Approximate models for cache analysis with correlated requests

Nicaise E. Choungmo Fofack

Orange Labs
38/40 rue du Général Leclerc, 92130
Issy-Les-Moulineaux, France
nicaise.choungmofofack@orange.com

1. Introduction

Content distribution and in-network caching have emerged as key features of several network architectures to accommodate the current content usage patterns (Video on Demand, User-Generated Contents, Web browsing, the recent Dynamic Page Caching concept of Akamaï, etc.) while reducing congestion and improving access speed as networks increase in size. In order to guarantee the latter performance, network designers and engineers need tools to quickly gain insights on the behaviour of the multi-cache systems that arise from their content-oriented architectures when deploying caches across the networks.

The analysis of these cache networks is significantly challenging due to their irregular topologies, the heterogeneity of their nodes (different replacement policies and capacities), and the statistical correlation of requests. Few modelling attempts have been proposed to solve this problem on general and heterogeneous Time-To-Live (TTL)-based cache networks where requests are described by Markov-renewal processes [1] or Markov-arrival processes (MAPs) [2]. However, the latter derivations suffer from a large complexity which limits their practical interest.

In this work, we aim at presenting an analytical methodology to approximate the performance metrics of heterogeneous networks where (i) caches are running the Least Recently Used (LRU), First-In First-Out (FIFO), or Random (RND) replacement algorithms, (ii) all requests are routed as a feed-forward (or hierarchical) network, (iii) streams of requests are described by simple MAPs [3] as briefly explained in Section 2.1.

2. Single cache approximation (SCA)

We consider a cache system that can accommodate C files requested from a catalog of size K ($K > C$). We assume that the cache is connected to a server where all files are permanently stored such that requests arrived first on the cache and the missed ones are forwarded to the origin server. We also assume a *zero delay* cache-server system i.e. request processing/forwarding and file downloading times are negligible in comparison to the inter-request times [1].
In this section we present a simplified traffic model which accounts for request correlations, then we recall the TTL-based model of LRU and FIFO/RND caches, and we provide the approximate metrics of interest (hit and occupancy ratios).

2.1. Workload model

Let us denote by $\{\mathcal{R}_k, k = 1, \ldots, K\}$ the general point process describing the sequence of requests for file k on the cache. Here, we aim at providing a simple and accurate approximation of $\mathcal{R}_k$ based on a *minimal available information* that engineers could easily measure or estimate at edge nodes of the network. As inputs of our process model, we choose the *request rate*, the *squared coefficient of variation* (scv), the *skewness* factor, and the *lag-1* (and possibly *lag-2*) autocorrolation. These quantities are general enough to capture the main statistical parameters (per-file popularity, temporal locality, and correlation coefficients) of the arrival process $\mathcal{R}_k$ with a MAP (Cf. the *MAP-Match* procedure in [3]). Therefore, all request streams in this paper will be approximated by the switched MAP$(\mathbf{D}_0, \mathbf{D}_1)$ described in [3, Sect.4].

2.2. Cache model

We recall a result from [1] that the characteristic time (i.e. the maximum inter-request time of a given file that yields a cache hit) converges in distribution (under some conditions that hold in practice) to deterministic and exponentially distributed random variables as $K \uparrow \infty$ for LRU/FIFO and RND caches respectively under general stationary request processes. Hence LRU, FIFO and RND caches may be analysed through their corresponding TTL-based models. In this work, we consider that TTLs are i.i.d with an Erlang distribution $E_r(m, \mu)$ (or a PH-distribution PH$(\alpha, \mathbf{T})$ with m transient states). For RND and LRU/FIFO caches we set $m = 1$ and $m \approx 10 \gg 1$ (sufficiently large to approximate a constant) respectively.

2.3. Performance metrics

In this section, we focus on the calculation of the file hit $\{H_k\}_{k=1}^K$ (resp. occupancy $\{O_k\}_{k=1}^K$) probabilities

defined as the stationary probability that a request for file k yields a cache hit (resp. the stationary probability that file k is in the cache at any time). To do so, we should find the rate μ of the TTL distribution. This is done by solving the following equation $\sum_{k=1}^{K} O_k = C$. Based on the exact results derived in [1, 2] we are able to derive the closed-form expressions of H_k (see [2, Lemma 5]) and O_k (see [1, Prop. 2.9]) needed in the latter equation.

3. Networks with feed-forward routing

In this section, we generalize previous results to cache networks where requests flow in the same direction (i.e. a cache does not query a cache from which it receives missed requests) by describing the main network traffic transformations in cache network deployments where requests are correlated and routed on a unique feed-forward network.

3.1. Miss process characterization

The SCA extends easily to tandem or **linear cache networks**. Indeed, one need to characterize the miss process of each cache, then apply the *MAPMatch* procedure on the miss streams, and calculate the metrics of interest as done in Section 2. If request processes are MAPs, so are miss processes of PH-distributed TTL models [2, Thms 3 & 4]. However, the complexity of the exact miss process characterization [2] is significantly reduced with our workload model.

3.2. Multiple sources

A cache may receive requests from several (not necessarily independent) sources such as caches or users. This is the common case for **tree networks**. The exact characterization of the aggregation of N MAPs requires a strong independence assumption among request streams being superposed, an important calculation to evaluate the Kronecker sum, and a huge memory to store the resulting MAP (limitations of the exact method in [2]). In this work, we propose to calculate the parameters needed for MAP matching [3] as weighted sum of corresponding parameters of MAP components (e.g the resulting scv is $c_v^2 = \sum_{n=1}^{N} \frac{\lambda_n}{\Lambda} c_n^2$ where λ_n and c_n^2 are the rate and the scv of the n-th MAP being aggregated, $\Lambda = \sum_{n=1}^{N} \lambda_n$ is the total request rate). Using the MAP matched of the overall process and the SCA, we obtain the performance metrics.

3.3. Several destinations

In many situations, a cache may select the next hop among two or more caches to forward its missed requests. The latter task may be performed based on some available informations (e.g. fault tolerance, multiple shortest paths, load balancing, etc.). We denote by r_j the probability that the outgoing-link j is chosen for request forwarding such that $\sum_j r_j = 1$ (e.g. r_j may be the fraction of time the outgoing-link j is up while others are down). If the cache miss processes are MAPs, so are the request processes on each link j (see [2, Lemma 4]). Finally, the MAPMatch procedure [3] is applied on each r_j-thinned request process.

4. Conclusion

In this paper, we proposed a methodology that can be used to quickly estimate the performance metrics of single TTL-based (e.g. LRU, FIFO, RND) cache when request streams are approximated by special Markov Arrival Processes. We have also explained how our modelling attempt extends heterogeneous cache networks where all requests are routed on a unique/same feed-forward network (e.g. linear, tree, polytree) built on top of the network topology. Ongoing work are devoted to provide a detailed description of the latter extensions, to perform an extensive evaluation of our models, and to investigate the case of general and heterogeneous caches networks with arbitrary routing.

References

1. Choungmo Fofack (Nicaise). – On models for performance analysis of a core cache network and power save of a wireless access network. –Ph.D. thesis, Univ. of Nice Sophia Antipolis, `http://tel.archives-ouvertes.fr/tel-00968894`, Feb. 2014.
2. Berger (D. S.), Gland (Philipp), Singla (Sahil) and Ciucu (Florin). – Exact Analysis of TTL Cache Networks : The Case of Caching Policies driven by Stopping Times. – Preprint ArXiv (CoRR, abs/1402.5987), `http://arxiv.org/abs/1402.5987`, 2014.
3. Horváth (Gabor). – Matching marginal moments and lag autocorrelations with MAPs. – In Proc. ValueTools'13, Torino, Italy, Dec. 2013.