The performance of a LRU cache under dynamic catalog traffic

Felipe Olmos[1], Bruno Kauffmann[2], Alain Simonian[2], Yannick Carlinet[2]

Orange Labs and CMAP École Polytechnique
[1]luisfelipe.olmosmarchant@orange.com
[2]firstname.lastname@orange.com

Abstract

We propose a simple traffic model featuring a dynamic catalog to construct a theoretical estimation of the hit ratio for a LRU cache offered such a traffic regime. We validate the accuracy of our theoretical estimates by computing the empirical hit ratio for real request sequences coming from traces of the Orange network.

1. Introduction

Caching performance evaluation is a relevant topic today in the context of Content Delivery Networks.

Most traffic models used to estimate the performance of a cache server are based on a fixed document catalog. As an example, the Independent Reference Model (IRM) further assumes that all requests are i.i.d., which allows to calculate theoretical estimates of the hit ratio, defined as the number of cache hits over number of requests.

However, in many applications, content is dynamic: It appears at a certain instant, its popularity varies over time and can eventually disappear. In this paper, we aim at building a tractable model that captures this feature (see [3] for details).

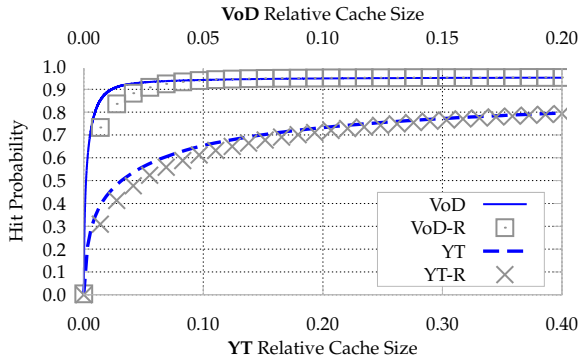

Figure 1: *Hit ratio of the original request sequence versus the result of semi-experiment (3).*

2. Semi-experimental driven modeling

We base our analysis on two datasets. The first dataset, called **YT**, comes from three months of YouTube traffic of the Orange Network in Tunisia. The second, called **VoD**, collects over three years requests from the Orange VoD service in France.

To discover the relevant features of our request sequences, we apply the semi-experimental methodology [2]. These semi-experiments aim to break specific structures of the request process by means of random shuffling; we then compare the resulting hit ratio, computed via simulation, with that of the original sequence and determine the importance of the broken structure. They consist in: 1) shuffling all requests, thus breaking all correlations; 2) shifting each document request sequence randomly, thus breaking the correlation between first requests; 3) fixing the first and last request of each document and shuffle all requests in between, thus breaking correlations within this sequence.

The results of semi-experiment (1) showed that the hit ratio curve differs considerably from that of the original sequence, thus confirming that IRM does not capture all the features of our datasets. In the case of (2) and (3), the curves do not differ considerably. This means that catalog arrivals and individual document request sequences can be modeled as homogeneous Poisson processes. For conciseness, we show in Figure 1 only the results of semi-experiment (3).

Taking into account the previous insights, we propose a two layered model. In the first layer, we consider a Poisson process Γ of rate γ, hereafter called the *catalog arrival process*. This process models the publications of documents in the catalog.

The second layer consists in the *document request processes*. Specifically, when a document d arrives to the catalog at time a_d, we model its request process as an homogeneous Poisson process of rate λ_d on the inter-

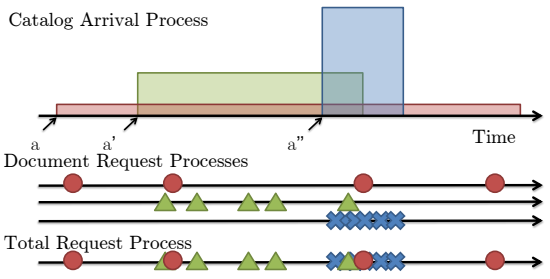

Figure 2: ***Top:*** *The boxes represent the lifespan and popularity.* ***Bottom:*** *Document request processes. Their superposition is the total request process.*

val $[a_d, a_d + \tau_d]$, where the random variable τ_d denotes the document lifespan. We further assume that the distribution of the pair (λ_d, τ_d) is stationary.
We consolidate the two layers into the *total request process*, which is the superposition of the document request processes. Figure 2 shows a schematic view of our model.

3. Hit ratio estimates and model validation

In order to estimate the hit ratio, we introduce the auxiliary process X_t, that counts the number of different requested objects up to time t. We then observe that for a LRU cache of size C, if an object is requested at time zero and later at time $\Theta > 0$, the latter request will be a hit if and only if $X_\Theta < C$. Equivalently, the request at time Θ will be a hit if and only if $\Theta < T_C$, where T_C is the first passage time of process X at level C.
We then invoke the so-called "Che approximation" [1], in which we assume that the random variable T_C is concentrated and thus can be approximated by a constant t_C.
To calculate the constant t_C, we notice that, by definition of T_C, we have the equality $X_{T_C} = C$. Thus, we impose that the constant t_C must satisfy this equality in mean, that is, $\mathbb{E}[X_{t_C}] = C$.
Process X is a non-homogeneous Poisson process. Define its mean function by $\Xi(t) = \mathbb{E}[X_t]$; as Ξ is an increasing function, t_C is then given by the formula $t_C = \Xi^{-1}(C)$. We make Ξ explicit in the following proposition.

Proposition 1 *The mean of the process X is given by*

$$\Xi(t) = \gamma\,\mathbb{E}\left[2t + \left(1 - e^{-\lambda t}\right)\left(\tau - t - \frac{2}{\lambda}\right)\mathbb{1}_{\{\tau \geq t\}}\right] + \gamma\,\mathbb{E}\left[2\tau + \left(1 - e^{-\lambda \tau}\right)\left(t - \tau - \frac{2}{\lambda}\right)\mathbb{1}_{\{\tau < t\}}\right]$$

where (λ, τ) is distributed as any (λ_d, τ_d).

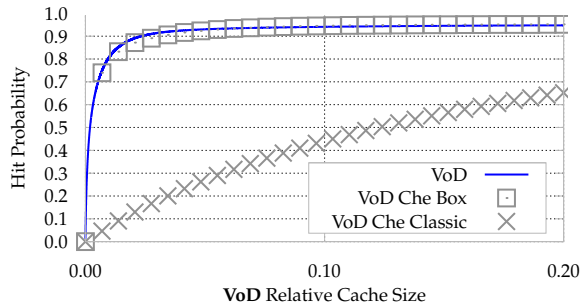

Figure 3: *Fitting for the Che estimation*

Finally, we write the expected number of hits for a document d in terms of its associated pair (λ_d, τ_d) and t_C:

Proposition 2 *Under the Che approximation, given pair (λ_d, τ_d), the expected number of hits $\overline{H}_d$ equals*

$$\overline{H}_d = \begin{cases} \lambda_d \tau_d - 1 + e^{-\lambda_d \tau_d} \\ \\ (\lambda_d \tau_d - 1)(1 - e^{-\lambda_d t_C}) + \lambda_d t_C e^{-\lambda_d t_C} \end{cases}$$

if $\tau_d < t_C$ and $\tau_d \geq t_C$, respectively.

For each document, unbiased estimates for λ and τ can be easily calculated; call these estimators $\widehat{\lambda}$ and $\widehat{\tau}$, respectively. Such estimators are valid only for documents with a number of requests $n \geq 2$; we cannot neglect, however, the documents that have only one request, since they compose a considerable part of the data. Due the fact that they add only misses, we can still incorporate them into the hit ratio as follows:

$$\mathrm{HR} = \frac{\mathbb{E}[H_d]}{\mathbb{E}[n_d]} = \frac{\mathbb{E}[H_d \,|\, n_d \geq 2]}{\mathbb{E}[n_d \,|\, n_d \geq 2] + \dfrac{\mathbb{P}(n_d = 1)}{\mathbb{P}(n_d \geq 2)}}.$$

Using the previous estimates, we compute t_C and $\mathbb{E}[H_d \,|\, n_d \geq 2]$ by plugging estimates $\widehat{\tau}$ and $\widehat{\lambda}$ into the formulas obtained in Proposition 1 and 2 and take averages.
We observe the result of this estimation on the VoD dataset in Figure 3, where we compare it to the actual hit ratio of the request sequence, obtained via simulation, and the estimation obtained via the "classic" Che approximation under the IRM setting. We observe that the hit ratio is noticeably underestimated in the latter case, whereas our model fits well the real hit ratio.

References

1. Hao Che, Ye Tung, and Z. Wang. Hierarchical web caching systems: modeling, design and experimental results. *Selected Areas in Communications, IEEE Journal on*, 20(7), 2002.
2. N. Hohn, D. Veitch, and P. Abry. Cluster processes, a natural language for network traffic. *IEEE Transactions on Signal Processing, special issue "Signal Processing in Networking"*, 51(8), Aug. 2003.
3. Felipe Olmos, Bruno Kauffmann, Alain Simonian, and Yannick Carlinet. Catalog dynamics: Impact of content publishing and perishing on the performance of a LRU cache. In *26th Internation Teletraffic Congress*. IEEE Communications Society, 2014. To appear.