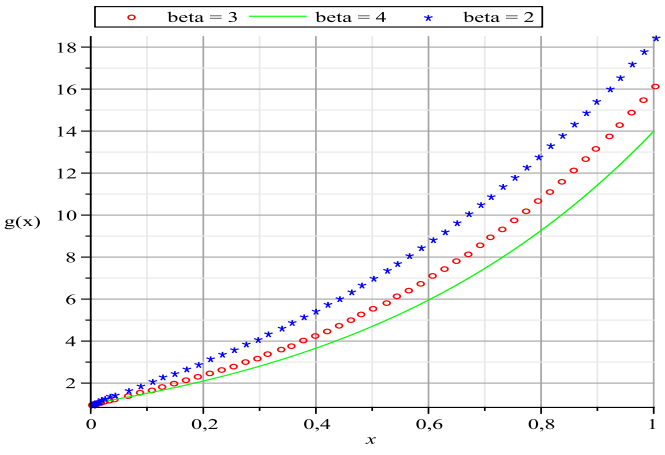

Fig. 2. System throughput as function of the number of users who already have a copy of the content for the plane geometrical system model.

In Figure 4, we plot the variation of the system throughput as function of the time for different pathloss parameter. Notice here that the system throughput is upper-bounded at a certain time. As an example, for $\beta = 2$ we gain 2 bits/sec/Hz of system throughput compared to the case where $\beta = 3$ and barely 4 bits/sec/Hz with respect to the case where $\beta = 4$.

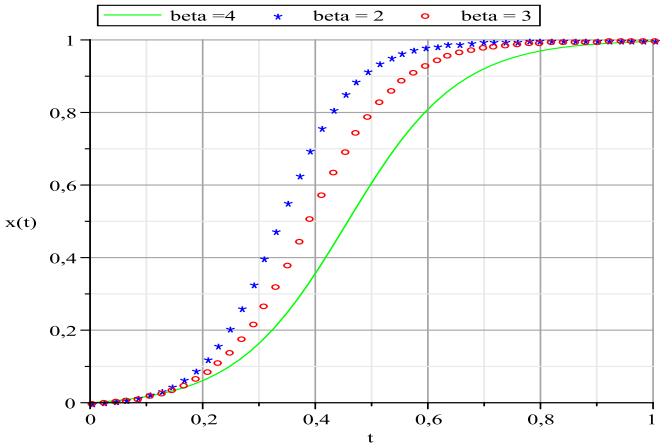

Fig. 3. Ratio of infected users as function of time for the plane geometrical system model.

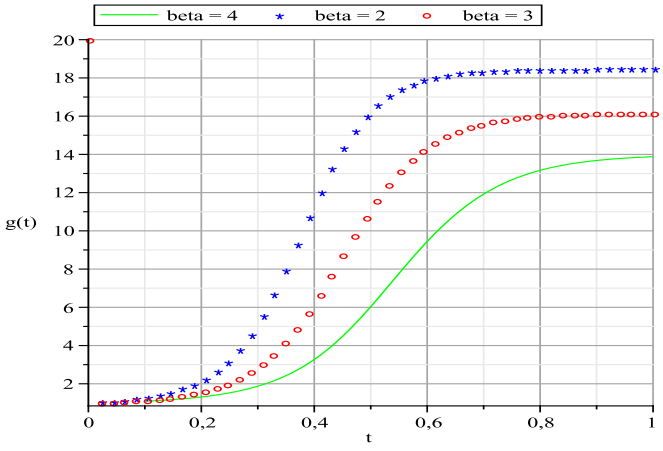

Fig. 4. System throughput as function of time for the plane geometrical system model.

VI. Conclusion

Modern Internet streaming services like YouTube have utilized various techniques to improve the quality of streaming media delivery. As an example, caching the most popular data at proxies close to clients is an efficient approach to save bandwidth and prevent user latency. In particular, the increased

availability of meta-data in media services like YouTube (a direct result of social networking) can and should be exploited to make such models more effective. In this paper, we have provided a dynamic model for the impact of popularity on the access speed due to caching policies of a service provider which allows us to gain insight on deriving more informative indications for quality of experience requirements of the Web traffic and services. As a future work, we intend to guarantee the accuracy of the approximate fluid model including the arguments justifying the existence of the ODE solution

REFERENCES

[1] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the edge," in *In: Proc. of IMC*, 2007.

[2] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of YouTube network traffic at a campus network - measurements, models, and implications," *Computer Networks*, vol. 53, no. 4, pp. 501 – 514, 2009.

[3] M. Cha, H. Kwak, P. Rodriguez, Y. yeol Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the world's largest user generated content video system," in *In Proceedings of the 5th ACM/USENIX Internet Measurement Conference (IMC07)*, 2007.

[4] R. Rejaie, H. Yu, M. H, and D. Estrin, "Multimedia proxy caching mechanism for quality adaptive streaming applications in the internet," 2000, pp. 980–989.

[5] J. Kangasharju, F. Hartanto, M. Reisslein, and K. W. Ross, "Distributing layered encoded video through caches," in *IEEE Transactions on Computers*, 2001, pp. 622–636.

[6] S. Chattopadhyay, L. Ramaswamy, and S. M. Bhandarkar, "A framework for encoding and caching of video for quality adaptive progressive download," in *Proceedings of the 15th international conference on Multimedia*, 2007, pp. 775–778.

[7] J. G. Proakis, *Digital Communications*, 4th ed. McGraw-Hill, August 2001.

[8] T. G. Kurtz, "Solutions of ordinary differential equations as limits of pure jump markov processes," *Journal of Applied Probability*, vol. 7, pp. 49–58, April 1970.