ENERGY SAVING IN CONTENT-ORIENTED NETWORKS

Conference Paper · October 2015

M.Yassine Naghmouchi
Paris Dauphine University

Nancy Perrot
Orange Labs

ENERGY SAVING IN CONTENT-ORIENTED NETWORKS*

N.C. Fofack[1], A.R. Mahjoub[2], M.Y Naghmouchi[3] and N. Perrot[3]
[1]INGIMA Régions
nchoungmofofack@ingima.com

[2]Université Paris Dauphine, France
mahjoub@lamsade.dauphine.fr

[3]Orange Labs, Issy-les-Moulineaux, France
{mohamedyassine.naghmouchi, nancy.perrot}@orange.com

ABSTRACT

By allowing in-network caching, content-oriented networks may significantly decrease the network congestion, shorten the access delays, and reduce latency when delivering contents. On the other hand, a massive deployment of caches may subsequently increase the operational expenditures (OPEX), and particularly the energy bill of telecommunication operators. In this paper, we address the energy saving problem in content-oriented networks. This consists in determining which caches and which links could be switched off to minimize energy consumption in such a way that all demands are met while respecting capacity constraints. We propose a novel Mixed Integer Linear Programming (MILP) Formulation of the problem to solve the related object caching and traffic routing problem on arbitrary graph-based network topologies. We use CPLEX to solve our model to optimality. Then, we assess several network performance metrics. After all, we develop a routing on shortest path-based heuristic in order to compare our solutions with those given by the standard shortest path-based routing. Finally, we discuss the numerical results. We show that: 1) The metrics of interest provide additional insights on the impact and/or gain of introducing energy-aware caches in a real telecommunication network; 2) The benefits of our model compared to a routing on shortest path-based model: 38.72 % of energy saving is reached using our MILP model.

Keywords: Content-Oriented Networks, Cache, Energy Saving, Mixed Integer Linear Programming.

1 INTRODUCTION

Recent years have seen, along with the growing popularity of video over Internet, a huge raise of traffic served by content-oriented networks. These network architectures such as Content Distribution Networks (CDN) and Content Centric Networks (CCN) operate by replicating contents at several locations of the caches network, reducing the delay of delivery and improving the Quality of Service (QoS). CDN has become a basic layer in the network architecture through efficiently distributing content. Furthermore, CCN is a novel design for content-oriented networks by providing new protocols centered on the data. More material about CDN and CCN can be found in [1] and [2], respectively.
The introduction of in-network caches generates additional energy costs due to the mass memory access. Meanwhile they deliver important fraction of traffic, content-oriented networks need to be energy efficient. There have been several works to improve energy efficiency in such networks.

*This work is supported by the European Celtic-Plus project CONVINCE and was partially funded by Finland, France, Romania, Sweden and Turkey

Different models have been developed. In particular, it has been shown that important energy savings can be achieved by turning off network caches and links [3].

Here, we go on this idea by also considering performance evaluation of the content-oriented networks in arbitrary graph-based networks. A major step towards the success of this novel paradigm is to analyze and compare its performance with respect to the classical architecture of IP networks and its routing standards usually based on shortest path between the sources and the destinations. In this paper, we give clear answers to this critical issue by proposing a methodology to assess how the innovative design of content oriented networks behaves as opposed to client-server architecture and its routing standards. At the same time, we study the impact of in-network cache parameters on energy consumption, on load, and on path length. The problem we address is called Energy Saving in Content-Oriented Networks (ESCON). It consists in finding the optimal subset of caches and links that could be turned off to minimize energy while finding a feasible routing in the network, under capacity constraints. Thus, we address the related object caching and traffic routing problem on arbitrary graph-based network topologies.

This paper is organized as follows. In Section 2, we discuss related works. In Section 3, we give a formulation for the problem in terms of graphs. In Section 4, we propose a mixed integer programming formulation and a routing on shortest path-based heuristic. This is used in Section 5 to show the impact and the gain of introducing energy-aware caches in a real telecommunication network, as well as the benefits of our MILP model compared to the one based on a shortest path solutions.

2 OPTIMIZATION MODELS AND ENERGY SAVING IN CONTENT-ORIENTED NETWORKS

Optimization models have been extensively used to study the performance of content-oriented networks. Particularly, they are extremely popular for addressing *the object placement problem* [3 - 9]. A description of these works is detailed in Subsection 2.1. In Subsection 2.2, we explain on which aspects our *object caching approach* differs from object placement problems.

2.1 Related Works: The Object Placement Problem

Here, the goal is to find the optimal locations of a set of objects to install in the network so that the total cost is minimum. Linear optimization technics are used in [4] to study the performance of a content distribution network modelled as a hierarchical cache system with a single origin server. A light-weight cooperative caches management algorithm was developed to maximize the traffic volume served from cache and to minimize the bandwidth cost. Mangili et al. [5] develop a novel optimization model to study the performance bounds of a content-centric network, by addressing the related object placement and routing problem. They show how the innovative design of a content-centric networking acts as opposed to the one proposed by content-distribution networks.

The authors in [8] propose to reduce energy consumption in CDN by turning off CDN servers. It is shown that it is possible to reduce the energy consumption of a CDN while ensuring a high level of availability that meets customer demands. In order to study the impact of different memory technologies on energy consumption, the work in [9] focuses on the energy efficiency considering data delivery and storage. The authors propose a genetic algorithm approach for finding an energy-efficient cache location. As a consequence, CCN yields greater energy savings for very popular content. A further work in [7] discusses the energy benefit of using CCN by comparison to CDN.

To optimize energy efficiency, operators try to switch off as many network devices as possible. Chiaraviglio et al. [6] consider the problem of minimizing power consumption for Internet Service Provider (ISP) networks, but they do not consider in-network caches. In particular, they propose and assess strategies to concentrate network traffic on a minimal subset of network resources. Given a telecommunication infrastructure, the aim here is to turn off network nodes and links while still guaranteeing full connectivity and maximum link utilization constraints.

A more closely related work is the one from Araujo et al. [3] who propose a model for saving energy by disabling devices. The problem is to minimize the total energy consumption by turning off links and caches in order to find a feasible routing in the network satisfying all the demands under the capacity constraints. The authors study the impact of using in-network caches and content delivery network cooperation on an energy-efficient routing. Arbitrary network structures are considered taking into account inrouter caches, while each cache serves only one city. Our problem differs from this optimization model in the sense that it enables sharing a single cache between multiple access points in a content-oriented network, but it also considers an “object caching approach”.

2.2 Our Work: An Object Caching Approach

Our problem can be considered as two joint subproblems, the *routing problem* and the *cache problem*. The routing problem aims at finding the links to be activated in order to satisfy all the demands with a minimum transmission cost. The cache problem is to find both the ON/OFF status of a cache and the volume that it intercepts for each demand, so that the storage cost is minimized. The objective of our work is to simultaneously optimize the traffic routing and the caching costs.

When inspecting the literature, we find that the related traffic routing and caching problem in their broadest sense have been widely studied. Nevertheless, object placement problems are based on static caches which are characterized by a binary hit ratio indicating whether an object is placed or not. This is a major limitation, because here the caching is neither adaptive to the demand changes nor to the popularity of contents since the objects are statically placed.

Thus, this paper is motivated by the fact that caching and routing should jointly adapt to traffic changes. In contrast to object placement which is static and based on placement decisions, our object caching approach is different. In fact, our caches are dynamic and self-adaptive to demand changes since they may run replacement algorithms such as Least Recently Used, Random Replacement, or Time-To-Live policies to add or remove contents from memories. Moreover, they are characterized by their average hit probabilities that indicate the fraction of demands served locally over all contents. Therefore, our work differs from object placement problems as treated in [3 – 9]. To the best of our knowledge, the related object caching and traffic routing problem on arbitrary graph-based network topologies has not been considered in the literature.

3 PROBLEM STATEMENT

We model the network by an undirected graph $G = (V, E)$, where V is the set of vertices and E the set of edges. A vertex can play the role of a provider, a user or a router. Let $P \in V$ be the set of providers, $U \in V$ the set of users, and $C \in V$ the set of routers. The index r will refer to the caches, as we suppose that a cache is installed on every router $r \in C$. A cache r has a streaming capacity $b(r)$, and can be turned on or off. Turning on a cache gives rise to a fixed energy cost and an increased energy consumption in terms of load. Let β denotes the power usage of a cache, i.e. the power consumption of a cache divided by the power consumption of a link [10]. We take $\beta \in [0,1]$. A cache is characterized by a *hit ratio*. As defined in [10], the hit ratio is the proportion of requests which are locally served by a cache with respect to the total number of requests. In our model, the hit ratio, denoted h_r, represents the maximal part of a demand served locally from a cache r. Each edge $uv \in E$ acts as a link and has a capacity $c(uv)$. We assume that each link uses one unit of energy. We note K the set of demands. Each demand $k \in K$ is a demand for several types of contents and is defined by a volume $D_k(gbps)$, a source $u \in U$, and a destination $p \in P$. We assume that the contents are aggregated and then, we determine the part of D_k locally served by a cache r. This is done without making a decision neither on the type nor on the volume of the locally served content. Therefore, our hit ratio does not depend on the type of content requested. It is an indication on the maximal part of the demand served locally over all contents.

The goal is to find which caches and which links to turn off in the network in order to minimize energy consumption, in such a way that all the demands are satisfied respecting capacity constraints. The total energy cost includes the energy used by the links and the energy used by the caches.

4 MIXED INTEGER LINEAR PROGRAMMING FORMULATION AND A ROUTING ON SHORTEST PATH BASED HEURISTIC

Let y_r be a binary variable that takes 1 if the cache in node $r \in C$ is turned on and 0 if not. Let x_{uv} be the binary variable which takes 1 if the link uv is activated and 0 if not. Let s_r^k, $r \in C$, $k \in K$, denotes the proportion of the demand k cached by r. Let z_r, $r \in C$ indicates the load rate of the cache in equipment $r \in C$, that is a fraction of the used bandwidth. Finally, let f_{uv}^k, $uv \in E$, $k \in K$, be the flow of the demand k on the edge uv. A portion of the total energy in the network is related to the transmission (each link uses one unit of energy), the other to the caches : we suppose that when a cache is turned on it uses fraction Υ of β and its power consumption grows linearly with the load to reach β when fully utilized. A MILP formulation for the ESCON problem is then :

$$Min \sum_{uv \in E} x_{uv} + \sum_{r \in C} [\beta \Upsilon y_r + \beta(1-\Upsilon) z_r]$$

$$s_r^k \le h_r D_k \qquad k \in K, r \in C \qquad (\mathbf{1})$$

$$\sum_{k \in K} s_r^k = b(r) z_r \qquad r \in C \qquad (\mathbf{2})$$

$$z_r \le y_r \qquad r \in C \qquad (\mathbf{3})$$

$$\sum_{s \in N_r^k} f_{rs}^k - \sum_{t \in N_r^k} f_{tr}^k = \begin{cases} -D_k, \ if \ r = u \\ s_r^k, \ if \ r \neq \{u,p\} \\ D_k - \sum_{w \in C} s_w^k, \ if \ r = p \end{cases} \qquad k \in K, r \in V \qquad (\mathbf{4})$$

$$\sum_{k \in K} (f_{rs}^k + f_{sr}^k) \le c(rs) x_{rs} \qquad rs \in E \qquad (\mathbf{5})$$

$$y_r \in \{0,1\} \qquad r \in C \qquad (\mathbf{6})$$

$$x_{uv} \in \{0,1\} \qquad uv \in E \qquad (\mathbf{7})$$

$$s_r^k \in \mathbb{R}_+ \qquad k \in K, r \in C \qquad (\mathbf{8})$$

$$z_r \in [0,1] \qquad r \in C \qquad (\mathbf{9})$$

$$f_{uv}^k \in \mathbb{R}_+ \qquad k \in K, uv \in E \qquad (\mathbf{10})$$

Constraints ($\mathbf{1}$) express the fact that a cache $r \in C$ serves locally a part of any demand k up to its maximum hit ratio h_r. By constraints ($\mathbf{2}$), the load of a cache is recorded. Constraints ($\mathbf{3}$), indicate that the load cannot exceed the capacity and should be zero if the cache is off. Constraints ($\mathbf{4}$) are the flow conservation constraints. Here, N_r^k is the set of adjacent cache nodes of the cache r. Finally, at the same time that they determine the ON/OFF status of links, constraints ($\mathbf{5}$) indicate that the flow on a link cannot exceed the capacity.

As a first step, we propose to solve the MILP formulation directly with CPLEX. Then, we propose a heuristic algorithm considering a standard routing based on shortest paths between the source-destination pairs. The objective of this heuristic is to compare our MILP model solutions with the ones based on the classical shortest path routing. Our heuristic is a polynomial-time algorithm that computes for each demand a shortest path between its source and its destination. The caching and the routing are then done along this path. The comparison of this heuristic and our model is analysed in the next section.

5 EXPERIMENTAL RESULTS

In this section, numerical results are discussed for a real telecommunication network. The solutions of our MILP are exploited to evaluate the network performance. Then, we focus on the sensitivity of these metrics to the network caches parameters. Finally, we give a comparison between the results obtained by the MILP model and the heuristic.

Table 1: Repartition of Users, Caches and Providers on the network

Nodes	Number	Level
Providers	0-3	4
Caches	{5-7 , 8-21 , 22 - 43}	{3, 2, 1}
Users	44-65	0

First, we define the diameter of a network d as the length of the longest path between a user and a provider in the network. Our instances are generated from a real French network of diameter $d = 4$ with $|V| = 66$ nodes and $|E| = 83$ edges. Table1 shows the repartition of nodes on the different hierarchical levels according to their role and number. For sensitivity analysis we vary the cache parameters h_r and β. We assign to the caches of the same level the same hit ratio. We denote by h_1, h_2, h_3 the hit ratios assigned to caches of level $1, 2$ and 3, respectively. To be closer to the reality, we take $h_1 > h_2 > h_3$. We define *the global hit ratio* $\widetilde{H}$ as the average of h_1, h_2, h_3. Finally, we consider 22 demands generated such that in average populations among the cities behave similarly. Thus, the total amount of demands originating from a city is proportional to its population.

We implement our formulation on the ILP solver CPLEX version 12.5 [11]. The majority of our instances were solved within a few seconds. Only 8% of instances are solved in a time greater than 15 min; those cases arise only for $\beta = 0.3, 0.4, 0.5$.

We define $\mathrm{PG(k)}$ as the subgraph of a demand k; that is the subgraph obtained by keeping edges rs for which flow variables f_{rs}^{k} are nonzero. Experimentations show that the subgraph of each demand is a tree. Actually, the way in which we define our performance measures depend on the nature of our solutions. In fact, there are many different ways to measure performance of a network, depending on its topology. Our proposal is detailed in the next Subsection.

5.1 Performance Evaluation

In order to evaluate the network performance after solving the MILP, we consider four key metrics namely energy consumption, used bandwidth, effective hit ratio, and path length.

5.1.1 Energy Consumption

The energy consumption EC, is $EC = \sum_{uv \in E} x_{uv} + \sum_{r \in C}[\beta \Upsilon y_r + \beta(1-\Upsilon) z_r]$. Figure 1 shows how energy consumption varies in function of cache parameters (global hit ratio $\widetilde{H}$ and power usage β). We obtain nearly the same energy consumption variation in terms of global hit ratio for all β. First, when $\widetilde{H}$ is zero, caches are not used, so the energy consumption is maximum. Second, when $\widetilde{H}$ increases between 0.1 and 0.8, EC decreases slightly. Finally, for caches with $\widetilde{H}$={ 0.9,1}, a fall of energy consumption is noted. Basically, a good cache usage allows reducing the energy consumption, whatever the values of the power usage β.

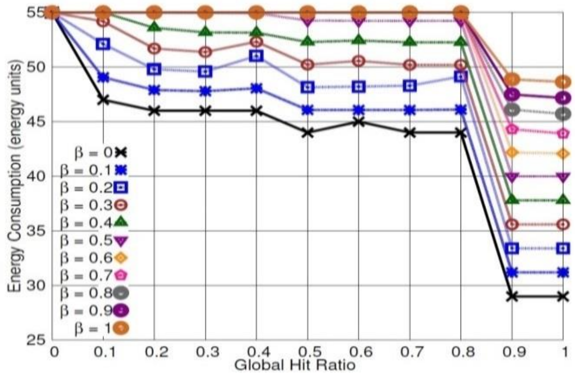

Figure 1: Impact of Cache Parameters on Energy Consumption

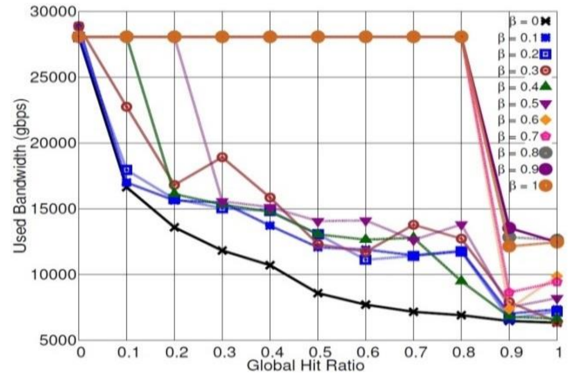

Figure 2: Impact of Cache Parameters on Bandwidth Utilization

5.1.2 Bandwidth Utilization

Bandwidth is considered as the total data transfer rate, i.e. the amount of data that can be carried from one point to another in a given time period. In this usage, bandwidth refers to the data rate that is supported by the network connections. Total bandwidth in the network is denoted Bw and given by $Bw = \sum_{k \in K} \sum_{rs \in E} f_{rs}^{k}$. Figure 2 displays the variation of total bandwidth in terms of global hit ratio and power usage. We show the benefit of introducing caches in the network as a way to control congestion in the network.

5.1.3 Effective Hit Ratio and Path Length

We propose to study the effective hit ratio for each demand. We distinguish the effective hit ratio and the real hit ratio. The real hit ratio, $\widetilde{h_r^k}$ is the probability of serving a demand k locally by a cache r. On the other hand, the effective hit ratio, h^{*k}_{r} is the probability of serving k locally by r knowing that it wasn't intercepted by its son in $\mathrm{PG(k)}$. Since $\mathrm{PG(k)}$ are trees, if $r \in \mathrm{PG(k)}$ there is a unique path between r and u. Let $\Pi_r = (r_n, r_{n-1}, \dots, r_1, r_0)$ be this path, where $r_n = r$ and $r_0 = u$. The effective hit ratio load as $h^{*k}_{r} = \begin{cases} 0, & r \notin \mathrm{PG(k)}, \\ \widetilde{h_r^k} \prod_{j=2}^{n} \left(1 - \widetilde{h_{r_{j-1}}^k}\right), & otherwise. \end{cases}$.

Here, $\widetilde{h_r^k}$ is given by $\widetilde{h_r^k} = \begin{cases} 0, & r \notin \mathrm{PG(k)}, \\ \frac{s_r^k}{\sum_{s \in A_r^k} f_{rs}^k}, & \text{otherwise.} \end{cases}$, A_r^k is the set of adjacent nodes of r in $\mathrm{PG(k)}$.

Now, the Path Length refers to the number of intermediate devices through which data must pass between source and destination (number of hops). Calculating the path lengths depends on whether or not there exist caches in the network.

Let $k \in K$ and $\mathrm{PG(k)}$ its subgraph. The path length in the case of no caching is defined as the $\mathrm{PG(k)}$ tree height in terms of links. Indeed, when there are no caches in the network the demand crosses every level to be served by the provider. Now, denote by L_i the different hierarchical levels in our real topology, $i = 1, \dots, d$. The path length with caches in the network load as $hop(k) = \sum_{i=1}^{d} \sum_{L_i \cap \mathrm{PG(k)}} (h^{*k}_{r} \prod_{j=1}^{n} \frac{f_{r_i r_{i-1}}^k}{\sum_{s \in A_{r_i}^k} f_{r_i s}^k})\, i$. Finally, we define an average path length for the whole network as $hop = \sum_{k \in K} \frac{D_k}{\sum_{k \in K} D_k}\, hop(k)$.

Figure 3 shows a comparative study based on path length and average path length metrics. We investigate the cases with and without caches in content-oriented network. We observe that for $h_1 = 0.4, h_2 = 0.3, h_3 = 0.2$ and $\beta = 0.3$, introducing caches provides a gain of 56.11% in terms of average distance. We can even reach 73 %of gain in term of distance with 4 of the demands.

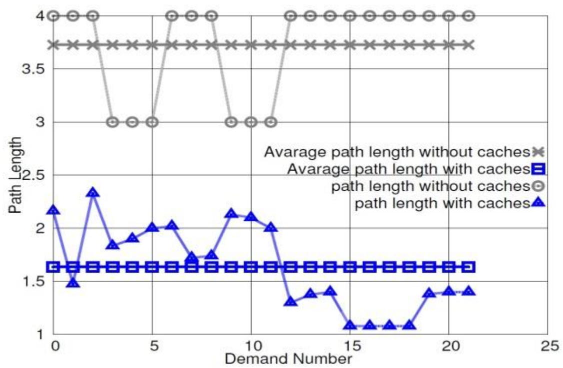

Figure 3: Path and Average Path Lengths of the Instance $\boldsymbol{H = (0.4, 0.3, 0.2)}$, $\boldsymbol{\beta = 0.3}$

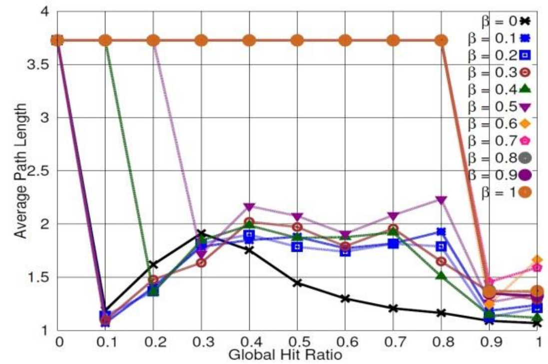

Figure 4: Impact of the Cache Parameters on Average path length

We exemplify now the impact of parameters of the cache. We look into how the average path length differs on changing values of the cache hit ratio h_r and of the cache power usage β. Figure 4 shows the gain in terms of distance provided by the introduction of caches in the network. Maximum gain of 76.49 % of the average path length, is obtained for $\widetilde{H} = 0.1$ and $\beta \leq 0.3$.

5.2 The MILP Model Compared to a Routing on Shortest Path-Based Caching Model

In this section we give a comparison between our MILP model and the routing on shortest path-based heuristic. In Figure 5 we compare the values of the objective function, while the computation time comparison is displayed in Figure 6. Here, the Gap indicates the percentage by which the solution found by the heuristic is worse and how much time is saved by using it. First, notice that when the network does not contain caches ($\widetilde{H} = 0$), it is possible to solve the MILP to optimality. For $\beta \leq 0.5$ we obtain solutions within at most 20 % of the optimum and save at least 68% of the CPU time. These solutions are more closely to the optimum (at most 5% of the optimum) when the global hit ratio $\widetilde{H} \geq 0.7$. Now for $\beta \geq 0.6$ the heuristic gives solutions within at most 38.72 % of the optimum while saving more than 95% of the CPU time.

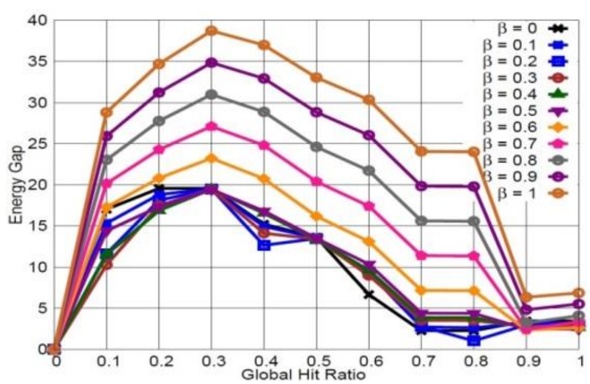

Figure 5: Comparison of the Objective Function Given by The Heuristic and The MILP Model

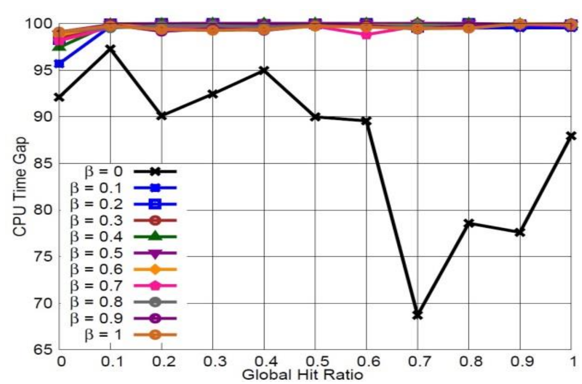

Figure 6: Comparison of the CPU Time Given by The Heuristic and The MILP Model

To conclude we can say that our model is the best choice compared to the heuristic. Certainly, our heuristic is an acceptable choice for the network with $\beta \leq 0.5$ especially with a very short computation time. However, for $\beta \geq 0.6$ the shortest path-based solutions are not a good choice since we can even reach $38.72\ \%$ of energy saving by solving our model.

6 CONCLUSION

We have proposed a new model for saving energy in content-oriented networks by disabling equipment. We have considered a caching object approach and jointly optimized the routing and caching object problem. Our model has been validated by solving instances based on real network topologies. Based on several network performance metrics, we have shown the impacts and the gains of introducing energy-aware on a real telecommunication network. Furthermore, we have proposed a heuristic which has allowed us to show the benefits of our model compared to the classical routing on shortest path-based caching model.

In future work a content analysis will be considered by enabling several types of content. In fact our caches are characterized by the average hit probabilities, indicating the ratio of demands served locally over all contents. An interesting direction is to study the optimization problem with uncertain hit ratio.

7 REFERENCES

[1] **G. Pallis and A. Vakali. 2006.** Insight and perspectives for content delivery networks, Communications of the ACM, vol. 49, no. 1, pp. 101–106.

[2] **V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard. 2009.** Networking named content, in Proceedings of the 5th international conference on Emerging networking experiments and technologies. ACM, pp. 1–12. 2009.

[3] **J. Araujo, F. Giroire, Y. Liu, R. Modrzejewski, and J. Moulierac.** Energy efficient content distribution, IEEE International Conference on Communications, 2013, pp. 4233-4238

[4] **S. Borst, V. Gupta, and A. Walid. 2010.** Distributed caching algorithms for content distribution networks, in INFOCOM, 2010 Proceedings IEEE. IEEE, pp. 1–9.

[5] **M. Mangili, F. Martignon, and A. Capone. 2013.** A comparative study of content-centric and content-distribution networks: Performance and bounds, Proceedings of IEEE Global Communications Conference (Globecom).

[6] **L. Chiaraviglio, M. Mellia, and F. Neri. 2012.** Minimizing ISP network energy cost: formulation and solutions, IEEE/ACM Transactions on Networking, vol.20, no. 2, pp. 463–476.

[7] **K. Guan, G. Atkinson, D. C. Kilper, and E. Gulsen. 2011.** On the energy efficiency of content delivery architectures, in IEEE International Conference on Communications, pp.1–6.

[8] **V. Mathew, R. K. Sitaraman, and P. Shenoy. 2012.** Energy-aware load balancing in content delivery networks, in INFOCOM, Proceedings IEEE. IEEE, pp. 954–962.

[9] **N. Choi, K. Guan, D. C. Kilper, and G. Atkinson. 2012.** In-network caching effect on optimal energy consumption in content-centric networking, IEEE International Conference on Communications, (ICC), pp. 2889–2894.

[10] **Y. Carlinet, E. Gourdin, B. Kauffmann, and Y. Levene, 2011**. Some keys for designing an efficient caching architecture, Orange Labs, Networks and Carriers, Tech. Rep.

[11] “Cplex.” [Online]. Available: http://www.ilog.com/products/cplex/