A Caching Policy Driven by Clusters of High Popularity

Natalia M. Markovich
Institute of Control Sciences, Russian Academy of Sciences
117997 Moscow, Russia
Email: markovic@ipu.rssi.ru

Udo R. Krieger
Faculty WIAI, Otto-Friedrich-University
D-96047 Bamberg, Germany
Email: udo.krieger@ieee.org

***Abstract*—Caching is applied to provide requested documents or Web contents quickly from a short memory. We consider the Cluster Caching Rule policy proposed recently by Markovich [12]. The idea of the rule is to keep only highly popular contents in the cache. Due to dependency in the inter-request process and heavy-tail distributed inter-request times, such frequently requested documents arise in clusters of popularity. The corresponding clusters of documents are loaded in the cache. If the requested document is present in the previous cluster, then it stays further in the cache. Otherwise, it is evicted from the cache. A mixture of $m-$dependent Markov and Poisson renewal processes is proposed as example of an inter-request time model. We present the hit/miss probabilities of such caching policy and consider cache size estimation.**

***Keywords*—Cluster Caching Rule, content popularity, clusters of extreme values, hit/miss probability, extremal index**

I. Introduction

Caching data is a widely used tool for an efficient information delivery to users. We study and develop the *Cluster Caching Rule (CCR)* policy proposed in [12] for fog computing or NDN networking at edge routers. According to this rule only highly requested popular documents may be stored in a cache of moderate size. This is really reasonable if to take into account that about 70% of contents is only once requested (cf. [5]). However, well-known rules like the Least-Recently-Used (LRU) (cf. [5]) and the Time-to-Live (TTL) (cf. [3], [7], [8]) allow to keep such documents in the cache which is not efficient. The Least-Frequently-Used (LFU) rule is similar to CCR since it stores only popular documents. But it does not take into account a possible dependence in the inter-request time (IRT) process and a random size of requested documents. The LRU, LFU and TTL rules are based on standard assumptions that may simplify the analysis. Namely, the so called Independent Reference Model (IRM) (cf. [8]) is widely accepted as a model of the request process. It implies that requests are assumed to arrive sequentially as a Poisson renewal process, i.e. IRTs are independent and exponentially distributed. Moreover, the popularity $q(n)$ of the nth object does not change that means time and spatial locality. The size of content is considered as a constant. That is not so strict if to believe that the information is requested by equal-sized chunks. The IRM has evident drawbacks: the content requests may be generated by users independently, but not necessarily exponentially distributed and independence is not the case for a hierarchy of caches when requests overflowing from lower layer caches to higher layer caches are correlated. Hence, Markov-modulated Poisson processes or Markov-arrival processes (MAPs) which model correlated requests are also popular as cache arrival models.

The CCR policy does not use the IRM and only a correlated IRT process is considered. The reason is that the independent IRT sequence does not contain clusters. Those may only appear due to dependence. Following [15] a cluster is a set of consecutive exceedances of an underlying process over a threshold arising between two consecutive non-exceedances. The CCR takes into account the cluster structure of the popularity process of the requested documents. Popularity clusters are caused by dependence in the IRT process. Heavy tails of the IRT distribution may strongly increase the cluster size and, hence, impact on the required cache size. The CCR is based on results from extreme-value theory related to clusters of exceedances over the threshold of a process of interest. All achievements follow from the probabilistic aspects of clusters of exceedances, namely distributions of their sizes. Such cluster approach allows us to explain effects related to the dependence impact on caching. The CCR may be extended to lines and trees of caches.

The objective is to obtain formulae for hit/miss probabilities and the cache size of the CCR. The hit/miss probabilities are the most important characteristics of the caching policy that reveal the probabilities to find the requested document in the cache or not, respectively. We focus on a single-cache case. Another goal is to move away from the Poisson process and to consider m-dependent Markov processes with heavy-tailed distributions as realistic models of IRT processes in the context of clusters of exceedances. To our best knowledge, such approach is novel. The rationality behind the CCR is to reduce the cache size required to keep only a finite number of popular documents and to increase the efficiency of caching. In contrast, the LRU and TTL rules decrease the miss probability by enlarging the cache and storing all requested objects.

The paper is organized as follows. Related work devoted to probabilistic aspects of caching is given in Section II. In Section III the CCR policy is defined and necessary results from extreme-value theory regarding clusters of exceedances are mentioned. In Section IV the modeling of an IRT process by the mixture of $m-$dependent Markov processes and Poisson processes is proposed as an example of correlated IRTs and the performance of the CCR policy is discussed. We obtain general analytic formulae of the hit/miss probabilities based on extreme-value theory taking into account possible repetitions of the same document in the cluster and its presence in consecutive clusters of high popularity. Based on theoretical

results regarding geometric models of cluster and inter-cluster size distributions, we also discuss several options to estimate the cache size. Some conclusions are presented in Section V.

II. Related work

The most popular caching policy is the LRU rule according to which the requested document hits the first position of the cache. The other documents located in the cache are shifted to its bottom. The object at the last position has to be evicted from the cache if the next requested document is new, i.e. it is not present in the cache. Otherwise it stays further at the last position in the cache. Despite of the simplicity, the probabilistic aspects of the LRU replacement of documents in the cache generate not easy problems. In [5] a simple exponential approximation is provided to get hit/miss probabilities of the LRU for lines and trees of caches under the IRM conditions. The TTL and LFU rules use the LRU combined with their specific properties. The TTL restricts the duration of the document to stay in cache. This time may be fixed for all documents or it may depend on the individual popularity of the document, cf. [3]. The LFU works optimal for the IRM conditions, cf. [10], and overcomes the LRU in terms of the byte hit probability, cf. [4]. But the LFU adapts poorly to changes of the popularity summarizing past high-request counts of stale documents, cf. [9], [16]. Moreover, if the document size is assumed to be fixed or, in other words, chunks of documents of a given size are requested then it is impossible to calculate the popularity of the chunks which may include different contents.

It is important how to define the popularity. One can define it as a high request rate of the document, cf. [7]. Zipf's law is accepted for the request frequency. Thus, the request frequency of a document with rank i is proportional to $1/i^{\alpha}$, where $\alpha > 0$ can be dynamically changed, cf. [4]. In [12] the popularity of the ith document at its jth request time $T_{i,j} = \sum_{n=1}^{j} \tau_{i,n}$ is defined by

$$X_i \quad = \quad j/N_{T_{i,j}}, \qquad i = 1, ..., N, \qquad j \geq 1, \qquad (1)$$

where $N_{T_{i,j}}$ is the total number of requests in time interval $[0, T_{i,j}]$. $\{\tau_{i,n}\}_{n \geq 1}$ are the IRTs of the ith document and N is the number of documents in the catalog. The popularity process of the ith object is then generated as j progresses. The common popularity process is built by such individual processes.

In [9], [10] it was derived that the cache missing probability is asymptotically the same for sufficiently large cache sizes regarding the LRU and LFU replacements, both with regard to dependent and independent identically distributed IRTs. For this purpose, the Markov-modulated Poisson process was considered as a request arrival process. The same conclusions may easily be obtained using the clustering of rare events (i.e. high popularity) if the cache size C is large enough. Since in practice C is finite, the dependence structure of the IRTs should be taken into account by means of the extremal index (see Section III-B) which defines the mean cluster size, [12].

III. The CCR cache replacement policy

A. Definition of the CCR policy

The idea of the CCR is explained in Fig. 1. Considering the popularity process $\{X_i\}$ in (1) built for request times $\{T_{i,j}\}$ of documents i from the catalog, we focus on the clusters of high popularity values exceeding the sufficiently high threshold u, see Fig. 2. Once the popularity of the document exceeds u, we put this object in the cache. If the same document appears several times in the same cluster, then the cache content does not change. The inter-cluster size $T_1(u)$ implies the number of inter-arrivals of observations running under u between two consecutive clusters of exceedances, see Fig. 2, and is determined by

$$T_1(u) \quad = \quad \min\{j \geq 1 : M_{1,j} \leq u, X_{j+1} > u | X_1 > u\},$$

where $M_{1,j} = \max\{X_2, ..., X_j\}$, $M_{1,1} = -\infty$. The cluster size $T_2(u)$ is the number of inter-arrivals of observations exceeding u between two consecutive non-exceedances, see Fig. 2, and is determined by

$$T_2(u) \quad = \quad \min\{j \geq 1 : L_{1,j} > u, X_{j+1} \leq u | X_1 \leq u\},$$

where $L_{1,j} = \min\{X_2, ..., X_j\}$, $L_{1,1} = +\infty$. Hence, we get $T_2^*(u) \leq T_2(u)$, where $T_2(u)$ is the cluster size and $T_2^*(u)$ is the number of exceedances in the cluster corresponding to different objects (the cluster size subtracting the document repetitions).

If the document popularity exceeds the threshold within several consecutive clusters, then such document will stay in the cache

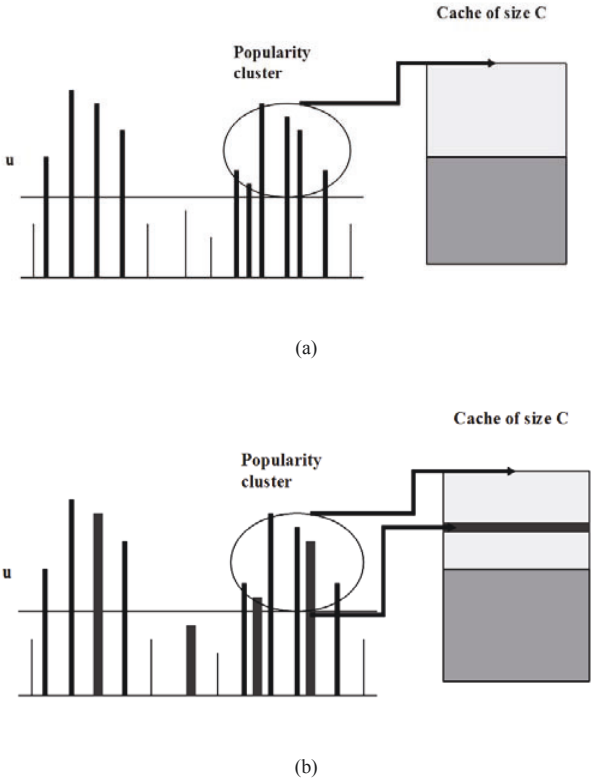

Fig. 1. The documents corresponding to a popularity cluster are uploaded in the cache, where the cluster contains all documents being different (a). In case the same document appears in one cluster repeatedly (its popularity is shown by a solid grey column), it is uploaded in the cache once (b). The same document may appear between clusters if its popularity falls down, but it will not be evicted from the cache during the inter-cluster time (an inertial effect).

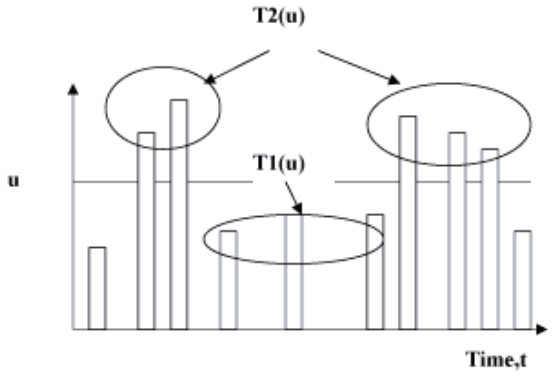

Fig. 2. Clusters of exceedances over threshold u of an underlying process.

within the duration of these clusters and corresponding inter-clusters. If the document is not found in the next cluster, it is evicted from the cache. *Durations of the cluster* structure $S_{T_2(u)}$ and *inter-cluster* structure $S_{T_1(u)}$ are determined by

$$S_{T_1(u)} = \sum_{i=1}^{T_1(u)} y_i, \qquad S_{T_2(u)} = \sum_{i=1}^{T_2(u)} y_i,$$

where $\{y_i\}$ are inter-arrival times between documents from the catalog in the common request arrival process. Then the minimal time for the ith document to stay in the cache (*the minimal document life time in the cache*) is equal to the duration of a cluster after the first exceedance of X_i plus the duration of the next coming inter-cluster:

$$T_i^{min}(u) \leq S_{T_1(u)} + S_{T_2(u)}.$$

The *maximal document life time in the cache* is formally not restricted and it depends on the random number N_c of consecutive clusters where the popularity of the ith document exceeds u, i.e. it holds

$$T_i^{max}(u) \leq \sum_{k=1}^{N_c} (S_{T_1(u),k} + S_{T_2(u),k}),$$

where $S_{T_1(u),k}$ and $S_{T_2(u),k}$ are durations of the kth inter-cluster and cluster. In this respect the CCR is similar to the TTL policy when the time to live in the cache depends on the popularity.
If some documents from the catalog are not requested, their popularity is equal to zero and they cannot enter the cache.

B. Clusters of exceedances and their probabilities

As the proposed CCR policy is derived from statistical properties of the popularity process induced by the request process of a document, we consider now clusters of exceedances of the popularity process over a threshold. The cluster is determined as a set of consecutive exceedances of the popularity between two consecutive non-exceedances. According to the theory of extreme values such clusters become independent for sufficiently high thresholds and the corresponding point process (i.e. epochs of the appearance of such rare events) proceeds as a renewal process, more exactly, as a compound Poisson process with specific intensity equal to $\theta\tau$. Here, the extremal index θ provides the reciprocal of the mean cluster size and plays a fundamental role in the theory. The constant $\tau > 0$ is determined in the following definition and it relates to the selection of the sufficiently high threshold u_n.

Definition 1: The stationary sequence $\{X_n\}_{n\geq 1}$ with common distribution function $F(x)$ and $M_n = \max\{X_1, ..., X_n\}$, is said to have extremal index $\theta \in [0, 1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers $u_n = u_n(\tau)$ such that

$$\lim_{n\to\infty} n(1 - F(u_n)) = \tau, \qquad \lim_{n\to\infty} P\{M_n \leq u_n\} = e^{-\tau\theta} \tag{2}$$

hold (cf. [11, p. 53]).

In case the random sequence is independent and identically distributed (iid) we have $\theta = 1$. The total dependence corresponds to $\theta = 0$. Taking into account that $1/\theta$ approximates the mean cluster size, $\theta = 0$ implies infinite size clusters. Those cannot be separated by any inter-cluster intervals. With regard to the maxima distribution we have

$$P\{M_n \leq u_n\} = P^{n\theta}\{X_1 \leq u_n\} + o(1), \qquad n \to \infty.$$

If $\theta = 0$ holds, then we get $\lim_{n\to\infty} P\{M_n \leq u_n\} = 1$, i.e. the maximum increases slower than in the iid case and it cannot exceed the threshold u_n that goes to infinity. The processes with $\theta = 0$ are not so unusual. The Metropolis algorithm allows to simulate Markov chains with given stationary distributions. Metropolis Markov chains with noise having some heavy-tailed distributions including the Zipf law (cf. [17]) and a Lindley process with subexponential jumps which is used as a model for waiting times in queuing systems (cf. [1]) provide examples of such processes. The subexponential distribution class includes all possible heavy-tailed distributions. Regarding our caching problem $\theta = 0$ is in contradiction to the IRM and it leads to non-optimal LRU, LFU and TTL rules. It implies that only an infinite cache size can be appropriate if the CCR rule is applied and the documents corresponding to clusters of popularity exceedances enter the cache.
Clusters of the popularity process are identical to inter-clusters of the IRTs, i.e. small values of IRTs running under a given threshold.
To get the hit/miss probabilities for the CCR, we need distributions of cluster and inter-cluster sizes for a given threshold u. In [15] and [14] geometric-like models for distributions of $T_1(u)$ and $T_2(u)$ which depend on the extremal index θ were derived and improved. Using high quantiles x_{ρ_n} of the levels $(1 - \rho_n)$ of the process X_t as u_n ($\rho_n \to 0$ as $n \to \infty$ due to (2)), the following approximations

$$P\{T_1(x_{\rho_n}) = j\} \approx \theta^2 \rho_n (1 - \rho_n)^{(j-1)\theta}, \tag{3}$$

$$P\{T_2(x_{\rho_n}) = j\} \approx \theta^2 q_n \left(1 - q_n\right)^{(j-1)\theta}, \tag{4}$$

were obtained for sufficiently large j and $\theta \in [0, 1]$, where $q_n = 1 - \rho_n$. Models (3) and (4) are derived in [15], [14] under specific mixing conditions. Those require a statistical independence of observations separated by a sufficiently large time. Such conditions are fulfilled particularly for regenerative processes where observations can be a partition at independent regenerative cycles.
Fig. 3 shows that the probability of $T_2(u)$ declines strongly for $j > 6$. As there are no results yet regarding the maxima

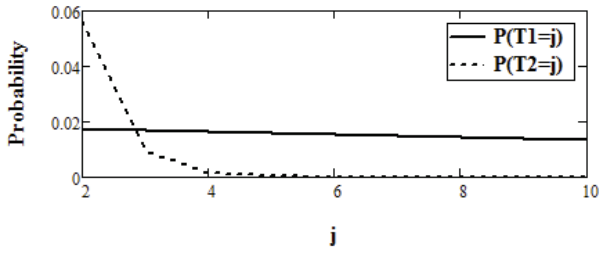

Fig. 3. Probability models (3) and (4) for $\theta = 0.6$ and $\rho = 0.05$ against j; both probabilities do not change much when $j > 6$.

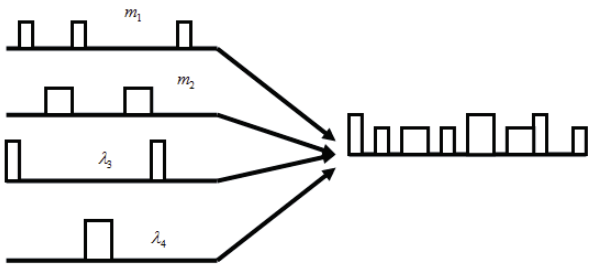

Fig. 4. Superposition of m-dependent Markov chains with m_1 and m_2 and Poisson renewal processes with λ_3 and λ_4 as request models of different documents.

distribution of $T_2(u)$, this allows to suspect that cluster sizes are likely less than 6. Then the cache size for the CCR can be determined in a similar way. In [12] it was proposed to take a mean cluster size of the popularity process as a cache size. Since by theory it holds

$$ET_2(u) \approx 1/\theta, \tag{5}$$

the cache size can be taken as $C = 1/\theta$. From (5) it follows that the estimate of the cache size depends on u, $\widehat{C} = \widehat{C}(u)$. This can be used for the CCR (see Section IV-C) to fill the cache better, i.e. to increase the cache utilization, cf. [12]. If the sufficiently high quantile x_{ρ_n} of the level $1 - \rho_n$ is used as u_n, then one can use the approximation in terms of ρ_n (cf. [15])

$$ET_2(u_n) \approx \theta^2(1-\rho_n)/(1-\rho^\theta)^2 \tag{6}$$

and use its right-hand side as the cache size $C(\rho_n)$. Models (3) and (4) work particularly well for $m-$dependent processes like the MM-process (cf. [15]) that is considered in Section IV-A as a model of the IRT process.

To find the probability to enter the cache at first time, we recall the first hitting time

$$T^*(u) = \min\{j+1 \geq 1 : M_j \leq u, X_{j+1} > u\}$$

and the model of its distribution

$$P\{T^*(x_{\rho_n}) = j+1\} \approx \frac{\theta^2\rho_n^2(1-\rho_n)^{\theta(j-1)}}{1-(1-\rho_n)^\theta}, \tag{7}$$

$j = 0, 1, 2, ...$, $M_0 = -\infty$, obtained in [13]. The probability to hit the threshold x_{ρ_n} twice, i.e.

$$P\{T^*(x_{\rho_n}) = j, T^{**}(x_{\rho_n}) = j+m\} = \tag{8}$$
$$P\{M_{j-1} \leq x_{\rho_n}, X_j > x_{\rho_n}, M_{j,j+m-1} \leq x_{\rho_n}, X_{j+m} > x_{\rho_n}\},$$

$m = 1, 2, ...$, may be approximated by the product $P\{\chi = j\}P\{\chi = m\}$, where χ is a geometrically distributed random variable with probability $\rho_n\theta$, cf. [13].

IV. Performance of the CCR replacement policy

A. Modeling the inter-request time process

Let $\{\tau_{i,t}, t \geq 1\}$ be the IRT sequence and $S_{i,j} = \sum_{t=1}^{j} \tau_{i,t}$ be the current time of the jth request of the ith document. Then $\{S_{i,j}\}$, $i = \overline{1,N}$, $j = 1, 2, ..., j_i$ are sequences of request times of documents from the catalog before time T, where $j_i = \arg\max\{j : S_{i,j} \leq T\}$ is the maximal number of requests of the ith document until the time T.

To construct a superposition of IRT series of all documents from the catalog one can reorder $\{S_{i,j}\}$ in increasing order denoting them as $Z_1 \leq Z_2 \leq ... \leq Z_l$, where l is a random number such that $l = \max\{j_i, i = \overline{1,N}\}$. Such IRTs are correlated. This allows to calculate the popularity by (1) using the label sequence $A_1, ..., A_l$ that indicates the numbers of the documents corresponding to $\{Z_i\}$.

Example 1: Let us model the IRT sequence by a mixture of m-dependent Markov chains and Poisson renewal processes with parameters λ regarding the document type. The first model is appropriate for news which are only important for a short time period. The m-dependence means the independence of the IRTs $\tau_{i,t}$ and $\tau_{i,t+m}$. The m determines the degree of interest or popularity duration. Documents related to science or culture may attract permanent interest for a long time and their requests are better fitted by the second model. Each document from the catalog of N objects has an own m or λ value. Reordering a catalog with k Markov chains with $m_1, ..., m_k$ values and $N-k$ Poisson renewal processes with $\lambda_{k+1}, ..., \lambda_N$, one can model their superposition selecting step by step documents with the minimal current times, see Fig. 4. For simplicity, we will assume that the set $m_1, ..., m_k$ does not change over time. Despite that news topics in the catalog are changed in time, the longevity of their popularity may be assumed the same.

As the m-dependent Markov chain of the ith document we propose the Moving Maxima (MM) process:

$$\tau_{i,t} = \max_{j=0,...,m_i}\{\alpha_j Z_{t-j}\}, \qquad t \in \mathrm{Z},$$

with nonnegative constants $\{\alpha_j\}$ such that $\sum_{j=0}^{m_i}\alpha_j = 1$ and iid standard Fréchet distributed r.v.s $\{Z_t\}$ with distribution function $F(x) = P\{Z_i \leq u\} = e^{-1/u}$. The distribution of $\tau_{i,t}$ is also standard Fréchet. The extremal index of the MM process is known to be equal to $\theta = \max_i\{\alpha_i\}$.

The extremal index of the Poisson renewal process is equal to 1 due to the independence of the corresponding exponentially distributed IRTs. Cluster properties of the obtained mixed IRT sequence are then totally determined by the MM components of the mixture and the hit/miss probabilities can be well modeled by (3) and (4). In Fig. 5 one can see that the cluster sizes of the MM process are the larger the smaller is the value θ. In Fig. 6 one can see the cluster structure of the popularity sequence corresponding to one mixed IRT sequence from our example.

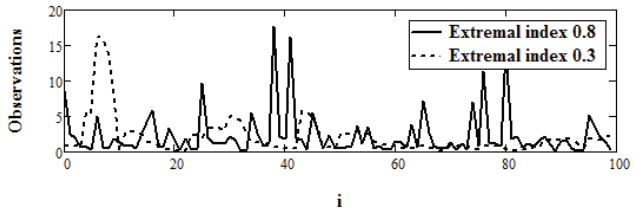

Fig. 5. Cluster structure of the MM processes with $\theta \in \{0.3, 0.8\}$.

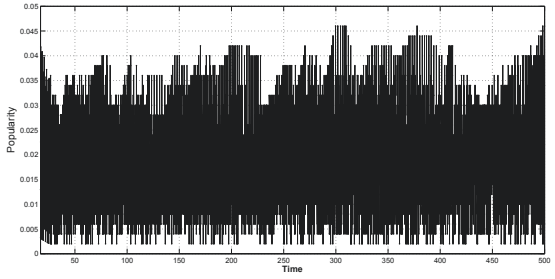

Fig. 6. Popularity (1) of documents modeled by the superposition of 90% MM processes and 10% Poisson processes against the time; the catalog contains $N = 100$ documents.

B. Estimation of the extremal index and statistical calculations

In practice, it is difficult to find analytic formulae of θ since precise models of the IRT and, hence, the popularity processes are usually unknown. Thus, one can apply nonparametric estimators of the extremal index. Well-known estimators are blocks, runs and intervals (cf. [2]) among others. We prefer the intervals estimator which has a good accuracy and depends only on the threshold u as parameter while the blocks and runs estimators require the block size as the second parameter. The intervals estimator proposed in [6] is determined by

$$\hat{\theta}_n(u) \qquad (9)$$
$$= \begin{cases} \min(1, \hat{\theta}_n(u)), & \text{if } \max\{T_i : 1 \le i \le N-1\} \le 2, \\ \min(1, \hat{\theta}^*_n(u)), & \text{if } \max\{T_i : 1 \le i \le N-1\} > 2, \end{cases}$$

where

$$\hat{\theta}_n(u) = \frac{2(\sum_{i=1}^{N-1} T_i)^2}{(N-1)\sum_{i=1}^{N-1} T_i^2},$$

$$\hat{\theta}^*_n(u) = \frac{2(\sum_{i=1}^{N-1}(T_i-1))^2}{(N-1)\sum_{i=1}^{N-1}(T_i-1)(T_i-2)},$$

$N = \sum_{i=1}^{n} 1(X_i > u)$ is the number of exceedances of u at time epochs $1 \le S_1 < \ldots < S_N \le n$ and the interexceedance times are given by $T_i = S_{i+1} - S_i$. The value corresponding to a stability interval regarding u is usually proposed as the estimate $\hat{\theta}$. In Fig. 7 we estimate the extremal index of the popularity process built by (1) using the simulated mixed process from Example 1. Using the obtained value $\widehat{\theta} = 0.8$ we can estimate the hit/miss probabilities as shown by (10). Since the mean cluster size is approximately equal to $1/\widehat{\theta} = 1.25$ due to (5), we can propose the cache size $C \ge 2$. The smaller θ and larger the mean cluster size, the larger cache is required. Since $\widehat{\theta} = 0.8$ corresponds on average to $1 - \rho \approx 0.3$, one can

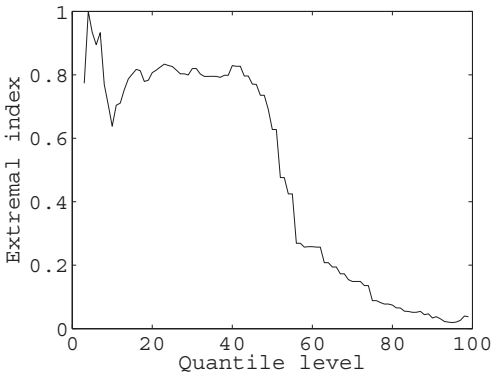

Fig. 7. Intervals estimator (9) of the extremal index against $(1-\rho)\cdot 100\%$, where $1-\rho$ is the quantile level of the threshold x_ρ: the stability interval $[20, 45]$ corresponds to $\theta = 0.8$ that can be selected as estimate.

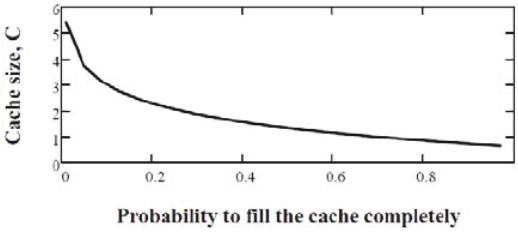

Fig. 8. Cache size C against χ for $a = 1.562$, $\theta = 0.8$, $q = 0.7$.

calculate the cache size by (6) for $\rho = 0.7$ as $C \ge 3$ since the mean cluster size is then 3.116.

C. Cache size estimation

A major problem concerns the cache size estimation to provide an effective cache utilization. Since the cluster size is random, it may happen that the cluster may be smaller or larger than the cache size, i.e. $T_2(u) > (\le) C$ for a given u. Then the cache may not be utilized completely (see the dark part in the cache in Fig. 1) or, on the contrary, some documents will not find a place in the cache. This may be the case when the cache size is selected as the mean cluster size by (5) or (6).

Several options are then possible: (1) we can change u to make clusters larger or smaller to fit the cache size; (2) having the cache size fixed, we may allow a part of documents with largest popularity from the previous cluster to stay in the cache longer to fill the cache completely; (3) if our cache is too small for the cluster we can send the rest of documents in the next cache if there is a line or a tree of caches at our disposal.

One can select C from (4)

$$P\{T_2(x_\rho) = C\} \quad = \quad a\theta^2 q\,(1-q)^{(C-1)\theta} = \chi$$

taking the right-hand side equal to χ close to 1. χ is the probability to fill the cache completely. Selecting the constant $a \ge \chi/(\theta^2 q)$ we get $C \ge 1$. For high thresholds corresponding to q close to 1 we obtain single exceedances in the clusters that corresponds to $C = 1$, see Fig. 8.

D. Hit/miss probabilities of the CCR policy

The probability to hit the cache by the ith document at the first time is determined by (7). This is the probability that its popularity X_i exceeds the threshold u at time epoch $j+1$ at the first time. The probability for the ith document to get in two consecutive clusters once is determined by (8). The probability of the second hitting time is equal to the probability of the ith document to be in cache longer than the sum of the durations of two clusters and two inter-clusters. Note, that in this case θ in (7) and (8) is equal to the extremal index θ_i of the popularity process of the ith document.

The cache *hit probability* for some set of j documents coincides with the probability for their popularity to exceed the threshold or to fall in the cluster. It is determined by (4), where j shows the number of exceedances in the cluster. The *miss probability* is then the probability for the popularity not to exceed the threshold and it is determined by (3). It is remarkable that these probabilities as (3) and (4) are the same irrespective of the distribution of the popularity process $\{X_i\}$ and they depend on the extremal index θ of the latter process of all requested documents. If the popularity of some document exceeds u several times within one cluster, then this does not change the content of the cache. Despite of the possible repetitions of exceedances of the same document in the clusters, their popularity will be different. Then (3) and (4) approximate the probabilities of cluster and inter-cluster sizes of distinct documents without repetition, i.e.

$$\begin{aligned} P\{T_1^*(x_{\rho_n}) = j^*\} &\approx \theta^2 \rho_n (1-\rho_n)^{(j-1)\theta}, \\ P\{T_2^*(x_{\rho_n}) = j^*\} &\approx \theta^2 q_n \left(1-q_n\right)^{(j-1)\theta}. \end{aligned} \quad (10)$$

An uncertainty of the miss probability arises when the popularity of some requested objects drops down below the threshold. Then the real miss probability may be less than (4). Empirical estimates of the hit/miss probabilities for the time t are still the same as for the LRU and TTL, i.e. $1 - M(t)/N(t)$ and $M(t)/N(t)$, where $M(t)$ is the number of missed objects, $N(t)$ is the number of all requests, cf. [3].

V. Conclusions

The Cluster Caching Rule proposed recently in [12] is studied. We consider the popularity process of requested documents and its clusters of exceedances over sufficiently high thresholds. The popularity at each request epoch is calculated as the ratio of the number of requests of a specific document and the total number of documents requested up to this moment. The clusters are caused by correlations in the IRT process of requested documents. Thereby, we avoid the usual Independent Reference Model and its assumptions like the constant document size, unchanged popularity and the independence of the IRTs.

The paper contains the following novelties. 1)We propose the mixture of $m-$dependent Moving Maxima and renewal Poisson processes as an example of a correlated IRT model since this is realistic for short- and long-term requested objects. 2) The hit/miss probabilities of the CCR policy are approximated by geometric-like probabilities of the clusters and IRTs, respectively. Significant is that the latter geometric models are invariant regarding the distribution of the IRTs. But it is sensitive to dependence that is represented by the extremal index. 3) Based on the geometric-like models of the hit/miss probabilities we propose the selection of the cache size which does not depend on the IRT distribution, too. We believe that the CCR policy has a better hit probability for small cache sizes than the TTL and LRU schemes.

The CCR approach is appropriate for any correlated IRT process. Its comparison with other caching rules by empirical data is a subject of our future work.

Acknowledgment

The first author thanks DAAD for the financial support of this research project by a scholarship.

References

[1] S. Asmussen, "Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities," *Ann. Appl. Probab.*, vol. 8, no. 2, pp. 354–374, 1998. [Online]. Available: http://dx.doi.org/10.1214/aoap/1028903531

[2] J. Beirlant, Y. Goegebeur, J.Teugels, and J. Segers, *Statistics of Extremes: Theory and Applications*. Chichester, West Sussex: Wiley, 2004.

[3] D. S. Berger, P. Gland, S. Singla, and F. Ciucu, "Exact analysis of TTL cache networks: the case of caching policies driven by stopping times," in *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '14, 2014.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1, 1999, pp. 126–134.

[5] H. Che, Y. Tung, and Z. Wang, "Hierarchical Web caching systems: modeling, design and experimental results," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305–1314, Sep 2002.

[6] C. Ferro and J. Segers, "Inference for clusters of extreme values." *Journal of the Royal Statistical Society, Series B*, vol. 65, pp. 545–556, 2003.

[7] N. C. Fofack, P. Nain, G. Neglia, and D. Towsley, "Analysis of TTL-based cache networks," in *6th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS) 2012*, Oct. 2012, pp. 1–10.

[8] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for LRU cache performance," *CoRR*, vol. abs/1202.3974, 2012.

[9] P. R. Jelenkovic and A. Radovanovic, "Least-recently-used caching with dependent requests," *Theoretical Computer Science*, vol. 326, no. 13, pp. 293 – 327, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030439750400475X

[10] ——, "Asymptotic optimality of the static frequency caching in the presence of correlated requests," *Operations Research Letters*, vol. 37, no. 5, pp. 307 – 311, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167637709000510

[11] M. Leadbetter, G. Lingren, and H. Rootzén, *Extremes and Related Properties of Random Sequence and Processes.* Springer, 1983.

[12] N. M. Markovich, "A cluster caching rule in next generation networks," in *18th International Conference on Distributed Computer and Communication Networks*, Oct. 2015, pp. 127–135.

[13] ——, "Clustering and hitting times of threshold exceedances and applications," *Submitted in IJDATS*, 2016.

[14] ——, "Erratum to: Modeling clusters of extreme values," *Extremes*, vol. 19(1), p. 139-142, 2016.

[15] ——, "Modeling clusters of extreme values," *Extremes*, vol. 17, no. 1, pp. 97–125, 2014.

[16] N. Megiddo and D. S. Modha, "Outperforming LRU with an adaptive replacement cache algorithm," *Computer*, vol. 37, no. 4, pp. 58–65, Apr. 2004. [Online]. Available: http://dx.doi.org/10.1109/MC.2004.1297303

[17] G. O. Roberts, J. S. Rosenthal, J. Segers, and B. Sousa, "Extremal indices, geometric ergodicity of Markov chains, and MCMC," *Extremes*, vol. 9, no. 3-4, pp. 213–229, 2006.