A Versatile and Accurate Approximation for LRU Cache Performance

Christine Fricker, Philippe Robert, James Roberts
INRIA, France
{Christine.Fricker,Philippe.Robert,James.Roberts}@inria.fr

Abstract—In a 2002 paper, Che and co-authors proposed a simple approach for estimating the hit rates of a cache operating the least recently used (LRU) replacement policy. The approximation proves remarkably accurate and is applicable to quite general distributions of object popularity. This paper provides a mathematical explanation for the success of the approximation, notably in configurations where the intuitive arguments of Che *et al.* clearly do not apply. The approximation is particularly useful in evaluating the performance of current proposals for an information centric network where other approaches fail due to the very large populations of cacheable objects to be taken into account and to their complex popularity law, resulting from the mix of different content types and the filtering effect induced by the lower layers in a cache hierarchy.

I. INTRODUCTION

The investigation of so-called information-centric networking (ICN) architectures is bringing renewed interest in the performance of caching. It is particularly important to understand the potential for trading off bandwidth for memory by implementing a network of caches and to develop tools that enable the optimization of such a network. The ICN application places particularly stringent requirements on evaluation tools since the population of content items available via the Internet is immense and caches are required to store content of diverse types, each type being distinguished by its peculiar popularity characteristics.

In recent work on cache performance in the context of ICN [7], we applied a tool from the literature that was particularly well adapted to requirements. This is an approximation for evaluating the hit rates of a cache under the least recently used (LRU) replacement policy proposed by Che, Tung and Wang in a 2002 paper [3]. The "Che approximation" proved extremely accurate, even in conditions where the authors' intuitive arguments were clearly not justified. The objective of the present paper is to provide more rigorous mathematical arguments allowing the scope of the approximation to be more clearly defined.

The Che approximation applies to the following model. Users request items from a population of N objects, first testing to see if the object is present in a cache of capacity C. If the object is present it is returned to the user. If not, it is obtained from some other source and copied to the cache as it is returned to the user. This object replaces the one that was least recently requested. The probability a request is for object n, for $1 \le n \le N$, is proportional to some popularity $q(n)$, independently of all past requests.

The hit rate $h(n)$ for object n, i.e., the probability this object is present in the cache, is approximated by

$$h(n) \approx 1 - e^{-q(n)t_C}$$

where t_C is the unique root of the equation

$$\sum_{n=1}^{N}(1 - e^{-q(n)t}) = C.$$

Related work

There is clearly a huge amount of related work on the performance of caching. We restrict ourselves here to a discussion of the papers most relevant to our work. Cited papers [4] and [12] may be consulted for summaries of significant earlier work.

Dan and Towsley [4] derived an iterative algorithm for calculating approximate hit rates for a cache of size C using the hit rates for a cache of size $C - 1$. Complexity is $O(CN)$ which can be prohibitive in ICN applications where N and C are very large. Dan and Towsley also propose a scheme comparable in complexity to the Che approximation for computing hit rates under FIFO replacement (i.e., the object replaced is the one that has been in the cache the longest). Recent work by Rosenweig *et al.* applies the LRU algorithm of [4] to analyse general networks of caches [16].

Jelenkovic provides closed-form asymptotic hit rate estimates for particular choices for popularities $q(n)$ [12]. These are namely, a generic light-tailed law, $q(n) = e^{-\lambda n^{\beta}}$ for $n > 0$, where λ and β are positive constants, and a generalized Zipf law, $q(n) = 1/n^{\alpha}$ for $n \ge 0$, with $\alpha > 1$. The latter law with $\alpha < 1$ and $\alpha = 1$ is considered by Jelenkovic *et al.* in [11] where it is shown that hit rates can be expressed in terms of a parameter defined as the root of a certain equation. The main disadvantage is that derived formulas are only applicable to particular popularity laws.

In the following, we first discuss the considered traffic model in Section II before presenting the Che approximation in detail in Section III. Reasons for its remarkable accuracy are elucidated in Section IV while Section V derives explicit results for the special case of Zipf popularity. An approximation for random replacement similar to the Che approximation is derived in Section VI and used in Section VII in an ICN application.

II. Traffic Model

We recall the independent reference model (IRM), discuss the nature of the popularity law $q(\cdot)$ and argue that the IRM is appropriate for modelling an information-centric network.

A. The independent reference model

Requests for objects occur in an infinite sequence where the object indexes required on the ith request, for $i > 0$, are independent random variables on $\{1, 2, \ldots, N\}$ with a common probability distribution. Specifically, the probability the required object has index n is proportional to $q(n)$, for $1 \leq n \leq N$. We refer to $q(\cdot)$ as the *popularity law*.

An alternative description of this stochastic is as follows (see Fill and Holst [6], for example). Let (τ_n) be a sequence of independent exponential random variables with respective rates $(q(n))$. At any arbitrary instant, the next object to be requested is n_0 where n_0 is the index such that τ_{n_0} is the minimum of the (τ_n).

B. Popularity laws

Cache performance depends crucially on the popularity law $q(\cdot)$. It is usual to order objects in order of decreasing popularity such that $q(1) \geq q(2) \geq \ldots \geq q(N)$. With this convention, the most frequently observed popularity law is a generalized Zipf law: $q(n) = 1/n^\alpha$ with $\alpha > 0$.

Examples of content types with reported Zipf law behaviour are web pages [1], [14], files shared using BitTorrent [7], YouTube documents [9], [2], video on demand movies [19]. The reason why content popularity follows the Zipf law remains unclear though the discussion by Mitzenmacher on generative models for power laws provides some possible explanations [15]. The estimated value of α is often around 0.8 though cases with $\alpha > 1$ have been observed.

In fact, agreement between observations and the Zipf law is not always entirely convincing. Sometimes the popularity law has a lighter tail where the least popular objects are very unlikely to be requested. As a simple example of a light-tailed law, we consider geometric popularity: $q(n) = \rho^n$. This choice is not based on any measurement results but is made rather with the intention of stressing the Che approximation: the approximation is in fact exact for a uniform probability law, $q(n) =$ constant; it is more likely to fail as the law becomes more accentuated, as with the geometric law.

While a Zipf law with suitable α might be a reasonable representation for a homogeneous set of content, caches in the Internet must be designed for a traffic mix. The popularity law should reflect this mix by weighting single type popularities by the proportion of requests due to that type. It is necessary also to account for significant differences in the size of objects of different types (see Section III-B).

In a network, there is typically a hierarchy where higher layer caches receive requests only for objects that are not found in lower layers. The popularity law is thus a filtered version of the initial law. Figure 1 shows how the popularity law at a second layer cache is deformed by the first layer.
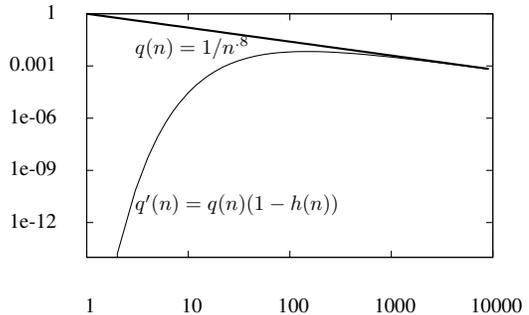

Fig. 1: A filtered popularity law: $q'(\cdot)$ is the popularity law for requests overflowing a size 1000 cache when $q(\cdot)$ is Zipf(0.8) and $N = 10000$.

Fortunately, the Che approximation is sufficiently versatile to account for both composite popularity laws and filtering [7]. Note that it is not necessary to order objects in decreasing order of popularity or to normalize the $q(n)$ and we do not make these assumptions in the following analysis.

C. Validity of the independent reference model

The independent reference model is a convenient abstraction that allows analytical modelling where a more accurate representation of reality would be intractable. It is important to understand the limitations of this model.

Adopting the IRM is to assume popularity does not change. This is clearly not true for content where not only the popularity of a given object, but also the catalogue of available objects changes over time. This is a manifestation of *temporal locality*. The IRM may still be considered acceptable if popularity variations are slow compared to the time scale of cache churn. Non-stationarity has a greater impact on the accuracy of measurements of popularity. For instance, statistics gathered over a period of weeks will hardly be representative of the popularity of a catch-up TV show that is only on-line for a few days.

A second cause of error in predicting popularity laws is *spatial locality*. Most types of content will have strong regional bias with, for instance, a network in France observing high popularity for movies in French. However, this does not so much invalidate the IRM assumption as require more care in specifying popularity laws.

The IRM is reasonable when content requests are generated independently by a large population of users. This is not the case, however, for requests seen by caches at second and higher layers in a hierarchy. The request process overflowing from lower layer caches is correlated. The IRM is nevertheless reasonable if a higher layer cache receives the aggregation of independent, low intensity overflows from many low layer cache instances. Moreover, even for the simple tandem of two caches considered by Jenekovic and Kang [13], the impact of correlation on hit rates was shown to be slight.

We conclude from the above that the IRM is a reasonable basis for evaluating cache performance, as long as care is taken

in specifying the popularity law.

III. THE CHE APPROXIMATION

We present the Che approximation and demonstrate its accuracy in comparison to the results of simulation.

A. A characteristic time

Consider a cache with capacity for C objects under the LRU replacement policy. We introduce the following random variables, for $t \geq 0$,

$$X_n(t) = \sum_{i=1, i \neq n}^{N} \mathbb{1}_{\{\tau_i < t\}}$$

and

$$T_C(n) = \inf\{t > 0 : X_n(t) = C\}.$$

$X_n(t)$ is the number of different objects requested up to time t, excluding object n, and $T_C(n)$ is the time at which exactly C different objects, other than n, have been requested.

Without loss of generality, suppose a request for object n occurs at time 0. The next request for object n will be a hit if fewer than C other objects are requested in $(0, \tau_n)$ (recall that τ_i is the generic, exponentially distributed inter-request interval for object i). In other words, there is a hit if $X_n(\tau_n) < C$. Now, $\{X_n(\tau_n) < C\} = \{T_C(n) > \tau_n\}$ so that,

$$h(n) = \mathbb{P}(T_C(n) > \tau_n) = \mathbb{E}\left(1 - e^{-q(n)T_C(n)}\right).$$

Since at $T_C(n)$ there are exactly C objects in the cache, we have

$$C = \sum_{i=1, i \neq n}^{N} \mathbb{1}_{\{\tau_i < T_C(n)\}},$$

and taking expectations,

$$C = \sum_{i=1, i \neq n}^{N} \mathbb{E}\left(1 - e^{-q(i)T_C(n)}\right).$$

A first approximation of Che *et al.* [3] is to assume for large C that the $T_C(n)$ are nearly deterministic. They replace random variable $T_C(n)$ by the constant $t_C(n)$ that solves

$$C = \sum_{i=1, i \neq n}^{N} \left(1 - e^{-q(i)t}\right)$$

and approximate the hit rates by

$$h(n) = 1 - e^{-q(n)t_C(n)}.$$

A second approximation in [3] is to assume $t_C(n) = t_C$ for $1 \leq n \leq N$ where t_C solves $C = \sum_{1 \leq i \leq N}(1 - e^{-q(i)t})$. This is arguably reasonable when individual popularities $q(n)$ are small relative to the sum $\sum_n q(n)$. Having verified numerically that this second approximation is generally very accurate, we adopt it for the remainder of the paper. This simplifies notation but it should be noted that our analysis could readily be adapted to preserve the dependence on n.

In summary, the Che approximation considered in this paper is as follows. Let t_C be the unique root of the equation

$$C = \sum_{i=1}^{N} \left(1 - e^{-q(i)t}\right). \tag{1}$$

The hit rate $h(n)$ for object n, for $1 \leq n \leq N$, is then

$$h(n) = 1 - e^{-q(n)t_C}. \tag{2}$$

Che *et al.* refer to t_C as the "characteristic time" of the cache.

Random variables $X(t)$ and T_C are defined as above without excluding object n. In particular, $X(t) = \sum_{1 \leq i \leq N} \mathbb{1}_{\{\tau_i < t\}}$. Since the Bernoulli events in the summation are independent, we readily derive the mean and variance of $X(t)$:

$$m(t) = \sum_{i=1}^{N}(1 - e^{-q(i)t}), \tag{3}$$

$$\sigma(t)^2 = \sum_{i=1}^{N} e^{-q(i)t}(1 - e^{-q(i)t}). \tag{4}$$

B. Variable sized objects

So far we have assumed the cache capacity is measured in objects. In reality content objects have different sizes and cache capacity is more reasonably measured in bytes. Suppose object n has size $\theta(n)$. Since the cache is intended to store a very large number of objects, $\theta(n) \ll C$ so that we can reasonably ignore boundary effects and adapt the Che approximation by replacing (1) with

$$C = \sum_{i=1}^{N} \left(1 - e^{-q(i)t}\right) \theta(i). \tag{5}$$

The hit rates are still given by (2).

An alternative way to account for variable size objects is to assume they are divided into constant sized chunks. This is the principle of content-oriented Internet architectures like CCN [10]. Let $\theta(n)$ be given in chunks and assume all chunks inherit the popularity $q(n)$ of their parent object. Applying the Che approximation to chunks, it is easy to see that equation (1) for t_C is then precisely the same as (5).

For some types of objects, it may be that chunks of the same object have different popularities (e.g., the first chunks of a video will be viewed more often than the last chunks). It makes more sense in this case to directly postulate a popularity law for chunks rather than objects.

We conclude that the Che approximation presented in Section III-A is appropriate also for variable size objects. It is not necessary to complicate the analysis in the next sections by introducing the size $\theta(n)$ (although to use (5) might be useful in practice as in [7]).
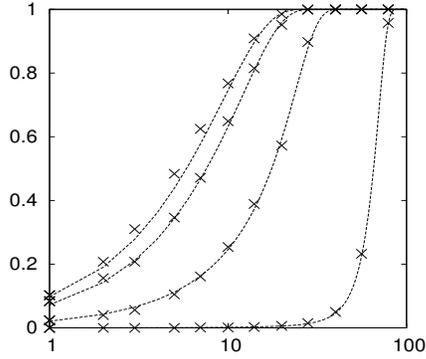
C. Accuracy

The accuracy of the approximation is typified by the results shown in Figure 2. Figures 2a and 2b plot the hit rates of objects ranked 1, 10, 100 and 1000 from a population of
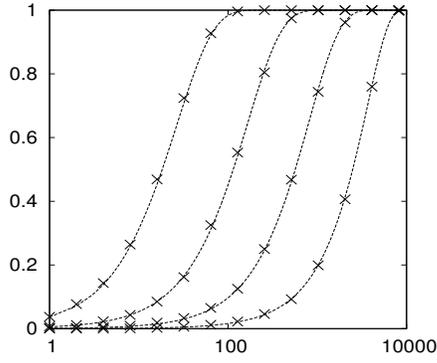
(a) $N = 10^4$, Zipf(0.8), objects 1, 10, 100, 1000
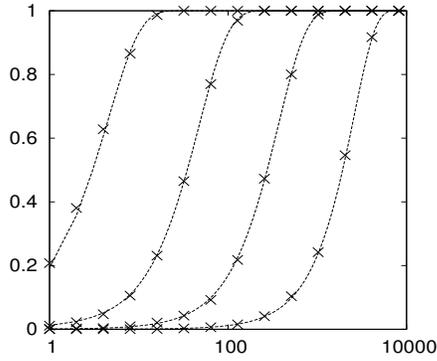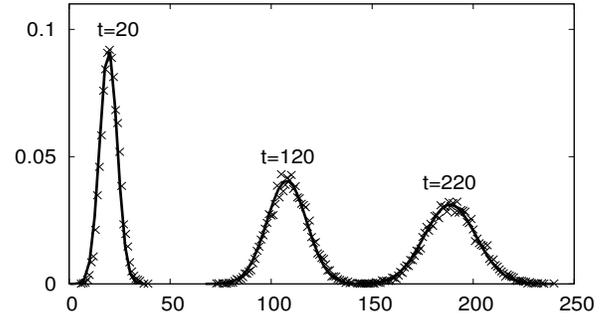

(b) $N = 10^4$, Zipf(1.2), objects 1, 10, 100, 1000
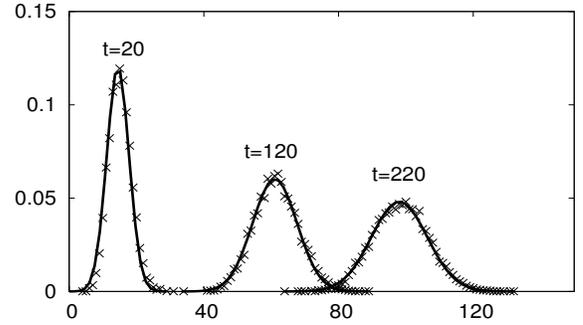

(c) $N = 100$, Geo(0.9), objects 1, 4, 16, 64

Fig. 2: Hit rate against cache size for selected objects, LRU replacement

(a) Zipf(0.8), $N = 10000$

(b) Zipf(1.2), $N=10000$

(c) Geo(0.9), $N=100$

Fig. 3: Distribution of $X(t)$, simulation and Gaussian approximation

10000, assuming Zipf popularity with $\alpha = 0.8$ and $\alpha = 1.2$, respectively. The crosses are the results of simulations with sufficiently long runs to ensure their high accuracy. The lines are derived from the Che approximation. Agreement is perfect, for all practical purposes.
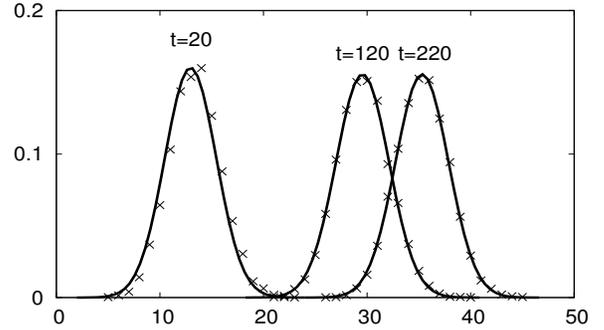
Figure 2c confirms the approach is accurate also for a case where the intuitive arguments of Che *et al.* do not apply. The population is only 100 and the popularity law is geometric with parameter $\rho = 0.9$. The figure plots the hit rates of objects ranked 1, 4, 16 and 64. Discrepancies are visible only for object 1 and these are very slight.

IV. WHY THE APPROXIMATION WORKS

We show that the Che approximation works so well essentially because $X(t)$ is approximately Gaussian, even for a small object population and a small cache size.

A. A Gaussian approximation for $X(t)$

Rather than attempting to study T_C directly, we consider $X(t)$ and exploit the elementary relation $\mathbb{P}(T_C > t) = \mathbb{P}(X(t) < C)$, for $t \geq 0$. Since the variable $X(t)$ is a sum of independent random variables, it is natural to expect its distribution to be approximately Gaussian.

Figure 3 plots the distribution of $X(t)$ for some values of t when the popularity distribution $q(n)$ is: 3a) Zipf(.8) for

$1 \leq n \leq 10^4$; 3b) Zipf(1.2) for $1 \leq n \leq 10^4$; and 3c) Geo(.9) for $1 \leq n \leq 100$. The figures plot simulation results as crosses with a superposed normal distribution of mean and variance given by (3) and (4), respectively. The figures clearly suggest that $X(t)$ is indeed Gaussian for the range of popularity laws and times t of interest.

B. A central limit theorem

The following proposition establishes conditions under which a Gaussian approximation for $X(t)$ is reasonable.

Proposition 1. *If* $Z(t) := (X(t) - m(t)) / \sigma(t)$, *where* $m(t)$ *and* $\sigma(t)$ *are given by (3) and (4), respectively, then*

$$\|\mathcal{L}(Z(t)) - \mathcal{L}(\mathcal{G})\| := \sup_{x \in \mathbb{R}} |\mathbb{P}(Z(t) \leq x) - \mathbb{P}(\mathcal{G} \leq x)|$$
$$\leq \frac{K}{\sigma(t)},$$

where $\mathcal{G}$ *is a centered normal random variable and* $K \leq 0.56$.

Proof: Let $Z_n(t) = [\mathbb{1}_{\{\tau_n \leq t\}} - (1 - \exp(-q(n)t))]/\sigma(t)$ and

$$Z(t) := \frac{X(t) - m(t)}{\sigma(t)} = \sum_{n=1}^{N} Z_n(t).$$

Then $\mathbb{E}(Z_n(t)) = 0$, $\mathbb{E}(Z(t)^2) = 1$ and we have,

$$\sigma(t)^3 \mathbb{E}\left(|Z_n(t)|^3\right) = e^{-3q(n)t}(1 - e^{-q(n)t}) + (1 - e^{-q(n)t})^3 e^{-q(n)t}$$
$$\leq e^{-q(n)t}(1 - e^{-q(n)t}).$$

Hence,

$$\sum_{n=1}^{N} \mathbb{E}\left(|Z_n(t)|^3\right) \leq \frac{1}{\sigma(t)^3} \sum_{n=1}^{N} e^{-q(n)t}(1 - e^{-q(n)t}) = \frac{1}{\sigma(t)}.$$

Berry-Esseen's Inequality, see Feller [5, p. 544], gives the relation

$$\|\mathcal{L}(W(t)) - \mathcal{L}(\mathcal{G})\| \leq K \sum_{n=1}^{N} \mathbb{E}(|Z_n(t)|^3) \leq \frac{K}{\sigma(t)},$$

where K is a constant. It has recently been shown that the value of K is no greater than 0.56 [17]. ■

This proposition confirms that $X(t)$ is asymptotically Gaussian as $t \to \infty$ if $\sigma(t)$ grows unboundedly. It is not wholly satisfactory, however, in that it does not explain the excellent fit illustrated in Figure 3 for small values of t and for geometric popularity where $\sigma(t)$ is not an increasing function. This appears to be a normal situation for central limit theorems where convergence to the normal distribution is often much better than predicted by the analytical bounds.

C. Approximating the hit rates

Starting with the Gaussian approximation for the distribution of $X(t)$, we argue that the Che approximation is indeed generally applicable.
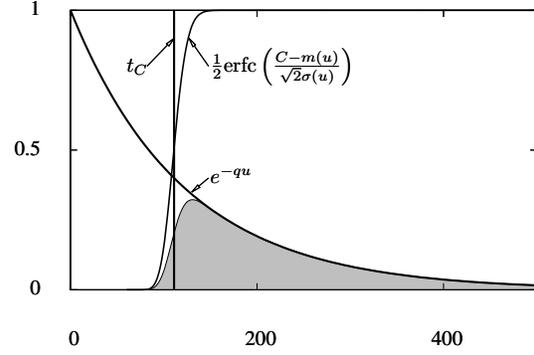

Fig. 4: Approximating the integral

Proposition 2. *Assuming* $X(t)$ *is Gaussian of mean* $m(t)$ *and standard deviation* $\sigma(t)$, *the hit rate for object* n *may be written,*

$$h(n) = 1 - \frac{1}{2} \int_0^{+\infty} \text{erfc}\left(\frac{C - m(u)}{\sqrt{2}\sigma(u)}\right) q(n)e^{-q(n)u}\, du \quad (6)$$

where $\text{erfc}(x)$ *is the complementary error function.*

Proof: Recall that $h(n) = 1 - \mathbb{E}\left(e^{-q(n)T_C}\right)$. Consider $\mathbb{E}\left(e^{-qT_C}\right)$ for some $q > 0$. By definition of T_C,

$$\mathbb{E}\left(e^{-qT_C}\right) = \int_0^{+\infty} \mathbb{P}(T_C < u)qe^{-qu}\, du$$
$$= \int_0^{+\infty} \mathbb{P}(X(u) > C)qe^{-qu}\, du$$

By the assumption that $X(t)$ is Gaussian, we can write

$$\mathbb{P}(X(u) > C) = \frac{1}{2}\text{erfc}\left(\frac{C - m(u)}{\sqrt{2}\sigma(u)}\right)$$

and the proposition follows. ■

Proposition 2 could be used directly to evaluate the hit rates. Instead, we derive the Che approximation as an approximation for the integral in (6). Note that the complementary error function in the integrand is an S-shaped function tending rapidly to asymptotes at 0 and 1 from a point of inflection at $m(u) = C$, i.e., at $u = t_C$. We therefore replace this function in the integral by the step function $\mathbb{1}_{\{m(u)>C\}}$ yielding the following approximation

$$\mathbb{E}\left(e^{-qT_C}\right) \approx \int_0^{+\infty} \mathbb{1}_{\{m(u)>C\}}qe^{-qu}\, du$$
$$= \int_0^{+\infty} \mathbb{1}_{\{u>t_C\}}qe^{-qu}\, du$$
$$= e^{-qt_C}.$$

The second step follows from $m(t_C) = C$ and the fact that $m(u)$ is increasing in u. This establishes the validity of the Che approximation on condition that replacing erfc by a step function is accurate.

This accuracy is illustrated in Figure 4 for a particular set of parameters: Zipf(0.8) popularity, $N = 10000$, $C = 100$ and $q = .83 \times 10^{-3}$. The exact integral is the shaded area. The approximation replaces this by the area to the right of t_C and under the exponential curve. Visibly, the approximation is good in this case. It is easy to convince oneself that this is generally true for the popularity laws of interest. However, it unfortunately does not seem possible to quantify the error due to the non-explicit nature of the erfc function argument in Proposition 2.

V. Zipf popularity and a large cache

For Zipf law popularity, we can prove the asymptotic validity of the Che approximation and characterize t_C directly, without solving equation (1).

A. Preliminary results

We first state two lemmas about the moments of $X(t)$.

Lemma 1. *For $t \geq 0$, the variance of $X(t)$ can be expressed in terms of its mean:*

$$\sigma(t)^2 = m(2t) - m(t).$$

The proof is a straightforward application of (3) and ((4).

Now consider the behaviour of the average of $(X(\cdot))$ at an appropriate time scale.

Lemma 2. *With $q(n) = 1/n^\alpha$ for $1 \leq n \leq N$, for any $\beta > 0$,*

$$m(\beta N^\alpha) = \psi_\alpha(\beta)N + o(N),$$

where

$$\psi_\alpha(\beta) = 1 - \int_0^1 e^{-\beta/x^\alpha}\, dx.$$

Proof:

$$m(\beta N^\alpha) = N \times \frac{1}{N}\sum_{n=1}^N \left(1 - e^{-\beta N^\alpha/n^\alpha}\right)$$
$$= N\left(\int_0^1 1 - e^{-\beta/x^\alpha}\, dx\right) + o(N).$$

∎

B. A Gaussian approximation for T_C

In the particular case of Zipf law popularity we show that T_C is asymptotically Gaussian and derive t_C as its expectation. In the following $\lfloor z \rfloor$ denotes the integer part of $z \geq 0$.

Proposition 3. *For Zipf(α) popularity, as N and C tend to infinity with $C = \lfloor \delta N \rfloor$, for $0 < \delta < 1$, we have*

$$t_{\lfloor \delta N \rfloor} = \psi_\alpha^{-1}(\delta)N^\alpha + o(N^\alpha), \qquad (7)$$

where $\psi_\alpha(\beta)$ is defined in Lemma 2.

Furthermore, the random variable

$$\frac{\psi_\alpha'(\psi_\alpha^{-1}(\delta))}{\sqrt{\psi_\alpha(2\psi_\alpha^{-1}(\delta)) - \delta}}\frac{(T_{\lfloor \delta N \rfloor} - t_{\lfloor \delta N \rfloor})}{N^{\alpha-1/2}}$$

converges in distribution to a centred Gaussian random variable.

Proof: The asymptotic relation for t_C is a direct consequence of Lemma 2. We have,

$$\mathbb{E}(X(\beta N^\alpha)) = \psi_\alpha(\beta)N + o(N)$$

and, by definition of t_C,

$$E(X(t_C)) = \lfloor \delta N \rfloor = \delta N + o(N).$$

Let $s_C = \psi_\alpha^{-1}(\delta)N^\alpha$ so that

$$\mathbb{E}(X(s_C)) = \psi_\alpha(\psi_\alpha^{-1}(\delta))N + o(N) = \delta N + o(N).$$

It follows that $t_C = s_C + o(N^\alpha)$ and the first statement of the proposition is proved.

Now consider the following relation, for $x \in \mathbb{R}$,

$$\mathbb{P}(T_C - t_C \geq x) = \mathbb{P}(X(t_C + x) < C) =$$
$$\mathbb{P}\left[\frac{X(t_C+x)-m(t_C+x)}{\sigma(t_C+x)} < \frac{m(t_C)-m(t_C+x)}{\sqrt{m(2(t_C+x))-m(t_C+x)}}\right].$$

The second equality follows on setting $m(t_C) = C$ and applying Lemma 1.

From lemmas 1 and 2 we know that $\sigma(\beta N^\alpha)$ goes to infinity as $N \to \infty$ for Zipf laws. Proposition 1 thus shows that the variable

$$\frac{X(\beta N^\alpha) - m(\beta N^\alpha)}{\sigma(\beta N^\alpha)}$$

is arbitrarily close to a centred normal variable if N is sufficiently large. The only thing to check is the asymptotic behaviour of the fraction

$$\frac{m(t_C) - m(t_C + x)}{\sqrt{m(2(t_C + x)) - m(t_C + x)}}$$

when $x = \beta N^{\alpha-1/2}$.

Using Lemma 2, write $m(t_C + x)$ as

$$m(t_C + x) = m\left(N^\alpha(\psi_\alpha^{-1}(\delta) + \frac{x}{N^\alpha} + o(1))\right)$$
$$= N\psi_\alpha\left(\psi_\alpha^{-1}(\delta) + \frac{x}{N^\alpha}\right) + o(N).$$

Expanding ψ_α about $\psi_\alpha^{-1}(\delta)$ yields the numerator

$$m(t_C) - m(t_C + x) = -\psi_\alpha'(\psi_\alpha^{-1})(\delta)N^{1-\alpha}x + o(N).$$

The denominator can similarly be written

$$\sqrt{m(2(t_C + x)) - m(t_C + x)} = \sqrt{\psi_\alpha(2\psi_\alpha^{-1}(\delta))N - \delta N + o(N)}.$$

The second statement of the proposition follows on substituting for x.

∎

Remark. The expression for $t_{\lfloor \delta N \rfloor}$ in Proposition 3 coincides with an equivalent quantity derived differently in Theorems 2 and 3 of Jelenkovic *et al.* [11] for $\alpha = 1$ and $\alpha < 1$, respectively. Theorem 1 of the same paper provides an explicit expression for t_C when $\alpha > 1$ and N is infinite. This expression proves significantly less precise than (7) when α is not much greater than 1 (1.2, say), even for N as large as 10000.

C. The Che approximation

Proposition 3 shows that there is a function $\Theta(N)$ such that, as $N \to \infty$, $\Theta(N)/t_C \to 0$ while $(T_C - t_C)/\Theta(N)$ converges to a Gaussian random variable. Thus in this case, T_C does indeed become deterministic (i.e., $T_C/t_C \sim 1$) and the original argument of Che *et al.* applies. We have $\mathbb{E}(e^{-qT_C}) \to e^{-qt_C}$ as C (and $N) \to \infty$.

D. Geometric popularity

It can be shown for geometric popularity, $q(n) = \rho^n$ for $n \geq 0$, that $m(t) = -\log t / \log \rho + O(1)$. Thus, by Lemma 1, $\sigma(t)^2 = \log 2 / \log \rho + O(1)$, i.e., the variance of $X(t)$ is asymptotically constant and small compared to $m(t)$. This explains why the Che approximation works for geometric popularity (applying the arguments in Section IV) although T_C is by no means deterministic.

VI. A "CHE APPROXIMATION" FOR RANDOM REPLACEMENT

The excellent accuracy of the Che approximation for LRU replacement motivates the search for a similar approach for other policies. In this section we consider random replacement and derive an approximation that is similar in accuracy and complexity to the Che approximation.

Random might be preferred to LRU because it is simpler to implement. When a new object is to be added to the cache, it overwrites a randomly chosen existing object independently of the popularity of the latter. This policy was shown by Gelenbe to have exactly the same hit rates as FIFO [8].

The exact analysis of Gelenbe is too complex for practical evaluation. A recent paper by Simonian *et al.* [18] derives some general hit rate expressions that are notably exploited to provide large cache asymptotics applicable for a Zipf popularity law with $\alpha > 1$. Dan and Towsley [4] propose an approximate evaluation for the hit rates of a FIFO cache.

Note that $h(n)$, the hit rate for object n, is the probability object n is in the cache at an arbitrary instant. This can be expressed by Little's formula as the product $\lambda(n) \times T(n)$ where $\lambda(n)$ is the frequency at which object n enters the cache and $T(n)$ is its average sojourn time.

Clearly, $\lambda(n) = (1 - h(n))q(n)$. We assume $T(n)$ is inversely proportional to the arrival rate of requests for objects other than n. This is only approximately true but is intuitively reasonable. We deduce, $h(n) = (1 - h(n))q(n) \times \tau_C / \sum_{i \neq n} q(i)$ or,

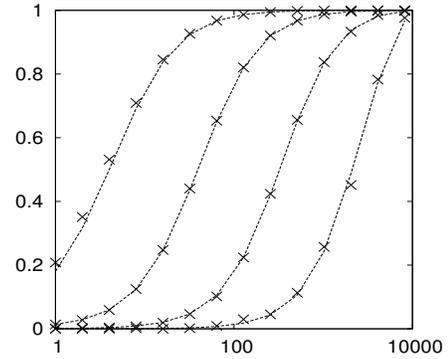$$h(n) = \frac{q(n)\tau_C}{\sum_{i \neq n} q(i) + q(n)\tau_C}, \tag{8}$$

for some unknown constant τ_C. Equating the sum of hit rates to the cache size C, as in Section III, yields the equation for τ_C,

$$C = \sum_{n=1}^{N} \frac{q(n)\tau_C}{\sum_{i \neq n} q(i) + q(n)\tau_C}. \tag{9}$$

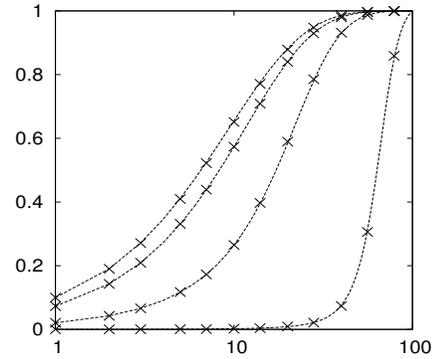Equation (9) is the 'random' equivalent to the LRU Che identity (1). Solving for τ_C yields the hit rates via (8).
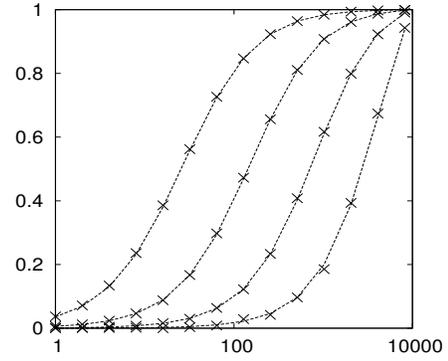
(a) $N = 10^4$, Zipf(0.8), objects 1, 10, 100, 1000

(b) $N = 10^4$, Zipf(1.2), objects 1, 10, 100, 1000

(c) $N = 100$, Geo(0.9), objects 1, 4, 16, 64

Fig. 5: Hit rate against cache size for selected objects, random replacement

Figure 5 shows results analogous to those of Figure 2 for LRU. The accuracy is clearly comparable. It largely remains to analyse why this is so but note that the assumption sojourn times are proportional to the request rate of other objects appears equally reasonable for all realistic popularity laws. The FIFO algorithm of Dan and Towsley [4] is similarly accurate.

VII. APPLICATION

In this section we present an application that is intended to illustrate the power of the Che approximation. We revisit the networking example introduced in [7] where users retrieve

	traffic share (p_i)	population (N_i)	object size (θ_i)	popularity (α_i)
Web	.18	10^{11}	10	0.8
File sharing	.36	10^5	10^6	0.8
UGC	.23	10^8	10^3	0.8
VoD	.23	10^4	10^4	1.2

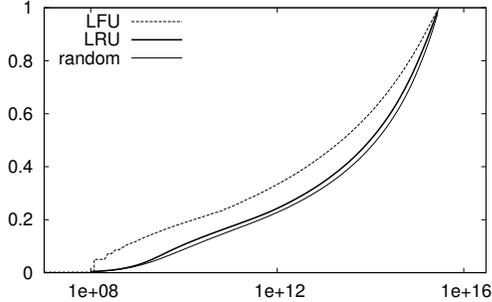TABLE I: Internet content traffic characteristics

Fig. 6: Hit rate against cache size (in bytes) for the traffic mix of Table I with LFU, LRU and random replacement

a mixture of web, file sharing, user-generated content (UGC) and video-on-demand (VoD) content via a cache.

Traffic characteristics and assumed popularity laws are presented in Table I. Note the very large populations and diverse popularity laws. These make other performance evaluation approaches, including simulation, impractical.

We suppose objects are divided into 1 KB chunks. For the sake of simplicity, objects of the same type i are supposed to have constant size θ_i chunks. We assume objects have Zipf popularity with exponent α_i and chunks inherit the popularity of their parent object. The proportion of type i traffic in bit/s downloaded by users is p_i.

Given these assumptions we deduce the popularity of chunk k of object n of type i, for $1 \leq i \leq 4$, $1 \leq n \leq N_i$ and $1 \leq k \leq \theta_i$,

$$q_i(n,k) = \frac{p_i/n^{\alpha_i}}{\sum_{j=1}^{N_i} \theta_i/j^{\alpha_i}}.$$

In applying the Che approximation we optimize summations over as many as 10^{11} objects in (1) and (9) by grouping successive terms which are nearly equal. Computation is then very rapid.

Figure 6 compares the overall hit rate for the traffic mix as a function of cache size for three different replacement policies: least frequently used (LFU), LRU and random. The LFU hit rate is calculated as in [7] while for LRU and random we use the Che approximations of Sections III and VI, respectively.

The significance of these and similar results is discussed in [7]. An additional observation is that random is hardly worse than LRU in this case. Our main objective in presenting these results is to stress that they are readily derived using the Che approximation when, in view of the huge populations and diversity of content objects, any other approach would be impracticable or inexact.

VIII. CONCLUSION

The Che approximation constitutes a versatile and highly accurate tool for predicting the hit rate performance of a cache with LRU replacement. We have demonstrated in the paper why the approximation works so well, even when the conditions suggested by its authors are not satisfied. The analysis lends confidence to using this tool to evaluate the performance of an information-centric network where the large populations and diversity of content catalogues preclude utilization of alternative approaches. Note, in particular, that the Che approximation can be usefully combined with the approach in [16] to evaluate large-scale, general cache networks.

REFERENCES

[1] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: evidence and implications. In *INFOCOM '99*, volume 1, pages 126 –134, March 1999.

[2] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 1–14, New York, NY, USA, 2007. ACM.

[3] H. Che, Y. Tung, and Z. Wang. Hierarchical web caching systems: modeling, design and experimental results. *IEEE JSAC*, 20(7):1305–1314, 2002.

[4] A. Dan and D. Towsley. An approximate analysis of the lru and fifo buffer replacement schemes. *SIGMETRICS Perform. Eval. Rev.*, 18:143–152, April 1990.

[5] W. Feller. *An introduction to probability theory and its applications*, volume II. John Wiley & Sons Ltd, New York, 2nd edition, 1971.

[6] J. A. Fill and L. Holst. On the distribution of search cost for the move-to-front rule. *Random Structures Algorithms*, 8(3):179–186, 1996.

[7] C. Fricker, P. Robert, J. Roberts, and N. Sbihi. Impact of traffic mix on caching performance in a content-centric network. In *IEEE Workshop on Emerging Design Choices in Name-Oriented Networking*, Orlando, USA, Mar. 2012.

[8] E. Gelenbe. A unified approach to the evaluation of a class of replacement algorithms. *IEEE Trans. Computers*, C-22(6), 1973.

[9] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 15–28, New York, NY, USA, 2007. ACM.

[10] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, and R. Braynard. Networking named content. In *CoNext 2009*, 2009.

[11] P. Jelenkovic, X. Kang, and A. Radovanovic. Near optimality of the discrete persistent access caching algorithm. In *International Conference on Analysis of Algorithms,*, 2005.

[12] P. R. Jelenkovic. Approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities. *Annals of Applied Probability*, 9(2):430–464, 1999.

[13] P. R. Jelenković and X. Kang. Characterizing the miss sequence of the lru cache. *SIGMETRICS Perform. Eval. Rev.*, 36:119–121, August 2008.

[14] A. Mahanti, C. Williamson, and D. Eager. Traffic analysis of a web proxy caching hierarchy. *IEEE Network*, pages 16–23, May/June 2000.

[15] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1(2):226–251, 2003.

[16] E. Rosensweig, J. Kurose, and D. Towsley. Approximate models for general cache networks. In *INFOCOM, 2010 Proceedings IEEE*, pages 1 –9, march 2010.

[17] I. G. Shevstova. An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, 82(3):862–864, 2010.

[18] A. Simonian, M. Gallo, B. Kauffmann, L. Muscariello, and C. Tanguy. Performance of the random replacement policy for networks of caches. In *Proceedings of Sigmetrics*, 2012.

[19] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. *SIGOPS Oper. Syst. Rev.*, 40:333–344, April 2006.