Computer Networks 125 (2017) 122–131

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Fair throughput allocation in Information-Centric Networks

Thomas Bonald[a], Léonce Mekinda[b,*], Luca Muscariello[c]

[a] TELECOM ParisTech, 23 Avenue d'Italie, 75013 Paris, France
[b] European XFEL, Holzkoppel 4, 22869 Schenefeld, Germany
[c] Cisco Systems, 11 Rue Camille Desmoulins, 92130 Issy-les-Moulineaux, France

ARTICLE INFO

Article history:
Received 15 October 2016
Revised 5 March 2017
Accepted 18 May 2017
Available online 20 May 2017

Keywords:
ICN
Caching
Fairness
Network performance analysis

ABSTRACT

Cache networks are the cornerstones of today's Internet, helping it to scale by an extensive use of Content Delivery Networks (CDN). Benefiting from CDN's successful insights, ubiquitous caching through Information-Centric Networks (ICN) is increasingly regarded as a premier future Internet architecture contestant. However, the use of in-network caches seems to cause an issue in the fairness of resource sharing among contents. Indeed, in legacy communication networks, link buffers were the principal resources to be shared. Under max-min flow-wise fair bandwidth sharing [14], content throughput was not tied to content popularity. Including caches in this ecosystem raises new issues since common cache management policies such as probabilistic Least Recently Used (p-LRU) or even more, Least Frequently Used (LFU), may seem detrimental to low popularity objects, even though they significantly decrease the overall link load [3]. In this paper, we demonstrate that globally achieving *LFU is a first stage of content-wise fairness*. Indeed, any investigated content-wise α-fair throughput allocation permanently stores the most popular contents in network caches by ensuring them a cache hit ratio of 1. As ICN caching traditionally pursues LFU objectives, content-wise fairness specifics remain only a matter of fair bandwidth sharing, keeping the cache management intact.

1. Introduction

Today's Internet owes its scalability to caching. Indeed, most of Internet contents cross Content Delivery Networks and significant research is pushing for a better solution, Information-Centric Networks. In ICN, and more specifically, Named-Data Networking (NDN) and Content-Centric Networking (CCN) [9], two leading ICN architectures, content objects are identified by their unique name. At every node/router, content Data packets are requested via matching Interest packets, through egress interfaces. Interests and their satisfying Data counterparts follow rigorously the same path. This feature would not be possible without the Pending Interest Table (PIT) structure that keeps track of every requesting interface and requested content. Naming Data packets allows storing them, on every traversed node, in a finite memory referred to as Content Store (CS) or cache and managed by an object eviction policy.

Caches and their eviction or management policies are the disruption that drives this paper. Traditionally, networks are modeled as interconnected queues with fair schedulers. The penetration of caching into the network layer clearly favors a few content objects, the most popular ones in case of the Least Frequently Used management policy (LFU) and its approximations such as (p-) LRU or LRU+Leave-Copy-Down [13]. Filling caches steadily with the most popular items, meaning keeping their hit ratio to their maximum *i.e.,* one, and letting other hit ratios be zero, entails the sacrifice of less popular objects [3]. This is at least a view discussed by state-of-art contributions on content-wise cache fairness [6,25]. These works observed the hit ratio on a single cache or a network of caches and prescribed an adaptation of the cache management policy for the purpose of fairness. For example, in [6], content-wise max-min fairness is only achievable if the hit ratios are forced to be equal for all content objects. In the same vein, proportional fairness requires that content hit ratio be proportional to their popularity. A consequence of this is that ICN cannot be fair to contents without revising its caching algorithms. From the viewpoint of these works, LFU is definitely unfair to lower popularity contents. By the way, remember *flow-wise* fairness means allocating resources such that every flow/route gets its fair share. On the other hand, by *content-wise* fairness, we denote allocating resources in such a way every content gets its fair share. This is the type of fairness this paper addresses.

Our paper analyzes the fairness of content delivery throughput in accounting for both cache hit ratio and link service rates, and

* Corresponding author.
E-mail addresses: thomas.bonald@telecom-paristech.fr (T. Bonald), leonce.mekinda@xfel.eu, mekleo@yahoo.fr, leonce.mekinda@telecom-paristech.fr (L. Mekinda), lumuscar@cisco.com (L. Muscariello).

http://dx.doi.org/10.1016/j.comnet.2017.05.019
1389-1286/

comes up with a different and optimistic conclusion. *ICN's traditional caching optimum leads to content-wise fairness as it is*. The better the convergence to LFU, the better the feasible content-wise fairness. The remaining task would consist in implementing content-wise fairness at the packet scheduling stage in ICN, similarly to flow-wise fairness in other networks [10]. Taking a network of caches as a whole, links and caches, the paper sheds new light on content-wise fair cache allocation. While previous works only considered caches and concluded that caching policies have to be adapted to be α-fair to contents, this work shows that LFU and its approximations are sufficient as they are, and content-wise α-fairness is the responsibility of network packet schedulers. This contribution brings α-fairness in ICN and α-fairness in traditional networks closer. Our results owe to the link service to the majority of contents that balances the rather permanent cache presence of a few contents. It is rather commonplace that persisting the most popular contents frees a maximal upstream link capacity to convey less popular objects. Another striking insight we got, is that a throughput-optimal content delivery network ends up being made up of autonomous caches that never forward their miss traffic. Such a network would not be committed to locally satisfy requests.

The main contributions of this paper are that: (i) it unifies caches and network queues into a single content service rate model; (ii) it tackles for the first time content throughput fairness in ICN in formulating that as a tractable nonlinear optimization problem; (iii) it provides closed-form expressions of α-fair hit ratios and link service rates; (iv) it indicates that today's LFU-approximating caches policies do not need to be replaced for ICN to become fair. We articulate these contributions throughout the paper as follows: Section 2 recapitulates previous contributions on fairness in the context of cache networks. In Section 3, we model the per-content throughput in unifying cache and network link contributions. Then we formalize α-fair allocations, key properties such as their Pareto-efficiency, and that LFU is an α-fair cache management policy, an important result. To ground the theory, a few trivial examples are analyzed in Section 4. They are followed in Section 5 with numerical evaluations that confirmed, by means of a nonlinear problem solver, our analytic insights.

2. Related work

Very few papers address the issue of fairness in networks of caches. In a paper dedicated to the subject some time ago [25], authors analyze the fairness in Content-Centric Networks from the viewpoint of object dissemination across the network. They expressed content-wise fairness as the total space contents occupy with respect to their popularity. The study concluded that medium-popularity content were favored as they spread linearly with their popularity whereas the most popular items spread sublinearly. This approach is definitely useful to map the asymptotic replica spatial distribution. However, it does not capture the throughput fairness. Shah and Veciana [21,22] tackled the impact of fairness on delivery time in large scale CDN but ignored the cache specifics. That work modeled cache networks as classical networks of file-serving queues. Files were assumed to have been pre-fetched and their long-term popularity was not taken into account.

Quite recently, [6] reverse-engineered popular LRU and LFU policy and found the utility function each policy optimizes. These utility functions achieve various classes of hit ratio α-fairness. Authors also provided algorithms for adapting Time-to-Live (TTL)-based caches to any given α-fair objective. Rapidly, [16] applied this work's reverse engineering approach to a special case of a novel class of latency-aware caching (LAC) policies previously introduced by Carofiglio et al. [5]. In [16], Neglia et al. show that LAC policies converge to the solution of a fractional knapsack problem (LFU) when their latency exponent tends to infinity.

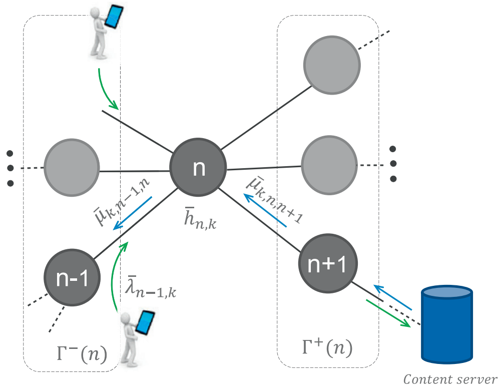

Fig. 1. Network conveying content k through cache n.

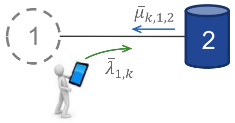

Fig. 2. Client/server topology.

Most of the existing literature on the subject, because of its focus on hit ratio, concluded that caching policies had to adapt to the content-wise fair objective. Our contribution is novel because it joins cache and link queue occupation in order to analyze the QoE-expressive throughput fairness. The QoE considered in the paper refers to how fair the user may perceive the throughput of the most popular content compared to those of less popular contents. We show that cache networks, and ICN in particular, can be α-fair, for any $\alpha \geq 0$, as soon as they couple the classical highest popularity content persistence *i.e.*, the global LFU cache management policy, with a proper content-aware α-fair packet scheduler.

3. Cache network model

First, we present a mathematical model that captures the dynamics of the entire network. The model views the latter as a network of queues where caches contribute to increase the network service rate. We aim at maximizing a utility function of the admissible exogenous traffic rate. Refer to Table 1 for the notation and to Fig. 1 for the model used hereinafter (Fig. 2).

3.1. Model assumptions

- Let the stochastic process $\{\lambda_{k,\,n,\,b}(t)\}_{0\,\leq\,t\,\leq\,T}$ be content k exogenous rate on link $(n,\,b)$ at time t. Let the stochastic process $\{\mu_{k,\,n,\,b}(t)\}_{0\,\leq\,t\,\leq\,T}$ be content k service rate on the link $(b,\,n)$ at time t. Let the stochastic process $\{h_{n,\,k}(t)\}_{0\,\leq\,t\,\leq\,T}$ be content k hit ratio on node n at time t. These processes are independent.
- The network routes based on a single prefix.
- Same object sizes. This is a widely adopted assumption in the caching literature [7]. It lies on the idea that the actual disparities among content sizes are embodied by the popularity factor q_k, which multiplies a content quantum (*e.g.*, a mean chunk size).
- Cache size is never zero.
- Content servers are not clients.
- The exogenous traffic on a given node is the one generated by a local application that is not satisfied by the local cache.
- We assume hop-by-hop congestion control *i.e.*, interests are sent in average at a rate equivalent to the link service rate.