Design and Evaluation of Cost-aware Information Centric Routers

Andrea Araldo
LRI-Paris-Sud University
91405 Orsay, France
araldo@lri.fr

Dario Rossi
Telecom ParisTech
75013 Paris, France
dario.rossi@enst.fr

Fabio Martignon
LRI-Paris-Sud University
91405 Orsay, France
fabio.martignon@lri.fr

ABSTRACT

Albeit an important goal of Information Centric Networking (ICNs) is traffic reduction, a perhaps even more important aspect follows from the above achievement: the reduction of ISP operational costs that comes as consequence of the reduced load on transit and provider links. Surprisingly, to date this crucial aspect has not been properly taken into account, neither in the architectural design, nor in the operation and management of ICN proposals.

In this work, we instead design a distributed cost-aware scheme that explicitly considers the cost heterogeneity among different links. We contrast our scheme with both traditional cost-blind schemes and optimal results. We further propose an architectural design to let multiple schemes be interoperable, and finally assess whether overlooking implementation details could hamper the practical relevance of our design. Numerical results show that our cost-aware scheme can yield significant cost savings, that are furthermore consistent over a wide range of scenarios.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Network communications, Packet-switching networks

General Terms

Algorithms; Performance; Design;

Keywords

Information Centric Networking; Cost-Awareness

1. INTRODUCTION

Information Centric Networks (ICN) let end-users' applications directly access named content, as opposite to addressable entities as in the current TCP/IP Internet. One among the expected benefits of ICN consists in *traffic reduction* through transparent caching, as opposite to deploying per-application "network accelerators" as typically happens nowadays.

Yet, benefits of ICN with respect to current technologies, such as caching at the network edge as in CDN, are so far unclear. On the one hand, recent research [12, 13] argues that benefits of ubiquitous ICN caching may, in reason of an unfavorable cache-to-catalog ratio, be neither sufficient[1], nor actually necessary[2]. On the other hand, before concluding that ICN has yet to convince, all the relevant factors need to be taken into account. These factors include, for instance: more optimistic ICN cache sizes due to algorithmic design [31] (rather than memory technology advances, which happen at a much lower pace), or the existence of a temporal correlation of the active catalog and requests [34] (that makes ICN caching more effective) or economic aspects [4] (since cost reduction is the ultimate goal of traffic reduction).

We argue that especially this latter aspect has yet to receive the attention it deserves in the ICN community. Namely, economic aspects [4] are perhaps the most important among ICN key performance indicators, and should be considered as a proxy of ICN success: capital expenditures for ICN deployment will be planned according to a direct measure of the expected operational ISP costs (and, especially, savings) under ICN. Yet, despite much research has focused on ICN performance *within* ISP boundaries, to date few works evaluate the effect of ICN on cost reduction *across* boundaries [24, 26, 17, 4].

In this work, we thus challenge the implicit simplifying assumption made in the literature that all inter-ISP links have equal cost, and address the *design and performance evaluation of cost-aware techniques*, whose main design goals are (i) flexibility to support multiple ICN architectures, (ii) interoperability with currently existing or future schemes, (iii) robust operation to ensure practical relevance of our proposal and, finally, (iv) simplicity to facilitate its adoption.

The rest of this work is organized as follows. In Sec. 2 we illustrate our system model, outline the guiding criteria of our design, and propose a simple distributed technique to achieve cost-effective ICN operations. We evaluate our proposal in Sec. 3, where we contrast it with traditional

ICN'14, September 24 - 26 2014, Paris, France.

http://dx.doi.org/10.1145/2660129.2660156.

---

[1] As [13] brilliantly points out, "changing the overall network architecture in order to tame the exponentially growing world of content with the logarithmic sword of caching seems a classical example of taking a knife to a gunfight: it may make for a great story, but it won't end well."

[2] In particular, [12] argues that most of the caching gain is attainable by simply (and painlessly) caching at the edge of the network, as in the current CDN model.

cost-blind schemes as well as the optimal solution, gathered in centralized settings, as a reference – showing that results are structurally similar, and performance very close, to that achieved by ideal policies. We then assess robustness of operation under implementation constraints, as well as over a wider range of scenarios, in Sec. 4. Finally, Sec. 5 places our proposal in the context of related effort, and Sec. 6 summarizes our main lessons.

2. SYSTEM MODEL AND DESIGN

In this section, we first introduce our model of economic interactions (Sec. 2.1). We then describe the principles (Sec. 2.2) that guide our design (Sec. 2.3). Finally, we introduce the terms of comparison, i.e., traditional cost-blind and optimal ICN strategies (Sec. 2.4).

2.1 Economic interactions

As shown in Fig.1, an ISP serves a rate λ_o of requests for a named object o belonging to the catalog $\mathcal{C}$. To serve these requests, the ISP possibly has to retrieve the object through one of its available *external links* (we denote them with the set $\mathcal{L}$), paying a related cost.

In case the ISP is operating caches, some of these requests can hence 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, according to the Forwarding Information Base (FIB). Indicating with $FIB(o)$ the result of the FIB lookup at the egress node for object o (i.e., lookup for content providers as in DONA [18], or for name prefix as in CCN [16]), the subset $\mathcal{C}_j$ of the original catalog $\mathcal{C}$ that is attainable through link j is thus $\mathcal{C}_j = \{o : FIB(o) = j\} \subseteq \mathcal{C}$. It follows that the load on the external link j will be (using unit object size for the sake of simplicity in the formulation):

$$\rho_j = \sum_{o \in \mathcal{C}_j} \lambda_o(1-h_o). \tag{1}$$

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 [10, 15, 32], we abstract the different types of interactions by distinguishing three categories of links, based on the cost associated to the traffic flow:

(i) **Settlement-free peering links** (e.g., connection between ISPs of the same tier) do not imply any economic transaction between the connected ISPs;

(ii) **Provider links** (e.g., transit link 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;

(iii) **Customer links** (e.g., links toward lower tier ISP, or CPs in multihoming [15, 21] or CDNs nodes) imply a revenue[3] for the ISP.

The maximization of the cache hit ratio, irrespectively 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 from the object cost: rather, the ISP aims at caching objects that lead to larger cost savings, i.e., objects that are accessible through the most expensive links.

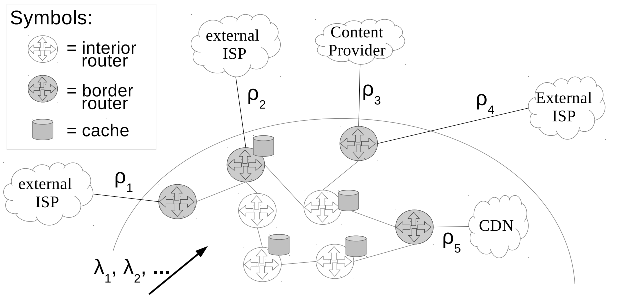


Figure 1: ISP model used throughout this work. The ISP is connected to third party networks through external links having prices π_j, and supporting a total traffic load of ρ_j.

Hence, unlike current literature that evaluates the cache vs bandwidth tradeoff within ISP boundaries [8], we instead assume as in [32] that these internal links have no cost. As commonly done in the literature and confirmed by very recent work [14] stating that the 95% charging model is still widely used, we consider that the cost incurred in retrieving objects is directly proportional to the traffic flowing on that link. Ultimately, the ISP operational cost jointly depends on the traffic load ρ_j crossing any given link j and the link price π_j:

$$\sum_{j \in \mathcal{L}} \pi_j \rho_j = \sum_{j \in \mathcal{L}} \pi_j \sum_{o \in \mathcal{C}_j} \lambda_o(1-h_o). \tag{2}$$

Thus, we argue that ISPs are interested in minimizing the above overall cost (2), considering not only the popularity λ_o but also the link prices π_j, as opposite to maximizing the overall hit ratio $\mathbb{E}[h_o]$ in a cost-blind fashion. In an ongoing related effort [5], we show these to be contrasting objectives in an optimization framework. In this work, we instead focus on a complementary perspective: the design of a distributed cost-aware mechanism, whose performance approaches the one gathered by the solution of a centralized optimization problem.

2.2 Cost-aware ICN guidelines

Our design of a cost-aware ICN is guided by a number of principles, namely (i) **Flexibility**, (ii) **Simplicity**, (iii) **Interoperability** and (iv) **Robustness**: these principles ensure that the resulting cost-aware design (i) can be fit in any existing ICN architecture, (ii) is simple enough to be worth implementing, (iii) is backward and forward compatible with extensions of any specific architecture and (iv) its implementation does not degenerate, under adverse conditions, in suboptimal behavior. In this section, we follow the above rationales in the selection process of the ICN architectural components that are apt to expose cost-aware functionalities.

---

[3] 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 [19, 4].

Following the taxonomy in [36], we namely consider the (i) **Naming**, (ii) **Routing**, (iii) **Forwarding** and (iv) **Caching** components: indeed, retrieval costs for named objects could (i) be embedded in the object name, and (ii) be possibly propagated via an ICN routing protocol; or (iii) be based on name resolution strategies, and consequently path or content-replica selection, which can be achieved in distributed settings by affecting forwarding decisions at each hop; or finally (iv) be embedded in caching-related components, by e.g., preferably storing the most costly objects.

2.2.1 Flexibility

Cost-aware ICN design should be general and flexible, so that it could be plugged as a component in any existing design, rather than requiring a complete redesign of the architecture. Since caching is a common point of most ICN architectures, a plausible option is to design cost-awareness around this component.

Conversely, exploiting peculiar naming schemes is not advisable, since this choice would break flexibility (as CCN-like prefix-based and DONA-like flat names are processed in different ways). Hence, it follows that exploiting the ICN routing component, as it is tightly coupled to naming, is not advisable either. Finally, exploiting the ICN forwarding component does not seem to be a good option, as this could reduce the degrees of freedom, and could compromise ICN efficiency: for instance, in terms of forwarding it would be advisable to exploit off-path cached copies via Nearest Replica Routing [12], which could be compromised by cost-aware solutions modifying the forwarding behavior.

2.2.2 Simplicity

Cost-awareness should be as simple as possible to implement, as simplicity is often a key ingredient to the success of an idea, and the KISS (Keep It Simple, Stupid!) is among one of the basic principles of computer science (and beyond [23]).

This guideline suggests that ICN components such as forwarding and routing are not ideal candidates. Indeed, forwarding operations already pose significant challenges to be performed at high-speed, and are matter of research per se [35, 37, 33]. Similarly, we can rule out routing, as, other than still being under definition, it is significantly complex (as testified by much valuable research on BGP). The simplicity goal thus indicates the caching or naming component as the natural target for cost-awareness: e.g., the former could exploit price information encoded in the latter to realize cost savings.

2.2.3 Interoperability

To maximize interoperability, the architectural design should allow multiple algorithms to transparently integrate, without mutually affecting their respective behaviors. As previously outlined, introducing cost-awareness in the forwarding component could break other desirable properties. Similarly, while routing weights are used to affect load within the ISP network, they may impact forwarding, which is thus not advisable. Finally, exploiting peculiar naming schemes is not advisable, not only because it would compromise security (as cryptographic signatures of the content are generally associated to names, so that verifiability would be lost), but also because it could compromise interoperability (as it is not straightforward to stack multiple modifications, in a furthermore invertible manner).

To ensure interoperability in the remaining component (i.e., caching), what is required is a syntactically rich way to let multiple independent strategies to transparently interoperate. This means, in particular, accommodating multiple caching policies beyond the cost-aware we propose in this work, such as policies driven by popularity (LCD [20], Unif [6], TwoHit [22]) or based on distance [27] or topological properties [9, 28]: since each of the above policies exploits different practical aspects, their benefits are possibly worth integrating.

We argue that a simple way to let these policies interoperate is via a standard packet format: i.e., border routers could tag packets with cost-related information for further processing in the network. We additionally notice that, since price information is domain-specific, packets would be tagged by border routers upon entrance in a new domain, ensuring safety of operation (e.g., against cheating neighboring domains).

2.2.4 Robustness

Finally, it would be desirable that cost-awareness is not compromised in practice when deployed in different scenarios (e.g., different popularity or cost settings), unexpected operational points (e.g., interaction with untested algorithms), or external constraints (e.g., packet framing format). In all the above situations, the expected behavior should hopefully be maintained, and in any case it must not deteriorate or adversely impact the architecture performance.

For instance, consider the packet framing formats. While it is totally out of the scope of our work to propose a format, which is indeed a matter of discussion at IRTF [1], we outline two possibilities to represent cost-related information: (i) to use a simple but rigid syntax, using a fixed-size field of a standard packet header format versus (ii) using a more complex but flexible syntax as Type Length Value (TLV) encoding.

Both implementations have pros and cons: experience with TCP/IP tells that while fixed-size fields are simpler (thus, faster) to handle, they also scale badly over time, and tend to become critical resources (e.g., IP TOS field). Moreover, while mechanisms to circumvent these limits exist (e.g., IP options), however they happen to be rarely used in practice. Conversely, flexibility (e.g., of TLV) comes at a price of increased complexity: historically, following the principle of pushing complexity to the edge, fixed framing has been preferred for lower layers of the protocol stack, that need to be treated within the network core, relegating syntactically more expressive formats to the application layer.

For our purpose, both solutions are in principle possible. For the sake of simplicity, during the design and evaluation phase it would be preferable to consider that border routers can tag packets with arbitrary information. However, this may not be true in practice, as the information bits available to express price differences may be limited. It follows that the architectural design should be stress-tested against such imposed limitations: in case benefits disappear, this can either be symptomatic of ill architectural design (requiring a redesign of some component), or be more general and thus worth bringing up as matter of discussion in the standardization process [11].

2.3 Cost-aware ICN design

Summarizing, the above principles identify the most flexible, simple, interoperable and robust design as the one embedding cost-awareness in the *caching component.* However, the design space is still fairly large.

As Fig. 2 shows, at every arrival of a new object, a *decision* has to be taken: whether to cache the new object (aka meta-caching), in which case a *replacement* policy is triggered to select a previously cached object to be discarded. Both replacement and decision are possible candidates to exploit cost-related information. We now discuss practical tradeoffs of cost-aware caching, that lead to our proposal.

2.3.1 Meta-caching vs Replacement policies

Intuitively, to reduce costs, it would be desirable for an ICN not only to cache 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 would be thus to bias the caching process toward more expensive objects. We argue this bias is better introduced in the cache decision (or meta-caching) policy, to avoid 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 [20, 22] or probabilistic [6, 22, 27, 9] meta-caching policies would be preferable. By extension, it would be better to bias the acceptance toward more expensive objects in the cache, than to bias the replacement process toward cheaper objects a posteriori.

Moreover, a cost-aware *replacement* process would require storing in the cache additional per-object metadata regarding the price of all objects, as their price needs to be accounted for to select the candidate for replacement. This is undesirable, since it increments complexity and costs. On the contrary, a cost-aware *decision* strategy is simpler to implement, as price-related information can be added within the packet header by the ISP border router once, and exploited independently by any router to take its meta-caching decision upon the reception of a new packet.

2.3.2 Cost-Aware (CoA) proposal

Overall, we design a cost-aware scheme based on modular meta-caching strategies (based on topological information, distance, popularity, cost, etc.). As exemplified in Fig. 2, composition can be simply achieved via product of functions, so that a meta-caching component driven by both cost and popularity will accept a new object with probability $\alpha(\cdot)\beta(\cdot)$, where $\alpha(\cdot)$ and $\beta(\cdot)$ jointly but independently weight popularity and price, respectively.

In practice, only *border routers* know the link through which objects enter the ISP domain, and can thus (i) compute a cost-related meta-caching probability $\beta(\cdot)$ and tag the packet accordingly; (ii) additionally, in case they are equipped with storage components, border routers take a caching decision according to $\alpha(\cdot)\beta(\cdot)$ prior to forwarding the packet. *Interior routers* along the path then (iii) take independent caching decisions based on the cost-related information tagged by border routers, and by any other information (e.g. centrality, distance), which possibly differs among routers.

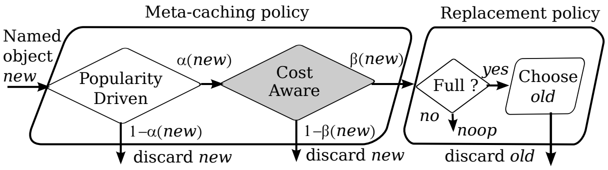


Figure 2: Cost-aware ICN design, plugged within the meta-caching decision policy of the caching component.

2.3.3 Popularity-driven vs Cost-aware decisions

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 trade-off: e.g., caching expensive but unpopular objects may not bring effective cost reductions. The design of the cost-aware function should thus permit to bias objects coming from links with different prices, but should still permit to differentiate between popular and unpopular objects. In other words, it would be useful to explicitly assign a weight between popularity and cost-awareness in the decisions. These observations lead to the following choice of function:

$$\beta(o) = M \cdot \pi_o^\kappa / \sum_{j \in \mathcal{L}} \pi_j^\kappa \qquad (3)$$

where π_o is the price of the link through which the border ICN router received the new object o and M is a constant that can be used to adjust the overall cache admission probability. Finally, the exponent $\kappa \in \mathbb{R}$ is used to tune the relative importance of popularity vs cost in the decision: indeed, the larger κ, the larger the skew toward costly objects, while for $\kappa < 1$ the importance of cost in the decision diminishes (note that the function degenerates into a uniform probability $M/|\mathcal{L}|$ when $\kappa = 0$).

We ensure that the average cache admission probability is equal in the cost-blind and cost-aware cases, choosing M in (3) such that $\mathbb{E}[\alpha(\cdot)\beta(\cdot)] = \mathbb{E}[\alpha(\cdot)]$. While this is a second order detail as far as the *design* is concerned, it is however important in order to clearly distinguish the benefits coming from cost-awareness and *fairly compare* cost-aware vs cost-blind schemes in Sec. 3.

2.4 Terms of comparison

We contrast our design, that we denote with CoA, against several terms of comparison: (i) cache-less systems, (ii) traditional ICN schemes where cost heterogeneity is not directly taken into account, (iii) ideal distributed decision policies with perfect knowledge of object popularity and (iv) optimal centralized solutions achieving provably minimum cost.

2.4.1 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.

2.4.2 *Cost-blind ICN*

Following our design, a natural term of comparison for cost-blind ICN consists in considering state-of-the-art meta-caching policies that ignore the costs of object retrieval (i.e., equivalent to setting $\beta(\cdot) = 1$). The popularity-driven meta-caching component could use Leave a Copy Everywhere (LCE, equivalent to setting $\alpha(\cdot) = 1$), Leave a Copy Down [20] (LCD, accepting new items only when they have traveled $d = 1$ hop in the network, expressed with the Dirac delta function $\alpha(\cdot) = \delta(d - 1)$), Uniform probabilistic decisions (Unif) [6] (where $\alpha(\cdot) = \alpha_0 \in [0, 1]$), or decisions based on distance [27], graph properties [9], correlation between consecutive requests [22], etc.

As it emerges from [22, 30], uniform probabilistic decisions are expected to be simple yet effective, and are thus preferable. Note that, while in the case of homogeneous prices a lower α_0 translates into better caching results (as it reduces eviction, due to less likely acceptance of rare objects, at the price of a slower convergence in learning the object popularity) this does however not hold in the case of heterogeneous prices: intuitively, a slower convergence also translates into more frequent downloads of objects before they are accepted into the cache, reducing the caching capability of absorbing costs. To gather a conservative estimate of cost-awareness benefits, we perform a preliminary calibration to find the most favorable setting in the scenarios under investigation, and fix $\alpha_0 = 1/100$.

2.4.3 *Ideal strategies*

We additionally consider strategies that ideally have perfect knowledge of object popularity, and that either explicitly take into account, or deliberately disregard, the object retrieval cost. Specifically, the decision whether to cache or not a new object is assisted by considering the eviction candidate of the Least Recently Used (LRU) replacement policy: the new object is accepted only if it is more "valuable" than the eviction candidate, which 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 savings, which happens by admitting only objects whose $\lambda_o \pi_o$ is larger than that of the LRU eviction candidate.

2.4.4 *Global optimum*

We finally find the minimal ISP cost by solving the optimization problem formalized in [5], where we minimize the cost incurred by an ISP by storing in the cache, a priori, objects o with the largest product of cost times popularity $\lambda_o \pi_o$.Since we use the optimum as a reference against our design, we deem its full formulation to be outside the scope of the paper, and refer the interested reader to [5] for more details.

3. 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 allows us to gauge the extent of possible improvements in our design.

With the exception of the global optimal solution, that we compute numerically, all strategies are implemented in ccnSim, an efficient and scalable [29] open-source ICN simulator available at [2]. In our assessment, we initially consider a simple scenario (Sec. 3.1), over which we cross-compare, at a glance, all the above strategies (Sec. 3.2), and additionally expose deficiencies of cost-blind strategies (Sec. 3.3). We instead defer the analysis of more complex scenarios to Sec. 4.

3.1 Evaluation scenario

Without loss of generality, we focus on a scenario similar to the one depicted in Fig. 1, where we only consider settlement-free and provider links, and additionally consider that different providers may have different pricing agreements.

Object popularity follows a Zipf distribution having skew parameter α, and we model request arrivals with a Poisson process of intensity λ_o for an object o having rank r_o, with $\lambda_o = \Lambda r_o^{-\alpha} / \sum_{o' \in \mathcal{C}} r_{o'}^{-\alpha}$, Λ being the aggregated request arrival rate.

We split the catalog $\mathcal{C}$ so that only disjoint portions are accessible behind each link. Specifically, we denote with $\mathcal{C}_i$ the set of objects that are accessible via link i and with s_i the corresponding fraction of objects[4]. By definition, we have that $\cup_i \mathcal{C}_i = \mathcal{C}$, that $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \forall i \neq j$, and $s_i = |\mathcal{C}_i|/|\mathcal{C}|$ with $\sum_i s_i = 1$.

For the sake of simplicity, in the reminder of the paper we limitedly 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., $\sum_{o \in \mathcal{C}_i} \lambda_o \neq \sum_{o \in \mathcal{C}_j} \lambda_o$). 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 having increasing price $\pi_3 \geq \pi_2 > \pi_1$. Specifically, one link models a settlement-free relationship ($\pi_1 = 0$), whereas the two other links represent a cheap ($\pi_2 = 1$) and an expensive link ($\pi = \pi_3 \geq \pi_2$, with π a free parameter). By a slight abuse of language, in the reminder of this paper we will refer to an "expensive object" as an object that has been gathered through an expensive link (despite there is no longer a notion of cost within the ISP boundaries after the object has been retrieved). This price diversity, coupled to catalog split settings $\vec{s} = (s_1, s_2, s_3)$, permits to gauge cost-awareness gain in rather diverse scenarios.

---

[4] 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 CoA gains.

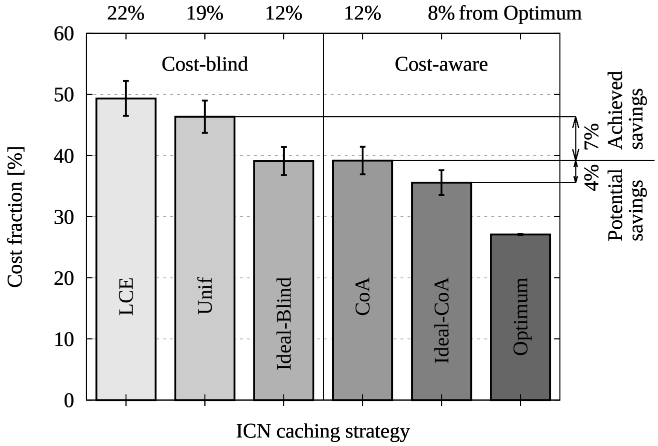


Figure 3: Benefits of cost-aware design. The cost fraction reported on the y-axis is calculated w.r.t. a cache-less system. Cost fraction difference from the global optimum is annotated on the top x-axis. Cost fraction difference of practical cost-aware policy (CoA) w.r.t. state of the art cost-blind policy (Ideal-Blind) and ideal cost-aware policy (Ideal-CoA) are annotated on the right.

Given our definition, it follows that a new object o is accepted in the cache with probability $\alpha(o)\beta(o)$. To ensure that the average cache admission probability is equal in the cost-blind and cost-aware cases, knowing the prices and the catalog split ratio, in (3) we fix $M = \sum_{j\in\mathcal{L}} \pi_j^\kappa / \sum_{j\in\mathcal{L}} s_j \pi_j^\kappa$. It follows that differences between the Unif and CoA strategies are solely due to the cost-aware bias in the meta-caching decision.

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.

3.2 Comparison at a glance

To evaluate the cost-effectiveness achieved by a caching strategy, we compute in each scenario a *cost fraction* as the ratio between the cost obtained by that strategy and the cost obtained by the cache-less strategy in the same scenario. Costs incurred by the ISP are evaluated in this steady state, where the same number of requests is handled by all different strategies. The cost is computed as the weighted sum of the link load ρ_i measured in the simulation, times the link price $\sum_{i\in\mathcal{L}} \rho_i \pi_i$. In case of a cache-less system, $\rho_i = \sum_{o\in\mathcal{C}_i} \lambda_o$ equals the aggregated arrival rate of the objects in $\mathcal{C}_i$, whereas in the case of ICN, ρ_i represents the aggregated miss stream. We express the *cost fraction* of a strategy X over the cache-less system as follows:

$$CF^X = \frac{\sum_{i\in\mathcal{L}} \rho_i^X \pi_i}{\sum_{i\in\mathcal{L}} \left(\sum_{o\in\mathcal{C}_i} \lambda_o\right) \pi_i} \tag{4}$$

with X being any of the strategies introduced earlier (i.e., LCE, Uniform, CoA, Ideal-Blind, Ideal-CoA, Optimum).

We start by considering a scenario with mild price variation $\vec{\pi} = (0, 1, 10)$, a uniform catalog split $\vec{s} = (1/3, 1/3, 1/3)$, a popularity skew $\alpha = 1$, and a cache to catalog ratio of $|c|/|\mathcal{C}| = 1\%$ (with $|\mathcal{C}| = 10^5$ and $|c| = 10^3$). Moreover, we initially set $\kappa = 1$. We instead assess gains in larger and more heterogeneous scenarios in Sec. 4.

Fig. 3 shows, at a glance, the cost fraction for cost-blind (left bars) and cost-aware (right bars) strategies. The figure is further annotated with the absolute distance (i.e., difference of cost fractions) for each strategy to the global optimum (top x-axis). Our strategy (CoA) can bring some sizable benefits, and these benefits appear even over 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 more important benefits with respect to information that is highly volatile and harder to infer (e.g., object popularity).

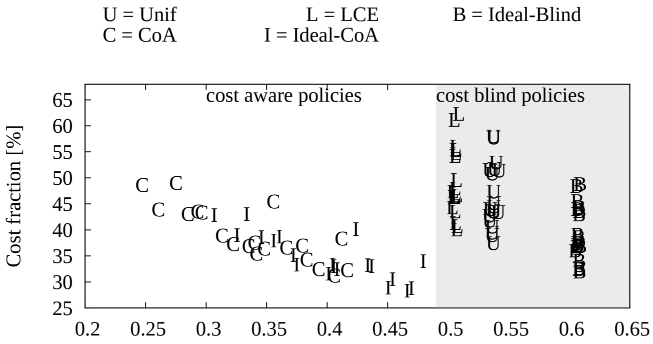
Additionally, consider the absolute distance from CoA to Uniform and Ideal-CoA, that is annotated in the right y-axis of Fig. 3: it turns out that (i) CoA brings a sizable improvement in terms of cost savings (7% of cost fraction reduction with respect to Uniform), and that (ii) there is still additional room for improvement (4% additional potential savings with respect to the Ideal-CoA scheme). Finally, notice that savings already achieved are larger than the additional potential saving, that are possibly tied to the popularity-driven component of the meta-caching policy of Fig. 2.

3.3 Root cause of cost saving

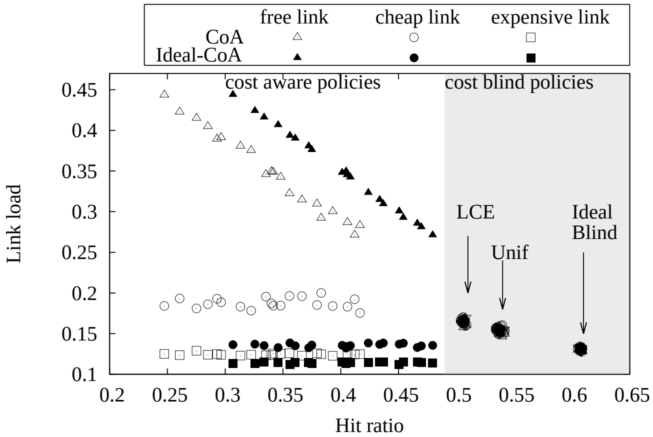
To understand the root cause of the performance gap, we start by showing a scatter plot of the cost fraction versus the cache hit ratio in Fig. 4-(a). In this figure, each point corresponds to a different simulation run: recall that, while the catalog is equally split over the three links, only the *number* of objects that can be attained behind each link is the same, but their *relative popularity* is not, hence the dispersion follows from the variability of aggregated demand in each sub-catalog.

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. Note that the partition of objects among the links at different prices changes from a run to the other. The cost fraction is sensitive to this partition, and this explains why all the policies exhibit high cost fraction variance (y-axis). Additionally, since the objects that are behind the expensive link are more likely to be cached by cost aware policies, the hit ratio of those policies depends on the object partition among links and exhibits high variance (x-axis). On the contrary, cost-blind policies are insensitive to the object distribution and their hit ratio has small variance.

To further assess the impact of cost-aware caching on the network, in Fig. 4-(b) we measure the traffic load over the free, cheap and expensive links, i.e. the number of the objects downloaded on that link divided by the overall amount of user requests. Both CoA and Ideal-CoA achieve structurally similar configurations. Specifically, cost-aware strategies reduce the load on expensive and cheap links (cir-



(a) Scatter plot of cost fraction versus hit ratio

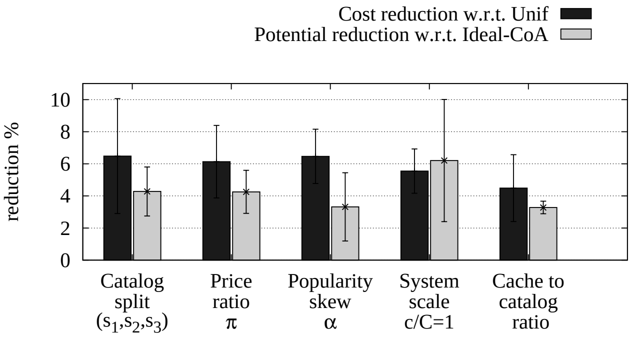


(b) Scatter plot of load over free, cheap and expensive links

Figure 4: Comparison of cost-aware vs cost-blind policies: (a) higher cache hit ratio does not necessarily imply lower cost and (b) cost-aware policies differentiate load on links with heterogeneous prices.

cles and squares in the figure), even if the average hit ratio on the network changes, at the expense of a load increase in the free link (triangles). Note that as the hit ratio decreases, the load on the free link increases: this means that all the additional miss stream includes only free objects. Finally, observe that Ideal-CoA and CoA induce a similar load on the free link, though Ideal-CoA has better hit ratio statistics in reason of 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).



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	$s_i \in \{1/3, h/4 \mid h \in \{0,1,2,3,4\}\}$ $\sum_i s_i = 1$
System scale $\lvert c\rvert/\lvert\mathcal{C}\rvert = 1/100$	5	$10^2/10^4, 10^3/10^5, 10^4/10^6,$ $10^5/10^7, 10^6/10^8$
Cache/catalog ratio $\lvert c\rvert = 10^3$	5	$10^3/10^5, 10^3/10^6,$ $10^3/10^7, 10^3/10^8$

Figure 5: Robustness against external factors.

As final remark, it is worth pointing out that the price differentiation operated by cost-aware policies permit 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.

4. ROBUSTNESS OF COST-AWARE DESIGN

While the previous sections have assessed potential benefits of cost-aware ICN routers operation, for the CoA design to be of any practical interest, the consistency of these gains has to be confirmed in the general case – which is the aim of this section. Specifically, we extend our evaluation to cover (i) a wider range of evaluation scenarios (ii) CoA settings and (iii) practical implementation aspects. We anticipate that gains are consistent, and despite our evaluation is thorough (overall, we perform over 500 simulation runs, accounting for over $8 \cdot 10^9$ requests), we will report it in the most compact way for the sake of synthesis.

4.1 External factors

For what concern evaluation scenarios, there are many factors that are unknown at best, that will likely change in unpredictable manner, and that are by the way not under the control of either manufacturers or ISPs. We therefore evaluate the CoA gain under a wider range of settings in terms of (i) the achieved gain over Unif (ii) the achievable gain to attain Ideal-CoA savings.

Detailed parameters and results are reported in Fig. 5. Clearly, each parameter concurs in determining the absolute savings: e.g., the absolute cost savings may be marginal for very low skew α, or when most of the catalog is accessible only through the most costly link, or when the cache is too small, etc.

Yet, we see that the gains resulting from biasing the cache admission policy along the cost dimension are consistent over all the parameter variations: on average, CoA obtains a cost fraction higher by 4% with respect to the ideal case, gaining 6% over Unif. Note also that these are *absolute* cost fraction differences. While, for the sake of clarity in the exposition, in this paper we refer mostly to absolute cost fraction differences, it is worth stressing that the *relative* gain is more interesting from an economic point of view.

In *relative* terms, the distance between CoA and Ideal-CoA is $(CF^{\text{CoA}} - CF^{\text{Ideal-CoA}})/CF^{CoA} = 10\%$, while the distance between Unif and CoA is $(CF^{\text{Unif}} - CF^{\text{CoA}})/CF^{Unif} = 14\%$. These gain can be interpreted by considering an ISP in which an 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 14% of the current operational expenditure for content retrieval, without making any additional expense. Indeed, while the installation of the ICN infrastructure implies a capital expenditure (capex), our CoA mechanism consists in simple tuning and does not requires additional capex. Yet, CoA offer the ISP a consistent saving in the operational expenditure (opex), that becomes sizeable as it accumulates over the years. Otherwise stated, CoA is expected to yield a +14% gain in the revenue of an ISP having deployed a state-of-the art ICN infrastructure, which is very appealing especially at times of world-wise economic crisis.

4.2 Internal settings

As we have shown, for an efficient cost reduction, the worth of an item should jointly weight popularity and price. Fig. 6 shows the achievable gains for three representative catalog splits, namely: (i) an optimistic scenario where half of the catalog is accessible behind a peering link and the remaining is equally split, (ii) a uniform scenario, (iii) a pessimistic scenario where half of the catalog is accessible behind the most costly link and the remaining is equally split. First, notice from Fig. 6 that already for very small values of κ, price discrimination brings sizable gains over completely blind strategy (when $\kappa = 0.1$, items having price 10 have about 10% more chance to be cached than items having unitary price). Second, notice that κ effectively tunes between three regimes (namely, 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). Finally, while largest gains are achieved by $\kappa \approx 1$, we also gather that performance smoothly varies on κ (so that its setting is not critical) and that ultimately $\kappa = 1$ offers a good performance and is thus a reasonable choice.

4.3 Implementation constraints

We have previously argued that limitations such as quantization of the cost information (due to the limited number of bits available in the packet header) can adversely impact the CoA gains. We set the link prices of the free, cheap and expensive link as $\pi_1 = 0, \pi_2 = z, \pi_3 = 10$ and we make z vary in $1, 2, \ldots, 10$. Effects are expected to be non trivial: for instance, when a single quantization bit is used (binary decision), objects of the cheap link are not cached (as if they were attainable through the free link) when $z < 5$, and are instead cached with the same probability of expensive objects when $z \geq 5$. Additionally, the magnitude of the impact, and not only the frequency of errors in the decision process, also depends on z. We thus represent the average cost fraction loss (with standard deviation) in Fig. 7 for different amount of quantization bits and z values w.r.t. the case when no quantization is applied: it can be seen that the performance degradation is less than 1% (0.1%) with 2 (4) quantization bits, which is an encouraging result. Yet, we point out that a more complete sensitivity analysis (larger number of links, where thus the CoA policy needs to discriminate prices at a possibly finer grain) is needed before a conservative estimate worth bringing up to standardization fora can be made.

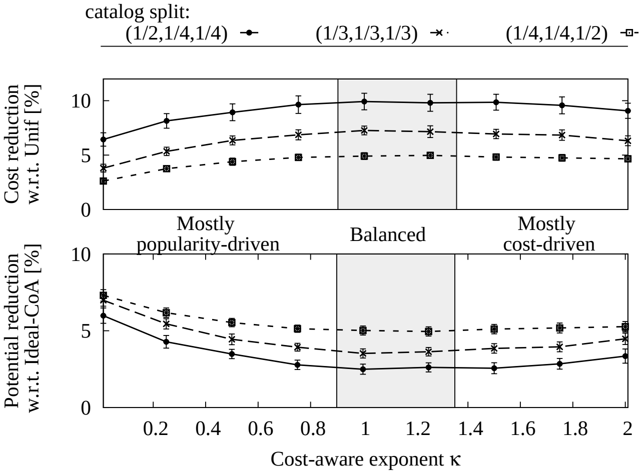


Figure 6: Robustness vs internal settings: impact of κ exponent for different catalog splits.

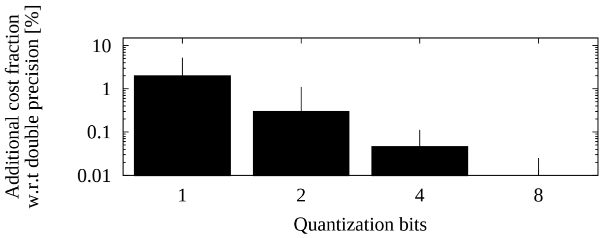


Figure 7: Robustness against implementation constraints: while price quantization affects accuracy of the decisions, the net effect is a negligible cost increase for the ISP.

5. RELATED WORK

We limit our discussion to recent literature, relevant to cost-aware solutions and ICN architectures. In terms of router design, we notice that ICN-capable routers are beginning to appear, with prototypes by Alcatel [35], Cisco [33] and Parc [3]. A first investigation on the possible architecture of an ICN router, with special attention towards computational issues related to the content store, appears in [6]. The work in [25] extends this analysis by presenting quantitative insights on the memory technologies that can be used to make wire-speed processing of ICN packets a reality. Both works focus on economic aspects, that however mostly relate to memory prices.

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 pricing of the equipment, a capital expenditure with respect to the ISP's viewpoint. Yet, our focus in this work is more on the cost savings that caching can bring or in other words, an operational expenditure viewpoint. Closer to our work under this perspective are [4, 24, 26, 7]. In more details, [4] presents an engineering and economic model to evaluate the incentives of different network players (including regulators) to deploy (or support) distributed ICN storage. In [26, 17], authors study the economic incentives in caching and sharing content in an ICN interconnection scenario, with a game theoretic approach. The interaction of autonomous cache networks, at the Autonomous System (AS) level, is addressed in [24], which investigates conditions that lead to stable cache configurations, both with and without coordination between the ASes.

Finally, [7] proposes to take into account the "cost" of objects in the caching mechanism. Yet, the notion of cost is a general one, where cost is a proxy to express a combination of general metrics such as download latency, object size, congestion status of the link used to download the object or the price paid to use that link. Our work differs from [7] in two aspects. First, our notion of cost is more specifically aimed at estimating the realistic cost savings of an ISP in Internet. Second, [7] proposes a replacement algorithm based on complex computation that would be impossible at line speed. On the contrary, we propose a decision policy that is light-weight and easily implementable in an ICN-router.

6. CONCLUSIONS

In this paper, we tackled a fundamental question currently overlooked in the design of Information Centric Networks (ICNs): the reduction of operational costs as consequence of the reduced load on transit and provider links.

To achieve this goal, we designed a cost-aware ICN mechanism: following architectural principles that let our design be simple, flexible, interoperable and robust, we argue that cost-awareness should be embedded as a configurable block of the meta-caching function.

We performed a thorough analysis of the proposed scheme, comparing it with traditional, cost-blind mechanisms, as well as with numerical results that provide upper bounds to the cost reduction achievable in any network scenario. Our results show that, in the scenarios under investigation, exploiting information already at hand that changes over longer timescales (i.e., the prices negotiated with different ISPs), brings as much benefit as information that is much harder to get and more volatile (i.e., item popularity). Results show that not only the structural cache distribution, but also the raw performance, both in terms of cost as well as hit ratio, are very close to those achieved by ideal policies.

Overall our proposed solution is simple, scalable and robust, providing consistent performance improvements and cost savings, thus representing a promising framework to integrate in all future ICN architectures.

While this paper opens a new interesting direction, it however leaves some open questions. Indeed, as we limitedly focus on the economic implications of content retrieval on caching, we neglect aspects that deserve future attention. For instance, ISPs achieve cost reduction by penalizing delay for some contents and users: a more fine-grained assessment of this tradeoff is thus necessary. Additionally, this work limitedly considers caching of monolitic objects: in case of chunk-level caching and applications with quality of service constraints, such as video streaming, futher care should be put to ensure per-object coherence, to avoid video stuttering and quality degradation.

Acknowledgment

This work was carried out at LINCS *http://www.lincs.fr* and funded by the DIGITEO project Odessa-CCN.

7. REFERENCES

[1] `http://trac.tools.ietf.org/group/irtf/trac/wiki/icnrg`.

[2] `http://www.enst.fr/~drossi/ccnSim`.

[3] CCNX Community Meeting (CCNxCon 2013) Keynote . Palo Alto, CA, USA, Sep. 2013. Speaker: Glenn Edens - `http://www.ccnx.org/events/ccnxcon2013/program/` Last accessed: May 2014.

[4] P. Agyapong and M. Sirbu. Economic incentives in information-centric networking: Implications for protocol design and public policy. In *IEEE Communications Magazine, vol. 50(12)*, pages 18–26, Dec. 2012.

[5] A. Araldo, M. Mangili, F. Martignon, and D. Rossi. Cost-aware caching: optimizing cache provisioning and object placement in ICN - `http://www.enst.fr/~araldo/cost_aware_caching.pdf`. Technical report, 2014.

[6] S. Arianfar, P. Nikander, and J. Ott. On content-centric router design and implications. In *ACM CoNEXT, Re-Architecting the Internet Workshop (ReArch)*, 2010.

[7] P. Cao and S. Irani. Cost-Aware WWW Proxy Caching Algorithms. In *Usenix symposium on internet technologies and systems*, 1997.

[8] G. Carofiglio, M. Gallo, and L. Muscariello. Bandwidth and Storage Sharing Performance in Information Centric Networking. In *ACM SIGCOMM, ICN Workshop*, 2011.

[9] W. Chai, D. He, I. Psaras, and G. Pavlou. Cache less for more in information-centric networks. In *IFIP Networking*. 2012.

[10] A. Dhamdhere and C. Dovrolis. The Internet is Flat: Modeling the Transition from a Transit Hierarchy to a Peering Mesh. In *ACM CoNEXT*, 2010.

[11] A. Y. Ding, J. Korhonen, T. Savolainen, M. Kojo, J. Ott, S. Tarkoma, and J. Crowcroft. Bridging the gap between internet standardization and networking research. *ACM SIGCOMM Computer Communication Review*, 44(1):56–62, 2013.

[12] S. K. Fayazbakhsh, Y. Lin, A. Tootoonchian, A. Ghodsi, T. Koponen, B. M. Maggs, K. Ng, V. Sekar, and S. Shenker. Less pain, most of the gain: Incrementally deployable icn. In *ACM SIGCOMM*, 2013.

[13] A. Ghodsi, S. Shenker, T. Koponen, A. Singla, B. Raghavan, and J. Wilcox. Information-centric networking: seeing the forest for the trees. In *ACM HotNets-X*, 2011.

[14] P. Gill, M. Schapira, and S. Goldberg. A survey of interdomain routing policies. *ACM SIGCOMM Computer Communications Review*, 44(1):28–35, 2014.

[15] T. Hau, D. Burghardt, and W. Brenner. Multihoming, content delivery networks, and the market for Internet connectivity. *Elsevier Telecommunications Policy*, 35(6):532–542, 2011.

[16] V. Jacobson, D. K. Smetters, N. H. Briggs, J. D. Thornton, M. F. Plass, and R. L. Braynard. Networking Named Content. In *ACM CoNEXT*, 2009.

[17] F. Kocak, G. Kesidis, T.-M. 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.

[18] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica. A data-oriented (and beyond) network architecture. *ACM SIGCOMM Computer Communication Review*, 37(4):181–192, 2007.

[19] J. Krämer, L. Wiewiorra, and C. Weinhardt. Net neutrality: A progress report. *Elsevier Telecommunications Policy*, 37(9):794–813, 2013.

[20] N. Laoutaris, S. Syntila, and I. Stavrakakis. Meta Algorithms for Hierarchical Web Caches. In *IEEE ICPCC*, 2004.

[21] D. Lee and J. Park. ISP vs. ISP + CDN : Can ISPs in Duopoly Profit by Introducing CDN Services? *ACM SIGMETRICS Performance Evaluation Review*, 40(2):46–48, 2012.

[22] V. Martina, M. Garetto, and E. Leonardi. A unified approach to the performance analysis of caching systems. In *IEEE INFOCOM*, 2014.

[23] W. Ockham. Summa totius logicae.

[24] V. Pacifici and G. Dan. Content-peering dynamics of autonomous caches in a content-centric network. In *IEEE INFOCOM*, 2013.

[25] D. Perino and M. Varvello. A Reality Check for Content Centric Networking. In *ACM SIGCOMM, ICN Workshop*, 2011.

[26] T.-M. Pham, S. Fdida, and P. Antoniadis. Pricing in Information-Centric Network Interconnection. In *IFIP Networking*, 2013.

[27] I. Psaras, W. Chai, and G. Pavlou. Probabilistic in-network caching for information-centric networks. In *ACM SIGCOMM, ICN Workshop*, 2012.

[28] D. Rossi and G. Rossini. On sizing ccn content stores by exploiting topological information. In *IEEE INFOCOM, NOMEN Worshop,*, 2012.

[29] G. Rossini and D. Rossi. ccnSim: a highly scalable CCN simulator. In *IEEE ICC*, 2013.

[30] G. Rossini and D. Rossi. Coupling caching and forwarding: Benefits, analysis, and implementation. Technical report, 2014.

[31] G. Rossini, D. Rossi, G. Garetto, and E. Leonardi. Multi-terabyte and multi-gbps information centric routers. In *IEEE INFOCOM*, 2014.

[32] S. Shakkottai and R. Srikant. Economics of Network Pricing With Multiple ISPs. *IEEE/ACM Transactions on Networking*, 14(6):1233–1245, 2006.

[33] W. So, A. Narayanan, D. Oran, and M. Stapp. Named Data Networking on a Router: Forwarding at 20Gbps and Beyond. In *ACM SIGCOMM*, 2013.

[34] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini. Temporal locality in today's content caching: why it matters and how to model it. *ACM SIGCOMM Computer Communication Review*, 43(5):5–12, 2013.

[35] M. Varvello, D. Perino, and J. Esteban. Caesar: A content router for high speed forwarding. In *ACM SIGCOMM, ICN Workshop*, 2012.

[36] 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 Communication Surveys and Tutorials*, 16(2):1024–1049, Jul. 2014.

[37] H. Yuan, T. Song, and P. Crowley. Scalable ndn forwarding: Concepts, issues and principles. In *IEEE ICCCN*, 2012.