Cost-aware caching: Caching more (costly items) for less (ISPs operational expenditures)

Andrea Araldo, D. Rossi, Fabio Martignon

Cost-aware caching: Caching more (costly items) for less (ISPs operational expenditures)

Andrea Araldo*†, Dario Rossi† Fabio Martignon*‡
* LRI, Université Paris-Sud, {first.last}@lri.fr
† Telecom ParisTech, {first.last}@enst.fr ‡ IUF, Institut Universitaire de France

***Abstract*—Albeit an important goal of caching is traffic reduction, a perhaps even more important aspect follows from the above achievement: the reduction of Internet Service Provider (ISP) operational costs that comes as a consequence of the reduced load on transit and provider links. Surprisingly, to date this crucial aspect has not been properly taken into account in cache design.**

In this paper, we show that the classic caching efficiency indicator, i.e. the hit ratio, conflicts with cost. We therefore propose a mechanism whose goal is the reduction of cost and, in particular, we design a Cost-Aware (CoA) cache decision policy that, leveraging price heterogeneity among external links, tends to store with more probability the objects that the ISP has to retrieve through the most expensive links. We provide a model of our mechanism, based on Che's approximation, and, by means of a thorough simulation campaign, we contrast it with traditional cost-blind schemes, showing that CoA yields a significant cost saving, that is furthermore consistent over a wide range of scenarios. We show that CoA is easy to implement and robust, making the proposal of practical relevance.

I. Introduction

Information Centric Networking (ICN) is a network paradigm having received increasing attention in the last decade, though its foundation can be traced back in the nineties [1]. The current Internet is composed of hosts, uniquely identified by IP addresses and exchanging data in a point-to-point communication. By contrast, ICN deploys caches over the network to store the most popular contents, and users send requests with the name of the content, without specifying the location of the content copy. It is expected that ubiquitous and transparent ICN caches can reduce the network load for ISPs, as well as reduce latency for users.

Despite these expected technical benefits, widely explored by the research community, ICN has so far remained only in the research literature, unlike Content Delivery Networks (CDN) or HTTP caches that enjoy a large deployment. In our opinion, the reason is that technical benefits are not sufficient to convince big Internet players (like ISPs and Content Providers) to switch to a new network paradigm, if they do not have clear *economic incentives*. As such, the investigation of the economic implications of ICN may have notable impact.

Under this light, we show that ICN can help ISPs reduce the cost related to inter-domain traffic: this is of crucial importance since inter-domain traffic grows by up to 60% every year [2], which is faster than cost reduction that current technology can offer [3]. This importance is confirmed by the flourishing literature on ISP cost [2]–[16]. While this reduction is usually achieved via routing [4], new peering interconnections with other ISPs [4] or traffic shaping [5], in this paper we argue that the *caching* function of ICN is also fit for the purpose.

Inter-domain traffic cost can be obviously reduced by increasing the cache storage space, but this would imply an increase in capital expenditure. This trade-off has already been tackled [6] and is not the object of our investigation. Our viewpoint is complementary: we aim to attain cost saving with respect to classic caching, leaving the cost of cache deployment unchanged and only by designing a proper cache strategy.

In order to do so, we adopt a different approach with respect to previous work [7], [17]–[26] in which optimizing caching efficiency is the *goal*. In this work, we instead advocate that caching should be considered *as a means* to obtain benefits not only from the user perspective (e.g., to reduce the retrieval distance or delay), but also in terms of ISP cost reduction. In other words, we argue that content retrieval cost should be explicitly taken into account in distributed network operation, making ICN economically profitable for ISPs. Note that we purposely adopt a rather extreme viewpoint by uniquely considering ISP cost: our aim is to gauge the extent of cost reduction that is achievable through simple caching techniques, which have been neglected so far in the literature (but see Sec. VII for a broader discussion).

We point out that the results and the mechanism presented in this work can be applied whenever caching is involved, i.e., apart from ICN, in Web proxy and CDN caching. However, we choose to primarily focus on ICN, because it represents the most challenging case for our proposal: since all decisions need to be performed at line rate, potential distributed solutions need to be effective, simple and scalable at the same time. Building on [27], that focuses on the ICN cost-awareness from an architectural perspective, this work extend its preliminary performance evaluation along several directions. Our key contributions can be summarized as follows:

- We motivate our research by showing that cost contrasts with hit ratio and that, therefore, it is necessary to design mechanisms that, contrarily to the classic ones whose primary goal is to maximize the hit ratio, directly attempt to reduce the cost (Sec. III-A).
- We design a Cost-Aware (CoA) scheme for a network of caches (Sec. III-B), of which we provide a model based on Che's [28] approximation (Sec. III-C). CoA is based on a novel decision policy and implies only simple and

uncoordinated operations at the individual caches: as a consequence it is implementable in nodes which require to operate at line-speed, as ICN routers.

- By means of simulation (Sec. IV), we contrast CoA with traditional cost-blind schemes and the cost-optimal solution in simple settings (Sec. IV-C) and additionally evaluate the performance benefits of CoA on realistic network scenarios (Sec. IV-E).
- We analyze how surrounding conditions (link price heterogeneity, popularity skew of contents, object reachability, catalog and cache size, etc.) and internal settings of CoA impact its performance. We verify cost saving to remain consistent across all the above factors (Sec. V).

The rest of the paper is organized as follows. in Sec. II we introduce our model of an ISP and its economic interactions. Sec. III motivates the need of a cost-aware caching mechanism, illustrates and models our proposal. Sec. IV and Sec. V evaluate our mechanism: the former quantifies the achieved saving at a glance and investigates its root cause, whereas the latter is a sensitivity analysis with respect to boundary conditions and internal settings. Related work is presented in Sec. VI. Sec. VII discusses the applicability of our proposal to different caching technologies and to a real deployment, where factors such as latency or internal load must be taken into account. Finally, Sec. VIII concludes the paper.

II. System model

In this section, we first describe a general model of an ISP network that includes aspects related to traffic and economic interactions with other ISPs (Sec. II-A). We then articulate the model to describe the scenarios considered in the reminder of this paper (Sec. II-B).

A. Economic interactions of an ISP

Fig. 1 illustrates the model adopted in this paper: an ISP serves a rate λ_o of requests for a named object o belonging to the catalog $\mathcal{O}$. Tab. I summarizes the notation used throughout this paper. To serve these requests, the ISP may need to retrieve the object through one of its available *external links* (we use the set $\mathcal{L}$ to denote them), paying a related cost.

In case the ISP is operating caches, some of these requests can be served within the ISP network: in this case, the incoming demand λ_o is filtered by caches within the network, so that the demand crossing the ISP boundary for object o is $\lambda_o(1-h_o)$, where h_o is the cache hit ratio for o. The demand for object o flows to a specific external link, and we denote with $\mathcal{O}_i$ the subset of the original catalog $\mathcal{O}$ that is attainable through the external link $i \in \mathcal{L}$. It follows that the load on i is (using unit object size for the sake of simplicity in the formulation):

$$\rho_i = \sum_{o \in \mathcal{O}_i} \lambda_o (1 - h_o). \quad (1)$$

In the current Internet, an ISP can retrieve content from other ISPs, CDNs or Content Providers (CPs) directly connected to the ISP network. As commonly done in the BGP literature [8], [29], we abstract the different types of interactions by distinguishing three categories of links, based on the cost associated to the traffic flow:

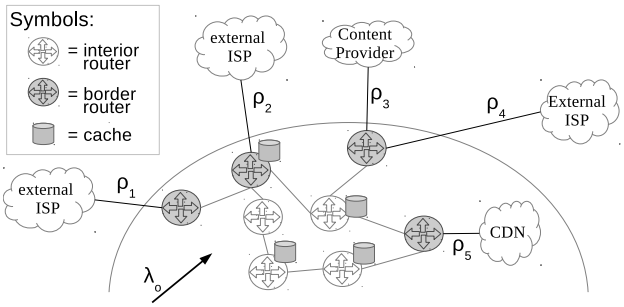


Figure 1. ISP model used throughout this work. The ISP is connected to third party networks through external links having prices π_i, and supporting a total traffic load of ρ_i.

- *Settlement-free peering links* (e.g., connections between ISPs of the same tier) do not imply any economic transaction between the connected ISPs;
- *Provider links* (e.g., transit links to a higher-tier ISP) involve a cost for the ISP, that is typically proportional to some properties (e.g., 95th percentile) of the traffic volume;
- *Customer links* (e.g., links toward lower tier ISPs, or CPs in multihoming [29] or CDNs nodes) imply a revenue[1] for the ISP.

The maximization of the cache hit ratio, irrespective of the link through which the requests exit the ISP network, has usually been the objective of ICN research. In contrast, we argue that the primary goal of an ISP is to minimize the cost associated to external links' utilization. In other words, by installing a limited amount of cache storage within its network, the ISP may not want to blindly maximize the hit ratio independently of the object cost: rather, the ISP aims at caching objects that lead to a larger cost saving, i.e., objects that are accessible through the most expensive links.

Hence, unlike current literature that evaluates the cache vs. bandwidth tradeoff within ISP boundaries [30], we instead do not associate any cost to the traffic on the internal links, as in [10], [11], since we focus on the inter-domain traffic cost, assuming capacities of internal links are sufficient to carry the required traffic, as in [8], [10]–[13]. Moreover, as [10], [13], we do not consider the cost of cache installation, because (i) it is a capital expenditure that is not related to the inter-domain traffic cost, which is the subject of our investigation, and (ii) we start from the assumption that a fixed amount of cache is already installed in the ISP network and we quantify the benefits achievable switching from classic cost-blind cache policies to our proposed cost-aware mechanism, with no difference in the cost of deployment of the cache infrastructure, thanks to the simplicity of our solution.

The 95% charging model is the most widely used among ISPs (see [3]–[5]): traffic volume on a provider link is sampled

[1]For correctness, it is worth specifying that usually CDNs pay ISPs to send them traffic only in case ISPs are sufficiently large. In the other cases, settlement-free agreements are established [9].

Table I
SUMMARY OF THE NOTATION USED IN THIS PAPER.

Variable	Meaning
$\mathcal{L}$	Set of external links
$\mathcal{O}$	Catalog
$\mathcal{O}_i$	Subset of the catalog accessible through link i
Λ	Aggregated request arrival rate
λ_o	Request rate for object o
h_o	Cache hit ratio for object o
ρ_i	Rate of requests crossing external link i
π_i	Price paid by the ISP for every object crossing external link i
π	Price ratio, i.e. the ratio between the expensive and cheap links
π_o	The price of the link which gives access to object o
α	Skew parameter of the Zipf popularity distribution
r_o	Rank of the object o
$\|c\|$	Cache size
$\vec{s} = (s_1, \ldots, s_N)$	Split vector; s_i is the fraction of objects that is behind external link i

every period, e.g. of five minutes, and the 95th percentile of the samples, computed over a larger time span, e.g. one month, is charged. However, as usually assumed in the literature, our traffic model is stationary, i.e. its statistics do not change over periods, and thus the 95% charging model is equivalent to the proportional one, in which the cost incurred in retrieving objects from a certain link is directly proportional to the traffic volume flowing on that link. Therefore, we will use the proportional charging model, conforming to literature (see [4], [6], [8], [10]–[13]). Ultimately, the cost of inter-domain traffic jointly depends on the traffic load ρ_i crossing any given link i and the link price π_i:

$$\sum_{i \in \mathcal{L}} \pi_i \rho_i = \sum_{i \in \mathcal{L}} \pi_i \sum_{o \in \mathcal{O}_i} \lambda_o (1 - h_o). \quad (2)$$

We argue that an interesting objective for ISPs is to minimize the above overall cost (2), considering not only the popularity λ_o but also the link prices π_i, as opposed to maximizing the overall hit ratio $\mathbb{E}[h_o]$ in a cost-blind fashion – that we show being contrasting objectives in Sec. III-A.

B. Network scenario

While the previous subsection provided a general model of an ISP, we need now to give some specific assumptions that allow us to both dissect the tradeoff between hit ratio vs. cost, as well as to enable a thorough and sensible performance evaluation. Since our goal is to study cost, we limitedly consider settlement-free and provider links, that fully covers the different pricing agreements that an ISP may have with its neighboring ISPs through their northbound interface (as shown in Fig. 1).

As commonly assumed in the literature [10], [13], [18], [21], [24], [30], [31], object popularity follows a Zipf distribution having skew parameter α. Denoting Λ the aggregated request arrival rate, we model the request arrival of each object o having rank r_o with a Poisson process of intensity λ_o:

$$\lambda_o = \Lambda \frac{1/r_o^\alpha}{\sum_{j \in \mathcal{O}} 1/r_j^\alpha} \quad (3)$$

We assume that each object is accessible through a single link, thus making $\mathcal{O}_i$ disjoint.[2] We denote with s_i the corresponding fraction of objects $|\mathcal{O}_i|/|\mathcal{O}|$. Throughout the paper we consider a random mapping between objects and links, tunable by varying the breakdown of objects behind each link, i.e., the catalog split vector $\vec{s} = (s_1, \ldots, s_N)$. An important point is worth stressing: clearly, even in case that partitions i, j contain the same number of objects (i.e., $s_i = s_j$), their aggregate request rates differ, as objects have skewed popularity (i.e., $\sum_{o \in \mathcal{O}_i} \lambda_o \neq \sum_{o \in \mathcal{O}_j} \lambda_o$). We cope with this imbalance of the aggregate link load resulting from a catalog split vector $\vec{s}$ by averaging results over multiple runs.

Without loss of generality, let us consider a scenario with three links modeling the following relationships:

- a settlement-free relationship, with price $\pi_{free} = 0$,
- a cheap transit link, with price $\pi_{cheap} = 1$ and
- an expensive link, with price $\pi_{exp} = \pi \geq \pi_{cheap}$

with π a *price ratio* parameter that indicates the ratio between the expensive and cheap link prices. Hereafter, by a slight abuse of language, we will refer to the price π_o of an object o as the price of the link which gives access to it. Consequently, we will refer to free, cheap and expensive objects – despite there is no longer a notion of cost within the ISP boundaries after the object has been retrieved. This price diversity, coupled with the catalog split settings $\vec{s} = (s_{free}, s_{cheap}, s_{exp})$, permits to exacerbate important differences when a fixed cache budget $|c|$ is managed in a cost-blind vs. cost-aware fashions.

III. COST-AWARE ICN DESIGN

In this section, we first provide the rationale behind our proposal, showing that hit ratio maximization and cost reduction are opposite goals (Sec. III-A) and that classic caching strategies focused on the hit ratio may be detrimental in terms of cost. We then illustrate our proposed Cost-Aware caching mechanism (Sec. III-B) and briefly comment on desirable properties of our design (Sec. III-D). For space reasons, we instead refer the reader to [27] for more insights concerning the architectural aspects, including rationales that led to a cost-aware design that exploits the caching function (e.g., as opposed to exploiting routing, forwarding or naming components).

A. The hit ratio vs. cost tradeoff

While generally caching schemes aim at maximizing the hit ratio, our goal is to minimize the cost of inter-domain traffic. Before going into the details of our mechanism, it is worth pointing out that these two goals are conflicting *by nature*. To structurally show this tradeoff, we consider two caching strategies:

- MAXHIT which caches a-priori the $|c|$ most popular objects, i.e. the objects o with the highest λ_o.
- MINCOST which caches a-priori the $|c|$ objects o with the highest $\lambda_o \cdot \pi_o$, where π_o is the price of the external link used to retrieve the object.

[2]While in the real Internet an object can be reachable through multiple links, we suppose that only the one at minimum cost is used, which yields a conservative estimate of the gains achieved by our mechanism.

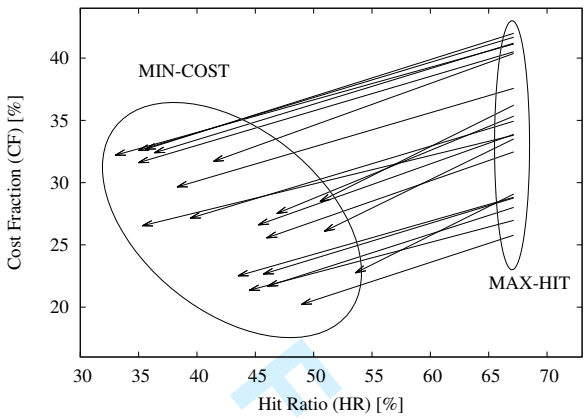


Figure 2. Hit ratio vs. cost trade-off. Values are numerically computed over 20 instances of the default scenario (see Tab. II). Each arrow is relative to a single instance and shows how cost fraction and hit ratio change when switching from MAXHIT to MINCOST.

Note that these two strategies require an a priori perfect knowledge of λ_o, which is not available in a real network: therefore, they are only useful for the purpose of illustration. We show in [14] that MAXHIT is optimal for maximizing the hit ratio, and MINCOST is optimal for minimizing the inter-domain traffic. Since both hit ratio and inter-domain traffic reduction depend on which objects are stored into the network-wide cache space, irrespective of their exact location, the optimality holds independently of the topology.

We compare these strategies in terms of hit vs. cost. The *network-wide hit-ratio* HR^X of strategy X (where $X \in \{\text{MAXHIT},\text{MINCOST}\}$) is the fraction of incoming requests that are satisfied by some cache in the network, computed as:

$$HR^X = 1 - \frac{\sum_{i\in\mathcal{L}} \rho_i^X}{\Lambda} \tag{4}$$

where ρ_i^X is the load on link i, i.e. the rate of requests that are not satisfied by any cache, when using strategy X.

The *cost fraction* CF^X of strategy X is the ratio between the cost incurred by X and the cost incurred by a cache-less system in the same scenario. The cost is computed as the weighted sum of the link load ρ_i times the link price: $\sum_{i\in\mathcal{L}} \rho_i^X \pi_i$, as in (2). In case of a cache-less system, $\rho_i = \sum_{o\in\mathcal{O}_i} \lambda_o$ equals the aggregated arrival rate of the objects in $\mathcal{O}_i$, whereas in the case of a caching mechanism X, $\rho_i = \rho_i^X$ represents the aggregated miss stream, as in (1), so that:

$$CF^X = \frac{\sum_{i\in\mathcal{L}} \rho_i^X \pi_i}{\sum_{i\in\mathcal{L}} \left(\sum_{o\in\mathcal{O}_i} \lambda_o\right)\pi_i} \tag{5}$$

Unless otherwise stated, in what follows we consider a simple yet instructive scenario, where a catalog of $|\mathcal{O}| = 10^5$ objects, having Zipf distributed popularity with skew $\alpha = 1$, are uniformly split between the free, cheap and expensive links $s_{free} = s_{cheap} = s_{exp} = 1/3$, with a ratio between expensive and cheap link prices $\pi = 10$. The ICN network has an overall cache budget of 10^3 objects and is modeled, for the time being, as a single cache – in this way, we avoid to jointly evaluate CoA and ICN routing, that we instead consider later in Sec. IV. For the time being, we are also interested in a relative comparison of MAXHIT vs. MINCOST, as opposed to a precise assessment of their absolute performance – while parameters of the default scenario are carefully set, we defer this viewpoint to the thorough sensitivity analysis reported in Sec. V.

The MAXHIT vs. MINCOST trade-off is illustrated in Fig. 2. As previously observed, there is an inherent variability across instances of the same scenario, which is tied to the different breakdown of the objects among external links in each instance. Results clearly show that cost fraction and hit ratio are conflicting goals: specifically, the arrow implies that a cache fraction loss is necessary to achieve a cost reduction gain in the corresponding random instance of the default scenario.

B. Design of Cost-Aware (CoA) decision policy

We propose a novel Cost-Aware (CoA) design to achieve significant cost reduction, that we illustrate with the help of Fig. 3. Simply speaking, any new object arriving at an ICN router is either cached or discarded, according to a *decision* policy; in the first case, a *replacement* policy is triggered to select a previously cached object to be evicted. We inject cost-awareness in the decision policy. The motivations behind this choice will be clearer after having described our design and will be discussed in Sec. III-D.

Intuitively, to reduce costs, a cache has to not only store the most popular objects (which results in *caching efficiency*) but also and especially those that are obtained through the most expensive links (which results in *cost reduction*). Otherwise stated, the aim of cost-aware caching is to bias the caching process toward more expensive objects. However, it is not to be forgotten that, beyond the price of individual links, content popularity still plays a paramount role. Indeed, popularity and cost factors are independent and may even conflict: e.g., caching expensive but unpopular objects may not bring effective cost reductions while, on the other hand, caching cheap but very popular objects may be worthwhile. Therefore, our goal is to consider price differences, but still differentiate between popular and unpopular objects.

For this purpose, we design a modular decision policy, which is the composition of a popularity-based module and a price-based one, represented by the functions $\psi(\cdot)$ and $\beta(\cdot)$, respectively. The composition of the two modules is achieved via product of the two functions, i.e. a new object is accepted with probability $\psi(\cdot)\beta(\cdot)$. This composition permits to jointly weight popularity and price. While $\psi(\cdot)$ can be any of the classic decision policies in the literature, we design the function $\beta(\cdot)$ whose specific role is to weight price, biasing the acceptance toward expensive objects, as follows:

$$\beta(o) = M \cdot \pi_o^\kappa / \sum_{i\in\mathcal{L}} \pi_i^\kappa \tag{6}$$

where π_o is the price of the external link through which the new object o crossed the ISP boundaries. The parameters M and κ have the following meaning:

- The constant M is set such that the average value $\mathbb{E}[\beta(o)]$, computed over all the new objects passing through the

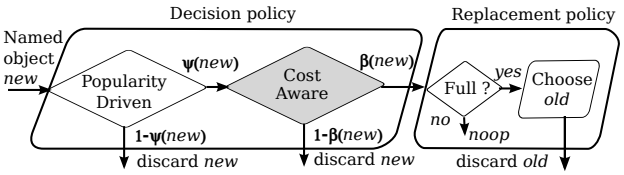


Figure 3. Cost-aware ICN design, plugged within the decision decision policy of the caching component.

cache, is 1. This guarantees that $\mathbb{E}[\psi(\cdot)\beta(\cdot)] = \mathbb{E}[\psi(\cdot)]$, i.e., the average acceptance ratio is not modified by function $\beta(\cdot)$. Otherwise stated, the cache accepts, on average, the same fraction of objects as without the cost-aware module β, with the only difference that it preferentially stores the expensive ones. Additionally, this ensures that convergence rates of Unif and CoA are the same. In the single cache case, the normalization factor can be computed as:

$$M = \sum_{i \in \mathcal{L}} \pi_i^\kappa / \sum_{i \in \mathcal{L}} s_i \pi_i^\kappa \tag{7}$$

- The exponent $\kappa > 0$ is used to tune the relative importance of popularity vs. price in the decision: indeed, the larger κ, the larger the skew toward expensive objects, while for $\kappa < 1$ the importance of price in the decision diminishes.

We observe here that classic decision policies approximate MAXHIT (Sec. III-A), trying to infer λ_o by means of function $\psi(\cdot)$ in order to cache the (locally) more popular objects. CoA approximates MINCOST, trying to infer $\lambda_o \pi_o$ by means of composition $\psi(\cdot)\beta(\cdot)$ in order to cache the objects that would generate the highest expenditure. In the following, we will consider a constant popularity based function $\psi(o) = \psi_0, \forall o \in \mathcal{O}$. Notice that in this case, by plugging (7) into (6) we can rewrite the CoA function as $\psi(o)\beta(o) = K\pi_o^k$, with $K = \psi_0 / \sum_{i \in \mathcal{L}} s_i \pi_i^\kappa$ a constant that depends on both uniform probabilistic decisions (numerator) as well as on object cost (denominator).

C. Model of Cost-Aware (CoA) decision policy

We now provide a simple model of CoA by extending the analysis of the *Unif* policy provided by [32] that is itself based on Che's approximation [28]. To this aim, we first restrict our attention to the subset $\mathcal{O}' \subseteq \mathcal{O}$ of objects having a chance to be cached, i.e. the objects whose price is non-zero, ignoring thus all the objects retrievable through a free link. By definition, the probability that CoA accepts an incoming object $o \in \mathcal{O}'$ in the cache is $\psi(o)\beta(o) = K\pi_o^k$. Considering a single *CoA* cache of capacity $|c|$ and whose incoming requests respect the Independence Reference Model (IRM), the hit probability for object an $o \in \mathcal{O}'$ is:

$$\mathbb{E}(h_o) = \frac{K\pi_o^k \cdot (1 - e^{-\lambda_o T_{|c|}})}{e^{-\lambda_o T_{|c|}} + K\pi_o^k \cdot (1 - e^{-\lambda_o T_{|c|}})} \tag{8}$$

$$= \frac{1 - e^{-\lambda_o T_{|c|}}}{1 - e^{-\lambda_o T_{|c|}} \left(1 - \frac{1}{K\pi_o^k}\right)} \tag{9}$$

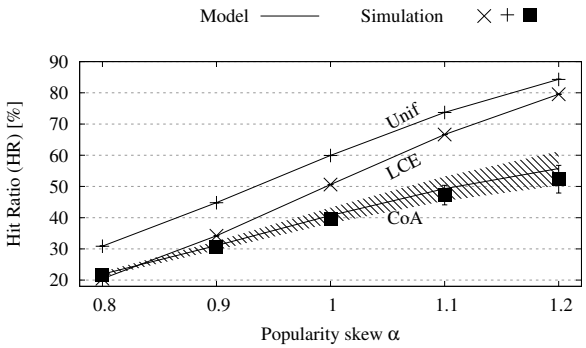


Figure 4. Model vs. simulation. Average hit ratio values over 20 instances of the default scenario (see Tab. II) and 95% confidence intervals are depicted. The average acceptance ratio for both Unif and CoA is 1% and $\kappa = 1$.

where the characteristic time $T_{|c|}$ is computed as in [28] by imposing that $\sum_{o \in \mathcal{O}'} \mathbb{E}(h_o) = |c|$. Notice that (9) degenerates into the original Che's approximation $\mathbb{E}(h_o) = 1 - e^{-\lambda_o T_{|c|}}$ for $\kappa = 0$ (i.e., when cost information is ignored). Clearly, for objects $o \in \mathcal{O} \setminus \mathcal{O}'$ retrievable through a free link, we instead have $\mathbb{E}(h_o) = 0$. The overall hit-probability, i.e. the expected hit ratio, can then be obtained over the whole catalog as:

$$\mathbb{E}(H^{CoA}) = \sum_{o \in \mathcal{O}} \frac{\lambda_o}{\Lambda} \cdot \mathbb{E}(h_o) \tag{10}$$

whose numerical solution is depicted in Fig. 4 alongside that of LCE [28] and Unif [32] models. As for LCE and Unif, comparison against simulation exhibit an excellent match (for CoA, we additionally remark for both model and simulation the variability tied to the catalog split early noted in Fig. 2).

D. Implementation considerations

Decision vs. Replacement policies. We motivate now why we introduce cost-awareness in the decision policy rather than in the replacement policy. First, a properly tuned decision policy avoids the proliferation of irrelevant content along multiple caches, which would happen in case any new content were systematically accepted in the cache (Leave a Copy Everywhere, LCE) and which would lead to an excessive number of repeated evictions. Therefore, deterministic [32], [33] or probabilistic [17], [18], [32], [34] decision policies are preferable. By extension, it is better to bias the acceptance toward more expensive objects in the cache, than to bias the *replacement* process toward cheaper objects a posteriori: in the latter case, each router should keep state of cached objects (i.e., additional price metadata), that it would need to manage at line speed (e.g., perform complex computations that take into account the price of all the cached objects, to select the cheapest one to evict). On the contrary, a cost-aware *decision* strategy, like the one we propose, is simpler to implement since it is lightweight and stateless (as price-related information can be carried in the packet header by the ISP border router once, and exploited independently by any router along the path), allowing the rest of per-object operation to remain simple (e.g., Least Recently Used or random eviction policies).

Implicit distributed coordination. In order to efficiently exploit the total cache budget contained in a network of caches, a form of coordination is required, to prevent, for example, a node to store an object already stored by some other neighbors. Contrarily to mechanisms realizing explicit coordination by message exchange over the control plane [19], [20], which has the downside of complexity and communication overhead, our mechanism achieves distributed coordination with implicit coordination. In other words, in our approach no information is exchanged over the control plane, but rather a minimum amount of information –i.e., a price indication– is carried via packets header directly in the data plane.

In practice, only *border routers* know the link through which objects enter the ISP domain, and can thus (i) tag the packet with a price indication; (ii) additionally, in case they are equipped with storage components, border routers take a caching decision according to $\psi(\cdot)\beta(\cdot)$ prior to forwarding the packet. *Interior routers* along the path then (iii) take independent caching decisions based on the price information tagged by border routers, and by any other information (e.g. centrality, distance), which possibly differs among routers. This price indication represents a negligible overhead, since it is marked only once and it travels together with the object, requiring the modification of only few bits of the header (as we have already shown in [14] and avoid reporting here).

No additional cost. From the simplicity of our design, it follows that deploying CoA does not imply higher installation and operation costs than a classic caching policy (e.g., LCE+LRU). Therefore, the whole saving in the operational costs comes for free, i.e., it does not require an increase in the capital expenditure, as is often the case.

IV. Benefits of cost-aware design

We now assess the benefits of our proposed cost-aware design against cost-blind and cost-optimum ICN strategies. On the one hand, comparison with cost-blind ICN schemes can be viewed as a direct measure of the return of investment following ICN deployment, and more precisely sizes the additional gain that can be attained by a cost-aware architecture. On the other hand, comparison with the optimal cost strategies allows us to gauge the extent of possible improvements in our design.

In this section we define the classic strategies that we contrast with CoA (Sec. IV-A), and the evaluation metrics (Sec. IV-B). We start by considering the default scenario to cross compare, at a glance, all the above strategies (Sec. IV-C). We next expose the deficiencies of cost-blind strategies (Sec. IV-D) and finally verify that the CoA saving is consistent over real ISP topologies as well as in synthetic topologies generated with the Watts-Strogatz model (Sec. IV-E).

A. Terms of comparison

We contrast our design against several terms of comparison, that represent (i) cache-less systems, (ii) traditional ICN schemes where price heterogeneity is not taken into account, (iii) ideal distributed decision policies with perfect knowledge of object popularity and (iv) MinCost achieving provably minimum cost. As the following table illustrates, these different designs provide an exhaustive coverage.

	Cache-less	LCE	Unif	Ideal-Blind	CoA	Ideal-CoA	MIN-COST
Cost-aware				✓	✓	✓	✓
Implementable	✓	✓	✓		✓		

Cache-less system. As naive benchmark, we consider costs incurred by systems that do not employ any kind of caching. We point out that, other than providing an upper-bound of the costs incurred by the system, considering a common reference significantly simplifies the assessment of the relative improvement between more sophisticated strategies.

Cost-blind ICN. Following our design, a natural term of comparison for cost-blind ICN consists in considering state-of-the-art decision policies that ignore the cost of inter-domain traffic (i.e., equivalent to setting $\beta(\cdot) = 1$). The popularity-driven decision component could use:

- Leave a Copy Everywhere (*LCE*), equivalent to setting $\psi(\cdot) = 1$.
- Leave a Copy Down (*LCD*) [33], accepting new items only when they have traveled $d = 1$ hop in the network, expressed with the Dirac delta function $\psi(\cdot) = \delta(d-1)$.
- Uniform probabilistic decisions (*Unif*) [34], where $\psi(\cdot) = \psi_0$, ψ_0 being a fixed probability.
- Decisions based on distance [18], graph properties [17], correlation between consecutive requests [32], etc.

In this work, to avoid cluttering the pictures, we limitedly consider a *Unif* policy, that is known to tend asymptotically to MaxHit for $\psi_0 \to 0$ under Independent Reference Model (IRM) [32] and is known to provide good enough results even when compared to more complex policies [26]. It follows that *Unif* is thus a reasonable term of comparison, representative of state-of-the-art cost-blind decisions. As a side effect, comparison of CoA and cost-blind Unif strategies can be done on a fair ground, i.e., on the same number of acceptance decisions as stated earlier.

Ideal strategies. We additionally consider two strategies that have perfect knowledge of global object popularity, and that thus constitute an ideal term of comparison on the single-cache scenario. Yet, we point out that since this knowledge is not available in real situations, these policies cannot be implemented and are introduced here only as benchmarks.

Specifically, the decision whether to cache or not a new object is assisted by considering the eviction candidate, i.e. the object that would be removed from cache to make room for the new one. We assume the replacement policy is Least Recently Used (LRU), and thus the eviction candidate corresponds to the object that was requested least recently. The new object is accepted only if it is more "valuable" than the eviction candidate. This is expected to increase the value of the overall cache content over time. We implement two notions of value, depending on whether they limitedly consider object popularity, or jointly consider popularity and link price.

The ideal cost-blind strategy (*Ideal-Blind*) strives to keep only the most popular objects, deterministically admitting a new object o only if its arrival rate λ_o is greater than the one of the LRU eviction candidate.

The ideal cost-aware strategy (*Ideal-CoA*), instead, jointly considers the arrival rate and the price of the link through which the object has to be fetched. The aim is clearly to cache only the objects that are expected to provide the largest saving, which happens by admitting only objects whose $\lambda_o \pi_o$ is larger than that of the eviction candidate.

Optimal. We finally consider the minimal cost incurred by the ISP, obtained via the MINCOST strategy. As opposed to the ideal strategies mentioned earlier, which take decisions on each packet arrival and are obtained via discrete event simulation, the MINCOST strategy is obtained via a centralized optimal solution. A further difference between *Ideal-CoA* and MINCOST is that MINCOST basically pre-fills caches, so that it provides a lower bound to the ISP expenditures.

B. Settings and metrics

To gauge the advantages introduced by CoA, we introduce two metrics beyond the cost fraction CF (5). Specifically, denoting $Cost^X = \sum_{i \in \mathcal{L}} \rho_i^X \pi_i$, we indicate with *Potential Saving* (PS) the room for improvement of our proposal, i.e., the percentage of additional saving that could be leveraged by switching to an *Ideal-CoA* policy:

$$PS = \frac{Cost^{CoA} - Cost^{Ideal-CoA}}{Cost^{CoA}} \tag{11}$$

We further indicate with *Achieved Saving* (AS) the percentage of expenditure which an ISP, currently running the state-of-the-art *Unif* policy, could save by switching to CoA:

$$AS = \frac{Cost^{Unif} - Cost^{CoA}}{Cost^{Unif}} \tag{12}$$

To perform a conservative evaluation, we need therefore to set the probability ψ_0 in *Unif*, to avoid overestimating the achieved saving. While [32] proves that *Unif* tends asymptotically to MAXHIT for $\psi_0 \to 0$, however care should be used when applying this theoretical result to a real scenario: indeed, for very small values of ψ_0, *Unif* suffers from a slow convergence in learning object popularity. In other words, *Unif* caches objects only after they are observed many times ($1/\psi_0$ times on average), and thus there is a long transient dominated by cache misses, which cannot be neglected as it entails a non-negligible cost for the ISP for very low ψ_0. Starting from these considerations we perform a preliminary calibration and identify in $\psi_0 = 1/100$ a value that is favorable to *Unif* in our scenarios, that we fix for the reminder of this work.

We point out that, since $\mathbb{E}[\beta(o)] = 1$ by design (see Sec. III-B), the average acceptance ratio is $\mathbb{E}[\psi(o) \cdot \beta(o)] = \mathbb{E}[\psi(o)] = \psi_0 = 1/100$ in both Unif and CoA. This ensures a *fair comparison*: indeed, the differences in performance cannot be ascribed to a different average cache admission probability, but are only due to cost-awareness, which is the main object of our investigation. Additionally, as the number of cache acceptance decisions taken by Unif and CoA is the same, their convergence speed is the same, despite the attained saving is different.

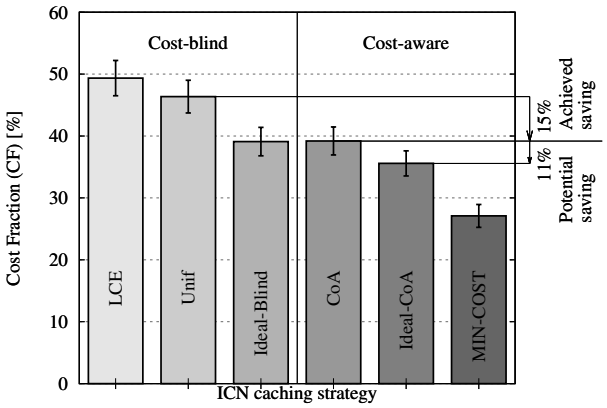


Figure 5. Benefits of cost-aware design. The cost fraction (5) obtained by each strategy is reported. Achieved and potential saving (expressions (12) and (11), respectively) are annotated on the right y axis.

C. Comparison at a glance

In this section, we still refer to the single cache scenario considered earlier (detailed values are highlighted with boldface in Tab. II), setting $\kappa = 1$ and $\psi_0 = 1/100$. With the exception of the MINCOST solution, that we compute numerically, all strategies are implemented in ccnSim, an efficient and scalable open-source ICN simulator that we make available at [35] along with simulation scripts to reproduce the results of this paper.

In the following we report the average results with 95% confidence intervals gathered from 20 runs for each setting; the duration of each run is sized to have statistically relevant results, and statistics are computed only after the initial transient period needed for the cache hit metric to reach a steady state. To evaluate the cost-effectiveness achieved by a caching strategy X, we compute in each scenario the hit ratio (HR^X) and the cost fraction (CF^X) as in (4) and (5), with X being any of the strategies introduced earlier (i.e., LCE, Uniform, CoA, Ideal-Blind, Ideal-CoA, MINCOST).

Fig. 5 shows, at a glance, the cost fraction for cost-blind (left bars) and cost-aware (right bars) strategies. Our strategy (CoA) brings sizable benefits over state-of-the-art cost-blind decision (about 15% of achieved saving over Unif), matching the performance of the Ideal-Blind strategy. This means that, exploiting information already at hand, and that changes over relatively long timescales (i.e., the prices negotiated with different ISPs), can bring benefits that are at least as important as those relative to information that is highly volatile and harder to infer (e.g., object popularity).

To interpret the practical relevance of the CoA benefits, consider the case of an ISP in which a state-of-the-art ICN caching system is already deployed, which is tuned in a cost-blind fashion to maximize hit-ratio. If the ISP decides to switch to CoA tuning, it will save about 15% of the inter-domain traffic cost, without facing any additional expense. Indeed, while the installation of the ICN infrastructure implies

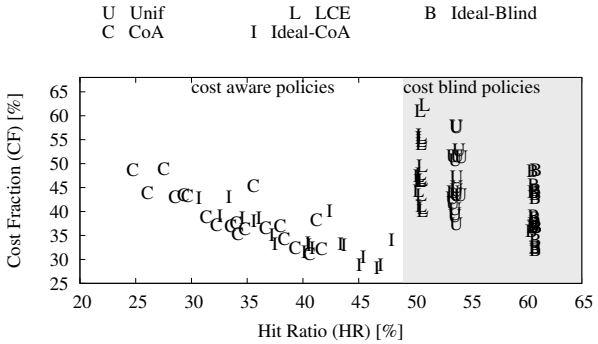


Figure 6. Comparison of cost-aware vs. cost-blind policies: Scatter plot of hit ratio versus cost fraction, confirming that higher cache hit ratio does not necessarily imply lower cost under a wider range of policies.

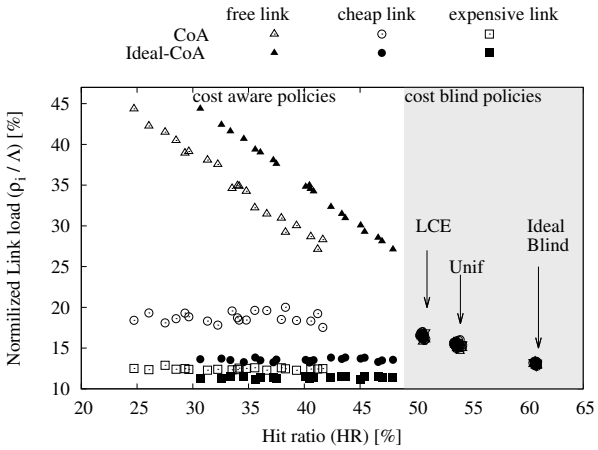


Figure 7. Scatter plot of hit ratio vs. load (normalized over the aggregate request arrival rate) on free, cheap and expensive links. Note that cost-aware policies differentiate loads on links with heterogeneous prices.

a capital expenditure (CAPEX), our CoA mechanism consists in a simple tuning and does not require additional capex. Yet, CoA offers the ISP a consistent saving in the operational expenditure (OPEX), that becomes sizable as it accumulates over the years.

At the same time, considering the distance from Ideal-CoA to MINCOST, we see that there is still additional room for improvement (11% of potential saving), which is however hard to reach, as it would require knowledge of object popularity.

D. Root cause of cost saving

To understand the root cause of the performance gap, we extend the previous representation of the MAXHIT vs. MINCOST tradeoff depicted in Fig. 2, to include the LCE (L), Uniform (U), CoA (C), Ideal-Blind (B) and Ideal-CoA (I) policies, that we represent with a capital letter in the scatter of Fig. 6. We generate 20 instances of the default scenario and run the different strategies on each instance. We observe that, despite the low hit ratio, cost-aware policies result in a lower cost fraction: this confirms that cost reduction does not only come from cache hit maximization, but is mainly due to price discrimination. This shows also that the tradeoff discussed when considering optimal strategies (Sec. III-A) also holds in practical implementations. Similarly to Fig. 2, the dispersion in Fig. 6 is caused by the object-to-link mapping randomly generated for each instance of the scenario.

To further assess the impact of cost-aware caching on the network, in Fig. 7 we report the *normalized traffic load* of the free, cheap and expensive links, i.e. the load (as in (1)) divided by the aggregated request rate, Λ. CoA and Ideal-CoA achieve structurally similar configurations; specifically, they reduce the load on expensive and cheap links (circles and squares in the figure), while increasing the load on the free link (triangles), since they both never cache free objects. Note that as the hit ratio decreases, the load on the free link increases while the loads on the cheap and expensive links are almost constant: this means that all the additional miss stream drains into the free link. Finally, Ideal-CoA exhibits better performance than CoA, in terms of both hit ratio and cost fraction, due to the perfect knowledge of object popularity.

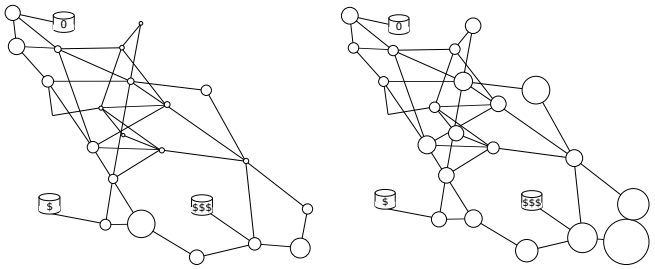
While cost-aware policies differentiate link load based on link prices, cost-blind policies uniformly distribute the load, resulting in overlapping points in the scatter plot. Note that, while reasonable, this result is not straightforward and is due to the cache filtering effect: in other words, despite the load in a cache-less scenario would not be uniform due to the variability of the aggregated demand in each sub-catalog, however, the cache equalizes the miss-stream over these links. This is intuitive, since in a uniform scenario, links with higher demand (before caching) are those behind which the most popular objects are accessible, thus, they will be most affected by load reduction due to caching.

To summarize, the price differentiation operated by cost-aware policies permits to cache only the objects that would result in a cost for the operator. This has two consequences: (i) it reduces cache efficiency in terms of hit ratio but, on the other hand, (ii) it limits ISP costs thanks to the diminished utilization of the costly links.

E. Performance on realistic network topologies

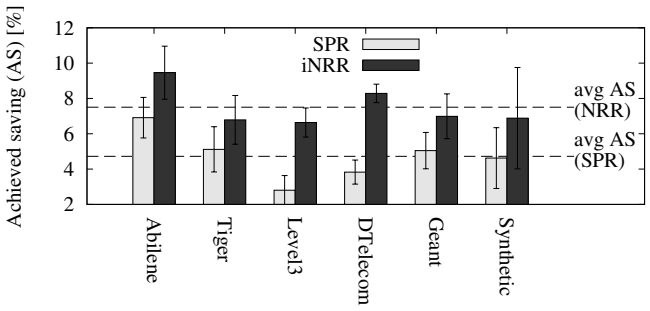
So far, we have analyzed the performance of a single cache operating with CoA. In this section, we show that cost reduction is consistent even in a distributed environment, consisting in a network of caches, each operating autonomously with CoA. We conduct a simulation campaign on both realistic (as in [23], [24]) and synthetic (as in [25]) network topologies (of which an example is depicted in Fig. 8 and that are described in Fig. 9) where, at each run, we attach the free, cheap and expensive links to randomly selected nodes. We allocate the total cache space uniformly among all routers (as in [24], [26]) and use the default values for the other parameters (bold values of Tab. II). We consider two forwarding strategies:

- *Shortest Path Routing (SPR)*: the path traversed by a request is the shortest between the origin and the egress router. The egress router is the one attached to the external link which gives access to the requested object.
- *ideal Nearest Replica Routing (iNRR)* [24]: if there exists a cache that is storing the requested object, and it is closer than the egress router, the request is sent to that cache.



(a) Unif (b) CoA

Figure 8. Geant topology. Node size is scaled based on their contribution to the overall saving when Unif (a) or CoA (b) is used. Objects are retrievable through the links connected to the nodes labeled with 0, $ and $$$, which are free, cheap and expensive, respectively.



#nodes	11	22	46	68	22	22
Avg node degree	2.5	3.6	11.7	10.4	3.4	4.0
Coeff. of variation of degree	0.2	0.2	0.9	1.3	0.4	0.4
Avg link propagation delay[ms]	11.3	0.1	8.9	17.2	2.6	2.6
Diameter	8	5	4	3	4	4

Figure 9. Achieved saving on different topologies, ordered from the smallest (with the fewest nodes) to the largest one. 95% confidence intervals are reported. The table reports the characteristics of the topologies.

With SPR, an interest can be matched only with the copies cached in one of the nodes along the shortest path. Therefore, a content may be downloaded through an external link even if a copy is present inside the network, which happens whenever the cached copies lay off the shortest path between the requestor and the repository. Due to the increased redundancy, and reduced efficiency in using a fixed cache space budget, we expect cost reduction in the SPR case to be smaller than that estimated in the previous section on a single case scenario.

This limitation is overcome by iNRR [24], which is able to exploit all the copies stored in the network. Even though iNRR is ideal, since it would require the knowledge of the objects cached in all the nodes, it can be easily approximated [26] in ICN and is thus worth considering. Opposite to the SPR case, we expect the iNRR cost reduction to be in line with that estimated in the single cache case.

In complex topologies, interesting mutual effects among nodes arise, whereas they are not observable when considering a single node. In particular, in the distributed case there is a mismatch between the *global popularity* of any object vs. its *local popularity* which accounts only for the requests received for that object by a specific node. In particular, the local popularity of an object observed by a node depends on (i) the routing policy, since not all the requests pass through that node and (ii) the cache filtering effect, due to cache hit at neighboring nodes. It follows that Ideal-Blind and Ideal-CoA policies are not effective in these scenarios, as they base their decisions on global object popularity.

We therefore exclude Ideal policies from the analysis, and limit our attention to comparing policies (namely, probabilistic cost-blind vs. cost-aware), under two routing schemes (namely, SPR vs. iNRR) on a range of topologies (namely, real vs. synthetic). Fig. 9 illustrates the saving obtained in five real topologies and synthetic topologies, generated with Watts-Strogatz model, matching Geant's characteristics. The achieved saving amount to 7.5% (4.7%) on average with iNRR (SPR). These results suggest that the achieved reduction is consistent even with realistic networks of caches, the size of the reduction depending on the topology. The advantages of CoA over Unif also depend on the forwarding strategy and are more evident with iNRR, as expected. Another interpretation of the SPR vs. iNRR performance gap can be given anticipating that our sensitivity analysis shows gain to grow proportionally to the cache space (see Fig. 14 in Sec. V-E): under this light, the gap follows from the fact that interests can leverage all the cache space under iNRR, while only the fraction of cache space included in the shortest path is exploitable with SPR.

Finally, in order to quantify the contribution of each node to the overall saving under the Unif vs CoA policies, we define the value of a cache node n as $v_n = \sum_{o \in \mathcal{C}_n} \lambda_o \cdot \pi_o$, where $\mathcal{C}_n$ is the set of objects stored by n, which is clearly the cost absorbed by that node. Fig. 8 depicts the (rescaled) cache value averaged over 20 simulation runs. Note that, when Unif is used, the majority of nodes has a small cache value: on the contrary, CoA tends to equalize cache values, allowing each node to give a substantial contribution to the overall saving (despite topological constraints, e.g., nodes being closer to more valuable content, still have a clear impact).

V. Sensitivity Analysis of Cost-Aware Design

The previous sections have delved into benefits of cost-awareness into a sensible yet specific scenario. We now extend the reach of the above findings by showing that CoA benefits are robust and consistent in a wide range of conditions – Overall, we performed over 4000 simulation runs, accounting for $O(10^{10})$ requests.

Specifically, we perform a sensitivity analysis of scenario parameters that are external and, in Sec. V-F, we show benefits to be smoothly varying with respect to internal CoA knobs, such as the κ parameter. We anticipate that our proposed Cost-Aware scheme provides a consistent and robust saving in all the considered network scenarios.

For what concerns evaluation scenarios, there are many factors that are unknown at best, which will likely change in unpredictable manner, and that are not under the control of either the manufacturers or the ISPs. We therefore perform

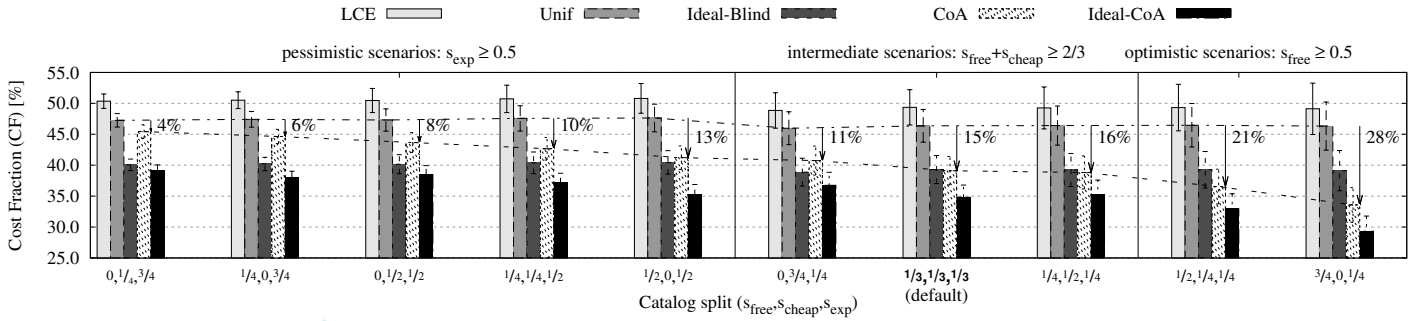


Figure 11. Impact of the catalog split: $(s_{free}, s_{cheap}, s_{exp})$ represent the probability that a content is retrievable through the free, cheap or expensive link, respectively. Arrows indicate the saving achieved by CoA over Unif.

Table II
PARAMETERS OF THE SCENARIO. BOLD VALUES REPRESENT THE DEFAULT SCENARIO USED THROUGHOUT THE PAPER.

Parameter	#	Values
Zipf skew α	3	0.8, **1**, 1.2
Price ratio π	5	1, 2, 5, **10**, 100
Catalog split $\vec{s}$	13	$s_i \in \{\mathbf{1/3}, h/4 \vert h \in \{0,1,2,3,4\}\}$ $\sum_i s_i = 1$
System scale $\vert c\vert/\vert\mathcal{O}\vert$	5	$10^2/10^4, \mathbf{10^3/10^5}, 10^4/10^6,$ $10^5/10^7, 10^6/10^8$
Cache/catalog ratio $\vert c\vert/\vert\mathcal{O}\vert$	5	$\mathbf{10^3/10^5}, 10^3/10^6,$ $10^3/10^7, 10^3/10^8$

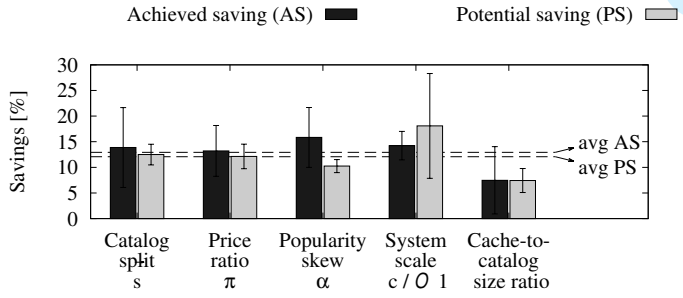


Figure 10. Robustness against external factors such as catalog split, price ratio, popularity skew, system scale, cache-to-catalog size ratio. Bars represent average and standard deviation of the achieved and potential saving over the full parameter space reported in Tab. II.

a thorough sensitivity analysis of the CoA performance on scenarios other than the default one investigated earlier. Tab. II reports the parameter values we consider in this section. For the sake of simplicity, since CoA performance under state-of-the-art iNRR routing is consistent with that of the single cache scenario, in this section we limitedly consider the latter.

Clearly, each parameter concurs in determining the CoA performance: e.g., we expect the achieved saving to be marginal for very low skew values (α), or when most of the catalog is accessible only through the most costly link, or when the cache is too small, etc. The impact of these parameters is summarized in Fig. 10, which represents the mean value of the achieved vs. potential saving and their standard deviation obtained by making the single parameters vary within their respective domains. We see that the gains resulting from biasing the cache decision policy along the cost dimension are consistent over all the parameter variations: on average, the achieved saving over Unif is 13%.

Hereafter, we investigate how each single parameter of the scenario impacts the CoA performance. If not otherwise stated, each configuration is obtained starting from the default one (see bold values in Tab. II) and varying only the parameter under analysis. Each configuration is evaluated providing the mean value of saving and the 95th percentile over 20 runs.

A. *Impact of catalog split*

At each simulation run, we place each object behind one of the three external links (free, cheap or expensive), on a probabilistic basis with s_i the probability that an object is assigned to link i. As a consequence, the catalog is split into free, cheap and expensive objects: we distinguish *pessimistic* scenarios in which at least half of the catalog is behind the expensive link, *optimistic* scenarios in which at least half of the objects are free, and *intermediate* scenarios.

Fig. 11 represents the cost fraction in 10 different scenarios, each characterized by a catalog split vector $\vec{s}$. The cost saving achieved by CoA over Unif is reported besides the arrows. As expected, cost-blind policies are insensitive to the catalog split, since they treat objects as they all had the same value. This is why in Fig. 11 their cost fraction is constant (with the exception of a slightly higher cost fraction in the pessimistic scenarios, that is due only to the fact that they are less favorable). On the contrary, the impact on cost-aware policies is evident. CoA performs better when a considerable part of the catalog is free or cheap: in this case, the achieved saving goes from 8% up to 27%. When half of the catalog is behind the expensive link, cost reduction is more modest, and this is due to the fact that there are inherently no gains to be exploited.

B. *Impact of price heterogeneity*

The price ratio $\pi = \pi_{exp}/\pi_{cheap}$ is the ratio of the expensive over the cheap link princes: the larger is π, the higher is the heterogeneity of the external link prices. We consider values of

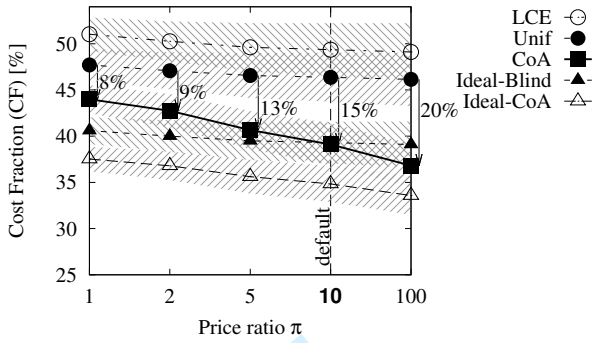


Figure 12. Impact of price heterogeneity on cost fraction. Arrows indicate the saving achieved by CoA over Unif.

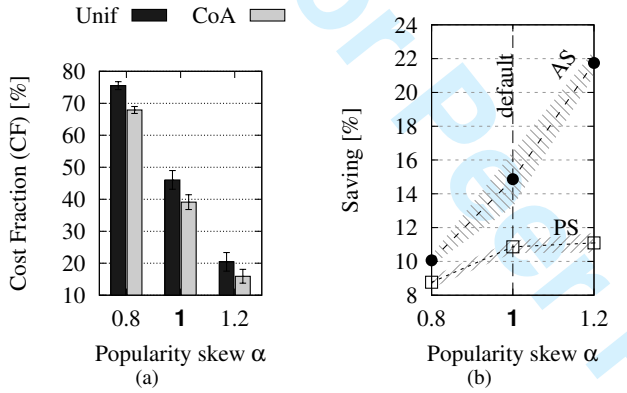


Figure 13. Impact of popularity skew: (a) Cost fraction of Unif and CoA and (b) achieved (AS) and potential (PS) saving.

price ratios ranging from 1 to 10, in line with values reported by both [4] and [15], who gathered information by publicly available data and by interviews with operators, respectively. We plot the cost fraction in Fig. 12, where the arrows report the achieved saving of CoA over Unif. For $\pi = 1$, cheap and expensive links have the same price: therefore, cost-aware policies achieve cost reduction only by avoiding to cache free objects, while the other policies tend to blindly cache them. The cost reduction of cost-aware policies becomes more evident as price heterogeneity increases: while cost-blind policies are insensitive to price ratio, cost-aware ones leverage it. Contrarily to [7], we argue that to reduce the ISP costs is not sufficient to blindly reduce inter-ISP traffic across external links, since price heterogeneity plays an important role and must be exploited. In order to depict an asymptotic behavior of the cost saving, we include in our analysis a price ratio of 100, showing that, already for practical $\pi = 10$ values, our CoA proposal gets most of the asymptotic benefits.

In addition, we observe that, for high price ratios, CoA equals or outperforms Ideal-Blind. It is interesting to underline that this holds even if CoA requires only the knowledge of the objects price (which changes very slowly in time and is easily traced by ISPs, and thus of practical use) as opposed to Ideal-Blind that requires a perfect and a priori knowledge of the popularity (which changes rapidly and is very difficult to infer properly, and thus impractical to exploit).

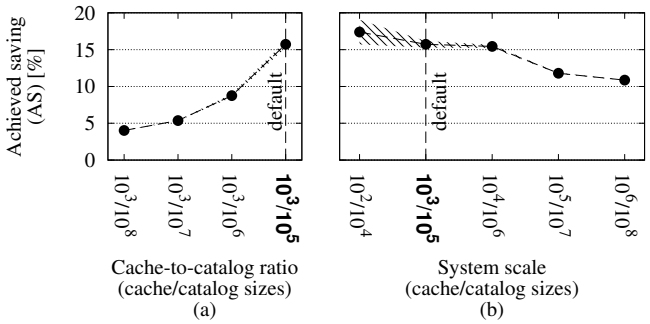


Figure 14. Impact of system scale and cache-to-catalog size ratio: Achieved saving of CoA over Unif.

C. Impact of popularity skew

We study how cost reduction is impacted by the popularity skew of the catalog. We let the Zipf exponent α vary along the range of values that are reported in recent work employing measurement from either a global CDN [24] or a local PoP of an ISP [36]. As expected, increasing the popularity skew plays in favor of caching, i.e. both Unif and CoA reduce their cost fraction, which can be seen in Fig. 13-(a).

Nonetheless, CoA consistently outperforms Unif. Indeed, even if the cost fraction of Unif decreases for an increasing skew, the CoA saving over Unif increase further: this clearly emerges from Fig. 13-(b), and is due to the fact that the denominator of the achieved saving (12) becomes smaller. Additionally, we see that the potential saving saturates after $\alpha > 1$, meaning that CoA is able to efficiently take advantage of the favorable conditions to caching represented by the high popularity skew.

D. Impact of cache-to-catalog size ratio

We next verify gain dependency on the relative scales of the cache vs. catalog sizes. We fix the cache size $|c| = 10^3$ and make the catalog size vary in $\{10^8, 10^7, 10^6, 10^5\}$, thus getting, respectively, cache-to-catalog size ratios varying in the 0.01% to 1% range, in line with [24], [26], [32].

Results reported in Fig. 14-(a) show that the reduction achieved by CoA increases with the cache-to-catalog size ratio: this means that the larger is the cache budget available for the ISP, the more attention is worth paying to its management, as the attainable cost saving is larger. The iNRR gain over SPR routing policies is partly explained by the same observation.

E. Impact of system scale

Finally, maintaining the cache-to-catalog size ratio size fixed to the default value $|c|/|\mathcal{O}| = 10^{-2}$, we change the scale of the simulation by varying simultaneously the cache size and the catalog size. The considered catalog sizes are representative of content providers of different dimension, as Video on Demand services or Youtube, and are based on previous work in literature [37], [38]. Results summarized in Fig. 14-(b) show that the achieved saving diminishes as the scale increases: yet,

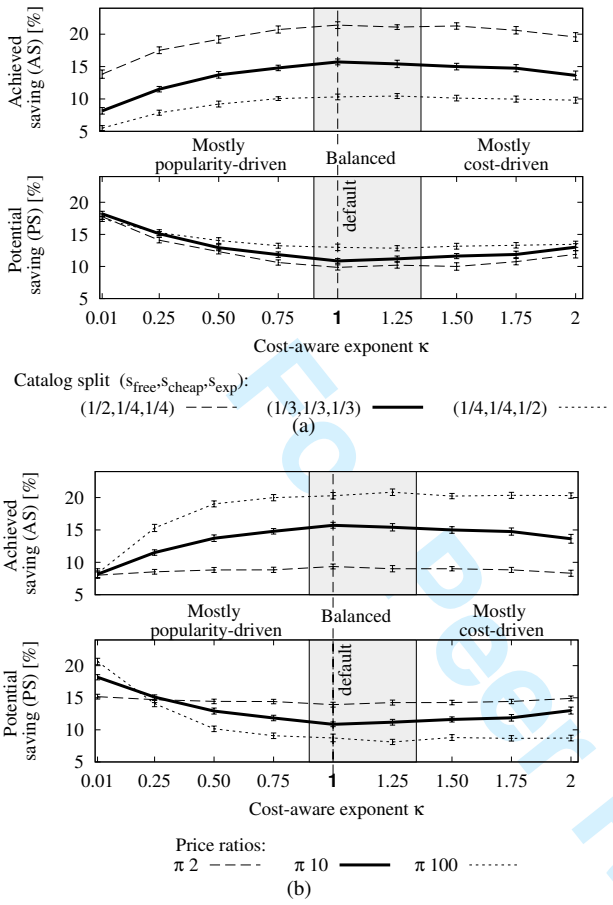


Figure 15. Impact of the κ exponent (that tunes the sensitivity to popularity or price), for different catalog splits (a) and different price ratios (b).

from smallest to the largest scale ($10^2/10^4$ to $10^6/10^8$), gains remain consistent (17% to 11%).

F. Impact of CoA Settings

As discussed earlier, for an efficient cost reduction the item worth should jointly weight popularity and price: the CoA parameter κ permits to tune this tradeoff giving more weight to popularity (low κ) or to price (high κ). It is thus important to perform a sensitivity analysis of κ, to assess to what extent its tuning is crucial in the correctness of CoA operation and to achieve the gain shown so far.

Fig. 15 illustrates the impact of κ on the achieved and potential saving, for different scenarios and price ratios. In particular, Fig. 15-(a) evaluates the impact in an optimistic scenario characterized by the prevalence of free objects $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, a uniform scenario $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and a pessimistic scenario in which most of the objects are behind the expensive link $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$. In Fig. 15-(b) the impact is measured varying the price heterogeneity, by letting π vary in $\{2, 10, 100\}$. Briefly, we observe that the value $\kappa = 1$ guarantees a good performance in all different conditions: indeed, (i) the achieved saving over Unif is close to the maximum value and (ii) potential saving over Ideal-CoA is close to the minimum. In more details, from Fig. 15-(a) we observe that, even for small values of κ, price discrimination brings sizable gains over completely blind strategies. Second, the parameter κ effectively tunes between three regimes (a *mostly popularity-driven* regime, a *balanced* one and a *mostly cost-driven* regime). As expected, gains are larger in the balanced regime (highlighted in gray in the picture), as it is the one that better jointly weights popularity and price, better inferring $\lambda_o \pi_o$. Finally, while largest gains are achieved by $\kappa \approx 1$, we also gather that performance smoothly varies with κ, so that its setting is not critical. Similar considerations hold by fixing the catalog split $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and varying the price ratio in Fig. 15-(b).

VI. Related Work

The design of caches (and cache network) and the economic implications of caching, have been treated as two orthogonal subjects so far. To the best of our knowledge this work is the first to jointly consider them, addressing the design and evaluation of practical schemes for cost-aware ICN routers and networks. Due to the segregation of related literature, we put our work in perspective by separately considering its design (Sec. VI-A) and economic aspects (Sec. VI-B).

A. Design of ICN routers and cache networks

In terms of router design, we notice that ICN-capable routers are beginning to appear, with prototypes by Alcatel [39], Cisco [40] and Parc [41]. The design of these devices demands for specific hardware and software solutions to make them operate at wire speed, which will likely have remarkable effects on the cost of the equipment, a capital expenditure with respect to the ISP's viewpoint. Yet, our focus in this work is more on the cost saving that caching can bring, or, in other words, an operational expenditure viewpoint.

Different aspects of caching systems, like cache sizing, replica placement and path selection have been studied via optimization models by [19], [22], [23], [31] and others. They have two limitations: (i) they provide only theoretical bounds and (ii) the high computational complexity limits their results to small-scale scenarios with simplifying assumptions. Conversely, our scheme is easily implementable and we evaluate it under realistic scenarios, checking its robustness under different boundary conditions.

Other authors focus on practical aspects of caching operation, as replacement [21] and decision policies [17], [33] or routing [24], [26]. In general, all these studies evaluate caching under cost-blind performance indicators, namely the overall network *hit-ratio* [7], [17], [24], the number of *hops* [17], [24]–[26], the latency [19]–[21], [24] or the *link load* [22]–[24]. Conversely, we show that focusing on these indicators does not permit to exploit the potential cost saving. To clarify the difference of our viewpoint, it is worth making a punctual example with e.g., the work in [7], which observes that, by increasing the hit ratio, the inter-domain traffic of an ISP decreases and cost is reduced. Our findings are different and more general: we observe that reducing inter-domain traffic blindly across all external links is not sufficient to get cost saving, as it is necessary to explicitly consider link price heterogeneity, to preferentially increase the hit ratio of objects that are downloaded through more expensive links.

B. Economic implications of caching

Seminal work in this area can be traced back to the late 90s, as for instance [42] that proposes to take into account the "cost" of objects in the caching mechanism. The cost can mean the download latency, the object size, the congestion status of the link used to download the object or even the price paid to use that link. Our work differs from it in two key aspects. First, we specifically focus on the monetary cost of inter-domain traffic, providing results on the realistic saving of an ISP. Second, [42] proposes a replacement algorithm based on complex computations that would be impossible at line speed. On the contrary, we propose a decision policy that is lightweight and easily implementable in an ICN-router.

In the ICN context, the economic implications of caching are considered by [9], [11]–[13], [16]. In more detail, [9] models the economic incentives of different network players (including regulators) to deploy (or support) distributed ICN storage. In [16], the economic feasibility of ICN is evaluated, contrasting it with client-server, peer-to-peer and CDN models.

Closest works to our are [11]–[13]. First, while [11]–[13] consider ISPs as atomic entities and focus on an inter-ISP view, we study the problem of cost saving from an intra-ISP perspective, and propose a scheme that ISPs can use to manage their own networks. Additionally, [11], [12] investigate new pricing models for ICN networks without looking at the caching policy to be used. Our focus is orthogonal: we instead focus on a novel cache strategy and we study its impact on the pricing model currently used in Internet. Finally, [10], [13] study how ISPs can reduce the transit traffic by sharing their cache content exploiting settlement-free peering links: while this reduction is blindly computed among all transit links, in this work we instead explicitly exploit price heterogeneity, that we show to have an important impact in practice.

VII. Discussion

The scope of our proposal is not confined to ICN but includes also other caching technologies such as Web proxies and CDNs. At the same time, we observe that Web proxy caching is becoming less effective, due to the growth of HTTPS traffic, which is inherently uncacheable, which translates in the increase of the inter-domain traffic and the relative cost. On the contrary, ICN and CDNs provide security guarantees while preserving the possibility of content replication. As for the applicability on CDNs, we have to consider that our proposal requires object replication based on the ISP's goal of inter-domain traffic cost reduction. This is a reasonable assumption when considering CDNs operated by ISPs [22] or collaborating with them [43], so that our proposal can also fit this purpose. However, requirements for line of speed operation in ICN represent an additional constraints, creating the need for simple yet effective solutions – challenge that we believe to have successfully tackled in this work.

While this work advocates to consider economic implications of caching as first class ICN citizen, it does so by simplifying the reality: indeed, we are well aware that other aspects behind the monetary cost of inter-domain traffic could be taken into account. These aspects include for instance the latency incurred by users (QoE), the traffic flowing inside the ISP network or to the repository (QoS), which are the very same aspects we decided to ignore in the first place. Yet, while these issues are commonly studied in the literature [7], [17], [19]–[26], the economic aspects addressed in this paper are less common. As such, this work aims primarily at raising interest on this so far neglected aspect, and designing a viable scheme that achieve cost reduction. At the same time, our proposal is explicitly designed to be modular, so that refinement of the policies can take into account a more holistic view, combining more classical QoS/QoE aspects with the notion of cost – which we believe to fall beyond the scope and aim of the present work.

VIII. Conclusion

In this paper, we tackle a fundamental question overlooked in current landscape of Information Centric Networks (ICN) research: namely, the reduction of operational costs as consequence of the reduced load on transit links due to ICN caching.

We argue that classic ICN schemes show fundamental limits, as they aim to optimize caching efficiency, which intrinsically tradeoff with the cost reduction objective. We therefore design a cost-aware mechanism, as a simple yet effective component of a cache decision policy. Our thorough analysis of the proposed scheme show sizable gains over traditional cost-blind mechanisms under a large number of settings and network topologies. Additionally, performance are good enough also compared to ideal and optimal schemes, that provide upper bounds to the cost reduction achievable in any network scenario.

Our results show that introducing a caching bias toward more expensive objects is a simple, scalable and robust solution: providing a significant cost saving at practically no additional complexity, it therefore represents a promising framework to integrate in next-generation ICN architectures.

Kudos

This work was funded by Labex DigiCosme (project ANR-11-LABEX-0045-DIGICOSME) operated by ANR as part of the program Idex Paris-Saclay (ANR-11-IDEX-0003-02) and carried out at LINCS http://www.lincs.fr.

References

[1] G. Xylomenos, C. N. Ververidis, V. a. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. V. Katsaros, and G. C. Polyzos, "A Survey of Information-Centric Networking Research," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 1024–1049, 2013.

[2] I. Castro, S. Member, R. Stanojevic, and S. Gorinsky, "Using Tuangou to Reduce IP Transit Costs," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1415–1428, 2014.

[3] L. Gyarmati, R. Stanojevic, M. Sirivianos, and N. Laoutaris, "Sharing the cost of backbone networks," in *ACM SIGCOMM IMC*, 2012, p. 509.

[4] M. Motiwala, A. Dhamdhere, N. Feamster, A. Lakhina, G. Tech, and C. Guavus, "Towards a Cost Model for Network Traffic," *ACM SIGCOMM CCR*, vol. 42, no. 1, pp. 54–60, 2012.

[5] R. Stanojevic, N. Laoutaris, and P. Rodriguez, "On Economic Heavy Hitters : Shapley value analysis of 95th-percentile pricing," in *ACM SIGCOMM IMC*, 2010.

[6] J. Roberts and N. Sbihi, "Exploring the Memory-Bandwidth Tradeoff in an Information-Centric Network," in *IEEE ITC*, 2013.

[7] C. Barakat, A. Kalla, D. Saucez, and T. Turletti, "Minimizing Bandwidth on Peering Links with Deflection in Named Data Networking," in *ICCIT*, 2013.
[8] S. Shakkottai and R. Srikant, "Economics of Network Pricing With Multiple ISPs," *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 1233–1245, 2006.
[9] P. K. Agyapong and M. Sirbu, "Economic Incentives in Information-Centric Networking : Implications for Protocol Design and Public Policy," *IEEE Commun. Mag.*, 2012.
[10] G. Dan, "Cache-to-Cache: Could ISPs Cooperate to Decrease Peer-to-peer Content Distribution Costs?" *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 9, pp. 1469–1482, 2011.
[11] F. Kocac, G. Kesidis, T. Pham, and S. Fdida, "The effect of caching on a model of content and access provider revenues in information-centric networks," in *IEEE SocialCom*, 2013.
[12] T. Pham, S. Fdida, and P. Antoniadis, "Pricing in Information-Centric Network Interconnection," in *IEEE IFIP Networking*, 2013.
[13] V. Pacifici and D. Gyorgy, "Content-peering Dynamics of Autonomous Caches in a Content-centric Network," in *IEEE INFOCOM*, 2013.
[14] A. Araldo, M. Mangili, F. Martignon, and D. Rossi, "Cost-aware caching: optimizing cache provisioning and object placement in ICN," in *IEEE Globecom*, 2014.
[15] V. Valancius, C. Lumezanu, N. Feamster, R. Johari, V. V. Vazirani, and G. Tech, "How Many Tiers ? Pricing in the Internet Transit Market," in *ACM SIGCOMM*, 2011.
[16] N. Zhang, T. Levä, and H. Hämmäinen, "Value networks and two-sided markets of Internet content delivery," *Telecommunications Policy*, vol. 38, no. 5-6, pp. 460–472, May 2014.
[17] W. K. Chai, D. He, I. Psaras, and G. Pavlou, "Cache Less for More in Information-centric Networks," in *IFIP Networking*, 2012.
[18] I. Psaras, W. K. Chai, G. Pavlou, and S. Member, "In-Network Cache Management and Resource Allocation for Information-Centric Networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2920–2931, 2014.
[19] S. Zaman and D. Grosu, "A Distributed Algorithm for the Replica Placement Problem," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 9, pp. 1455 – 1468, 2011.
[20] N. Laoutaris, O. Telelis, V. Zissimopoulos, and I. Stavrakakis, "Distributed Selfish Replication," *IEEE Trans. Parallel Distrib. Syst.*, vol. 17, no. 12, pp. 1401–1413, Dec. 2006.
[21] M. Badov, A. Seetharam, and J. Kurose, "Congestion-Aware Caching and Search in Information-Centric Networks," in *ACM SIGCOMM ICN*, 2014.
[22] D. Tuncer, M. Charalambides, R. Landa, and G. Pavlou, "More control over network resources: An ISP caching perspective," in *CNSM*, 2013.
[23] M. Mangili, F. Martignon, and A. Capone, "A Comparative Study of Content-Centric and Content-Distribution Networks: Performance and Bounds," in *IEEE Globecom*, 2013.
[24] S. K. Fayazbakhsh, Y. Lin, A. Tootoonchian, and A. Ghodsi, "Less Pain , Most of the Gain: Incrementally Deployable ICN," in *ACM SIGCOMM, ICN Workshop*, 2013.
[25] Y. Wang, Z. Li, G. Tyson, S. Uhlig, and G. Xie, "Optimal Cache Allocation for Content-Centric Networking," *ICNP*, 2013.
[26] G. Rossini and D. Rossi, "Coupling Caching and Forwarding : Benefits, Analysis, and Implementation," in *ACM SIGCOMM ICN*, 2014.
[27] A. Araldo, D. Rossi, and F. Martignon, "Design and evaluation of cost-aware information centric routers," in *ACM SIGCOMM ICN*, 2014.
[28] 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.
[29] T. Hau, D. Burghardt, and W. Brenner, "Multihoming, content delivery networks, and the market for Internet connectivity," *Telecommunications Policy*, vol. 35, no. 6, pp. 532–542, Jul. 2011.
[30] L. Muscariello, G. Carofiglio, and M. Gallo, "Bandwidth and storage sharing performance in information centric networking," *ACM SIGCOMM, ICN Workshop*, 2011.
[31] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, and K. K. Ramakrishnan, "Optimal Content Placement for a Large-Scale VoD System," in *ACM CoNEXT*, 2010.
[32] V. Martina, M. Garetto, and E. Leonardi, "A unified approach to the performance analysis of caching systems," in *IEEE INFOCOM*, 2014.
[33] N. Laoutaris, S. Syntila, and I. Stavrakakis, "Meta algorithms for hierarchical Web caches," in *IEEE IPCCC*, 2004.
[34] S. Arianfar, P. Nikander, and J. Ott, "On content-centric router design and implications," in *ACM CoNEXT, ReArch Workshop*, 2010.
[35] "ccnSim." [Online]. Available: http://www.enst.fr/~drossi/ccnSim
[36] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Temporal locality in today's content caching," *ACM SIGCOMM CCR*, vol. 43, no. 5, pp. 5–12, Nov. 2013.
[37] 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 *ACM SIGCOMM IMC*, 2007.
[38] H. H. Liu, Y. Wang, Y. R. Yang, H. Wang, and C. Tian, "Optimizing cost and performance for content multihoming," *ACM SIGCOMM CCR*, vol. 42, no. 4, p. 371, Sep. 2012.
[39] D. Perino, M. Varvello, B. Labs, L. Linguaglossa, R. Laufer, and R. Boislaigue, "Caesar: A Content Router for High-Speed Forwarding on Content Names," in *ACM/IEEE ANCS*, 2014.
[40] W. So, A. Narayanan, D. Oran, and M. Stapp, "Named Data Networking on a Router: Forwarding at 20Gbps and Beyond Categories and Subject Descriptors," in *ACM SIGCOMM*, 2013.
[41] N. Solis, G. Scott, and G. Edens, "CCN 1.0 keynote," in *ACM SIGCOMM ICN*, 2014.
[42] P. Cao and S. Irani, "Cost-Aware WWW Proxy Caching Algorithms," in *Usenix symposium on internet technologies and systems*, 1997.
[43] B. Frank, I. Poese, Y. Lin, G. Smaragdakis, A. Feldmann, B. M. Maggs, J. Rake, S. Uhlig, and R. Weber, "Pushing CDN-ISP Collaboration to the Limit," *ACM SIGCOMM CCR*, vol. 43, no. 3, pp. 35–44, 2013.

Andrea Araldo (S'14) is a PhD candidate at Université Paris-Sud and Télécom ParisTech. He received his MSc in Computer Science from Università di Catania in 2012. He worked in European research projects, such as FP7 Ofelia and FP7 mPlane. In 2012 he worked for the Italian academic consortium CNIT. He participated to the implementation of CONET, an Information Centric Network architecture. His current research interests are Information-centric networking and Internet traffic measurement.

Dario Rossi (SM'13) received his MSc and PhD degrees from Politecnico di Torino in 2001 and 2005 respectively, and held a visiting researcher position at University of California, Berkeley during 2003-2004. He his currently Professor at Telecom ParisTech (Paris, France) and Ecole Polytechnique (Palaiseau, France). He has coauthored over 120 conference and journal papers and participated in the program committees of over 50 conferences including ACM CoNEXT and IEEE INFOCOM. His current research interests include Internet traffic measurement, Information-centric networks and high speed networking.

Fabio Martignon (M) received the MSc and the PhD degrees in telecommunication engineering from the Politecnico di Milano in October 2001 and May 2005, respectively. He has been associate professor at University of Bergamo, and he is now Full Professor in the Laboratory for Computer Science, Paris-Sud University, and member of Institut Universitaire de France. His current research activities include information-centric networks, network planning and game theory applications to networking problems.