Energy Efficient Content Locations for In-Network Caching

Satoshi Imai
Fujitsu Laboratories Ltd.
4–1–1 Kamikodanaka, Kawasaki-shi,
Kanagawa, 211–8588 Japan
Email: imai.satoshi@jp.fujitsu.com

Kenji Leibnitz
NICT, CiNet
1–3 Yamadaoka, Suita,
Osaka, 565–0871 Japan
Email: leibnitz@nict.go.jp

Masayuki Murata
Osaka University
1–5 Yamadaoka, Suita,
Osaka, 565–0871 Japan
Email: murata@ist.osaka-u.ac.jp

***Abstract*—As various multimedia services are being provided on networks, broadband traffic is growing as well. Reducing traffic is important because power consumption in networks has been increasing year by year. Meanwhile, content caching is expected to reduce data traffic by storing content replicas on the network nodes and it is beneficial in view of energy efficiency. However, in order to realize an energy efficient network, it is necessary to allocate content replicas effectively in consideration of both power consumption of content caching and transmission of traffic. In this paper, we propose a design method which derives the energy efficient cache locations for content dissemination. Furthermore, we demonstrate the effectiveness of our proposed method.**

I. Introduction

The growing variety of multimedia services provided on networks is leading to an increase in network traffic. As a result, also the power consumption of network systems is increasing year by year. In the reports of the Ministry of Economy, Trade, and Industry [1], broadband traffic in Japan is growing at an annual rate of 25% and network power consumption is expected to occupy 20% of the total ICT power consumption by 2025.

Recently, power-saving mechanisms of network devices [2] have been studied in order to realize *Energy Proportional Networks* [3], [4] in which power consumption of each device is proportional to its usage. In addition, energy efficiency can be improved by reducing the traffic flow in the network, because the traffic decrease can improve the effect of the power-saving mechanism in each device or can prevent that frequent incremental deployments of the network devices are required. As technologies for reducing network traffic, *Content Delivery Network* (CDN) architectures in metro/access networks, such as Akamai-CDN and Web-Proxy, are well known. The CDNs can manage the content delivery at the edge of the networks by allocating content replicas in cache servers which are in geographical proximity to users.

On the other hand, a content dissemination architecture called *Named Data Networking* (NDN) [5], utilizing caching functionality on routers, has recently been proposed. Content-Centric Networks (CCN) for NDN use a receiver-driven protocol where data is only sent in response to a user's request for a content name. A user's request is forwarded on some content routers (*CR*s) until the requested content can be found. When the requested content is found on a *CR*, the content data are transmitted on a reverse route for the request. Furthermore, data are cached on some *CRs* along the transmission route based on the specific replacement strategy such as *Least Recently Used* (LRU), *Least Frequently Used* (LFU) or another effective method [6], [7].

In the content caching, network traffic is influenced by the cache locations because each content generates a different amount of traffic depending on its popularity. Moreover, many storages such as DRAM or SSD are required for content replicas. Therefore, in order to realize an energy efficient resource management for content dissemination, the appropriate cache locations should be realized in consideration of the balance of traffic volume for delivering each content and memory used for storing content replicas.

In this paper, we aim to derive reference locations for energy efficient caching strategies, and propose a design method to minimize the sum of *power consumption of storage devices for content caching* and *power consumption of network devices for content delivery*. Furthermore, we evaluate the energy efficiency in case of realizing optimal content locations for some scenarios.

The remainder of this paper is organized as follows. Section II discusses general issues and sketches the proposed approach followed by Section III which summarizes related work. We describe the design method to solve the cache location problem in Section IV. Section V demonstrates our simulation results and finally we conclude the paper in Section VI.

II. Issues and Approach

Caching strategies can reduce network traffic by storing content replicas in many locations. However, this can result in inefficient storage of content replicas, when the number of requests for content is small. Therefore, in order to realize energy efficiency of the network under the condition that network devices and memory can be deployed in proportion to their usage, the appropriate cache allocation should be executed in consideration of the *power consumption of memory for storing content replicas* and the *power consumption of network devices to accommodate traffic*.

Moreover, most caching structures for IPTV or CCN networks tend to have a logical hierarchy [8], which is constructed

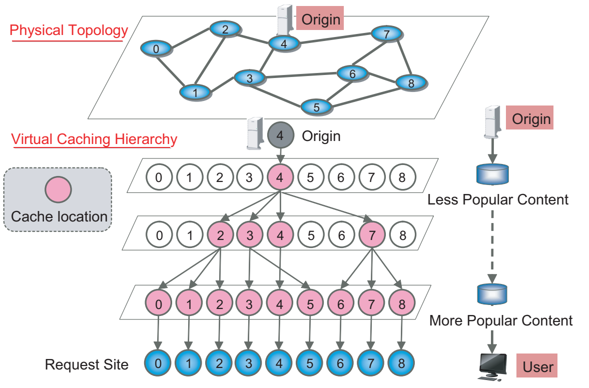


Fig. 1. Caching hierarchy

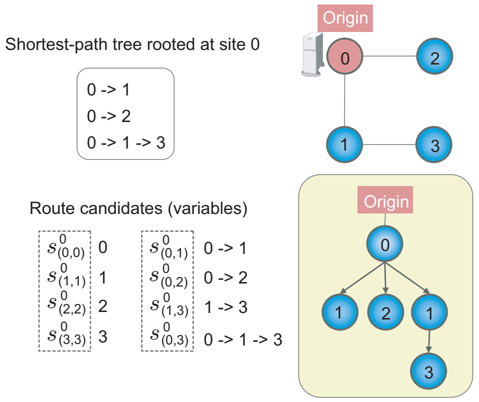


Fig. 2. Example of definition of route candidates

by caching content in some nodes on a tree rooted at an origin site for the content. The caching hierarchy is constructed by routes between an origin site, caching nodes and users, such that

- less popular content is cached on nodes near to the origin site, and
- more popular content is cached on nodes near to users.

Therefore, we propose a design method which can derive energy efficient locations of content on constraints of the caching hierarchy (cf. Figure 1) rooted at the content's origin site so as to minimize the sum of *cache allocation power*, i.e., the memory power consumed by storing the content, and *traffic transmission power*, i.e., the total power consumed by network devices when data are transmitted. As a result, our algorithm can provide reference locations to realize energy efficiency for cache strategies. In this paper, an adaptive cache management to allocate content on the optimal locations is another research topic.

III. Related Work

Recently, energy efficiency in CCN has been attracting a lot of attention [9], [10], [11]. Lee *et al.* [9], [10] survey the energy efficiency of various network devices deployed in access/metro/core networks. Furthermore, they evaluate the power-saving effect in the entire network when the deployment ratio of CCN-enabled edge/core routers is changed. As a result, they show that CCN is able to improve the energy efficiency of current CDN.

Furthermore, Guan *et al.* [11] build energy models of traffic transmission power and caching power for content delivery architectures such as "Conventional and decentralized server-based CDN", "Centralized server-based CDN using dynamic optical bypass", and CCN. Using their energy models, they analyze the energy efficiency of each of these architectures. Those energy models are approximations based on the relations between the topological structure and the average hop-length from all sites to the nearest cache location. The appropriate number of cache locations for content can be estimated in order to save the network power consumption when the target topology, the required average hop-length, and the number of requests for content are given.

In those papers energy efficiency is represented differently depending on the cache locations for content.

Meanwhile, traditionally there are content placement algorithms [12], [13], [14], [8] as a solution for *File Allocation Problems* [15] which minimize cost imposed for content storage and queries (requests), or maximize performance such as distance to content. Baev *et al.* [12] propose an Integer Linear Programming (ILP) model which minimizes content placement cost and an approximation solution using a linear relaxation. Furthermore, Qui *et al.* [13] develop and compare with some replica placement algorithms to solve a K-median problem for CDNs.

In contrast to the above-mentioned content placement problems, Leff *et al.* [14] propose a distributed algorithm based on coordinating local information and local searching. Borst *et al.* [8] formulate an ILP model based on a hierarchical structure for content locations to minimize bandwidth costs and a distributed solution. Moreover, they evaluate the cost-saving effect for a hierarchical topology which has symmetric bandwidth cost for a parent node and some leaf nodes. However, these works don't discuss an energy efficient design for content locations considering that some caching hierarchies, which have different origin sites for content and asymmetric routes, are multiplexed in a target network.

Therefore, we propose a new ILP model to design the most energy effective cache locations in consideration of the multiplexed caching hierarchies for content dissemination.

IV. Proposed Algorithm

In this paper, we aim to design content locations in order of descending popularity so as to minimize total power consumption based on *Energy Proportional Networks* The proposed algorithm derives the optimal cache locations of a content item based on a 0-1 ILP. Moreover, we consider multiplexing cache hierarchies which have a different origin site for content. Recently, there are several solvers available to solve large-scale ILP problems at high speed. Therefore, they can be used to solve our 0-1 ILP model.

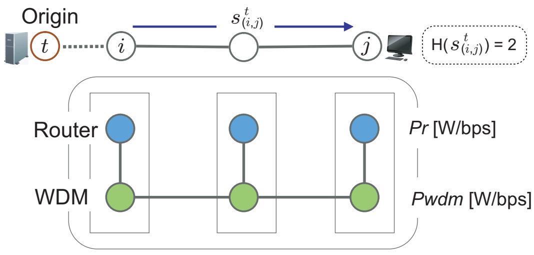


Fig. 3. Network model

A. Design Flow

The cache location design for each content is executed in the following steps.

S1. Get content popularity.

S2. Select a target content (k-th popular content) in the order of content popularity.

S3. Extract route candidates to deliver the target content from a cache location to a requesting user on the shortest-path tree rooted at the origin site of the target content and define the route candidates as the design variables, cf. Figure 2.

S4. Design the optimal cache locations of the target content based on the route candidates in consideration of the pre-designed routes for more popular content which have the same origin site as that of the target content.

S5. Return to **S2** and execute the design of the next popular content ($[k+1]$-th popular content).

B. Objective Function

We consider a network composed of *CRs* and Wavelength Division Multiplexing (WDM) nodes in Figure 3 and we design cache locations and content delivery routes for the delivery tree, which consists of a set of vertices (sites) V and route candidates $\mathrm{R_c}$. The objective is to minimize the sum of *cache allocation power* $Ca_{k,i}$, i.e., the power for storing the (k-th popular) target content in *CR* on site i, and *traffic transmission power* $Tr_{k,(i,j)}$, i.e., the total power of routers and WDM nodes when the target content is delivered on the route from site i to site j.

$$\textbf{Minimize} \sum_{i\in \mathrm{V}} \{Ca_{k,i} \cdot u_i\} + \sum_{(i,j)\in \mathrm{R_c}} \left\{Tr_{k,(i,j)} \cdot s^t_{(i,j)}\right\} \quad (1)$$

The design variables are defined as follows.

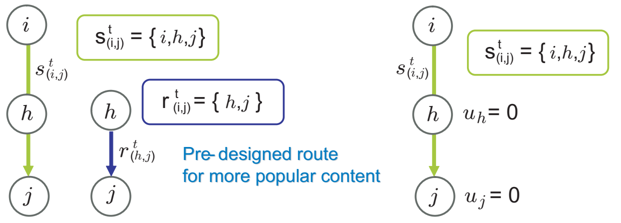
- $u_i \in \{0,1\}$ indicates whether or not to store the target content in *CR* on site i.
- $s^t_{(i,j)} \in \{0,1\}$ describes whether or not to select the route from site i to site j defined on the shortest-path tree rooted at the origin site t of the (k-th popular) target content.

1) Cache Allocation Power: Power consumption $Ca_{k,i}$ when the target content is cached in *CR* on site i ($u_i = 1$) is defined as

$$Ca_{k,i} = D_k \cdot P_{ca}, \quad (2)$$

TABLE I
VARIABLES IN THE PROPOSED MODEL

Variable	Design/ Given	Definition
u_i	Design	Binary variable for whether to store a designed content (k-th popular content) in *CR* on a site i or not
$s^t_{(i,j)}$	Design	Binary variable for whether to select the route from site i to site j defined on a shortest-path tree rooted at origin site t of target content (k-th popular content) or not
P_{ca}	Given	Power density for storage [W/byte]
P_r	Given	Power density of a router [W/bps]
P_{wdm}	Given	Power density of a WDM node [W/bps]
D_k	Given	Data size of target content (k-th popular content) [bytes]
B_k	Given	Transmission rate, i.e., required throughput of target content (k-th popular content) [bps]
$R_{k,j}$	Given	The number of requests for target content (k-th popular content) from a destination site j



(a) Hierarchical route (b) Cache location

Fig. 4. Constraint conditions

where P_{ca} is the memory power density [W/byte] and D_k is the data size [bytes] of the target content (k-th popular content).

2) Traffic Transmission Power: As shown in Figure 3, the power consumption when the target content (k-th popular content) is delivered on the candidate route $s^t_{(i,j)}$ from site i to site j on the shortest-path tree rooted at origin site t, is defined as

$$Tr_{k,(i,j)} = B_k \cdot R_{k,j} \cdot (P_r + P_{wdm}) \cdot \left\{\mathrm{H}(s^t_{(i,j)}) + 1\right\}, \quad (3)$$

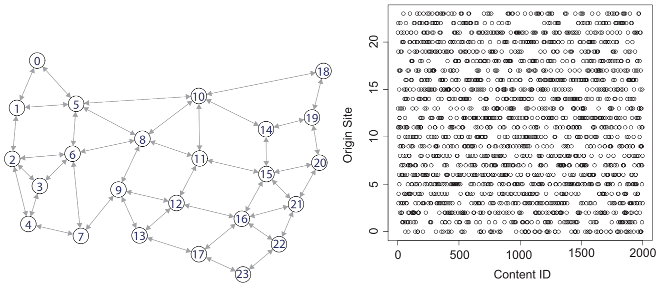
where P_r and P_{wdm} are the power densities [W/bps] of a router and of a WDM node, respectively, and B_k and $R_{k,j}$ are the content delivery rate and the number of requests from site j of the target content (k-th popular content). Furthermore, we define $\mathrm{H}(s^t_{(i,j)})$ as the hop-length of route $s^t_{(i,j)}$. All variables and functions in the proposed model are summarized in Table I.

C. Constraint Conditions

We now define the constraints for the proposed 0-1 ILP model.

- **Route selection constraint:** The transmission routes to site j, which requests the target content having origin root t, should be created.

$$\sum_{i\in \mathrm{V}} s^t_{(i,j)} = 1 \quad \forall j \in \mathrm{V} \quad (4)$$



(a) Test topology

(b) Origin site of content

Fig. 5. Test Conditions

TABLE II
POWER DENSITY PARAMETERS.

Device (Product)	Power / Spec	Power Density
DRAM	10 W / 4 GB	$P_{ca} = 2.5 \times 10^{-9}$ W/byte
Router (CRS-1)	4185 W / 320 Gbps	$P_r = 1.3 \times 10^{-8}$ W/bps
WDM (FLASHWAVE9500)	800 W / 480 Gbps	$P_{wdm} = 1.67 \times 10^{-9}$ W/bps

- **Relation constraint between cache location and transmission route:** The starting site of the transmission route should be the cache location of the target content.

$$s^t_{(i,j)} \leq u_i \quad \forall i \in \mathrm{V},\ \forall (i,j) \in \mathrm{R_c} \tag{5}$$

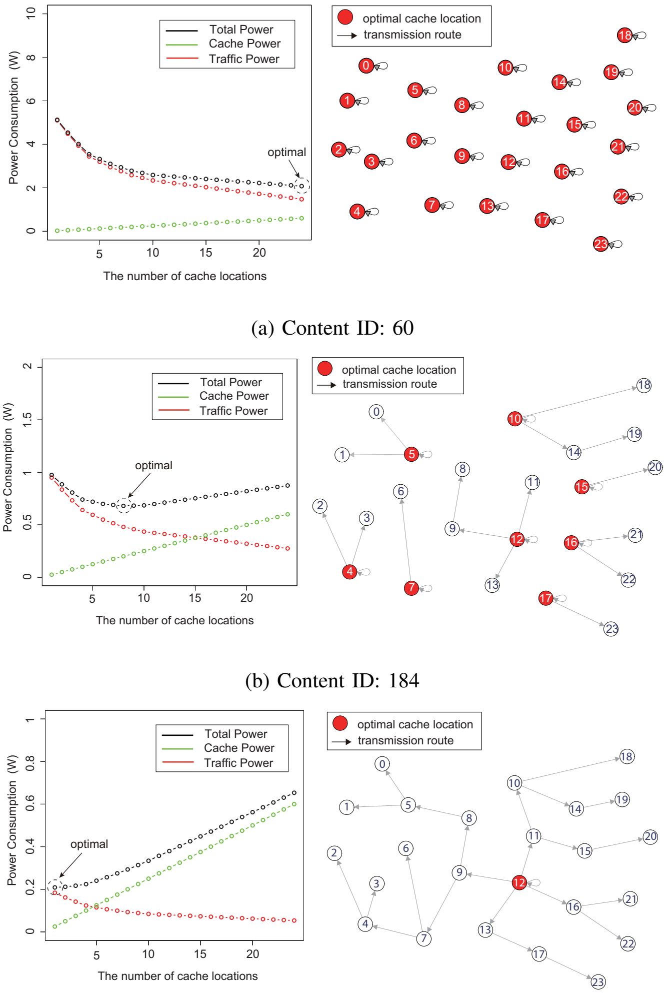
- **Hierarchical route constraint:** The target content should be cached on the shortest-path tree rooted at origin site t of that. Furthermore, there should be more popular content, having the same origin site t as the target content, on the designed transmission route and caching hierarchy should be constructed as shown in Figure 1.

$$s^t_{(i,j)} = 0 \quad \begin{array}{l} \text{if } \mathbf{r}^t_{(h,j)} \notin \mathbf{s}^t_{(i,j)} \\ \vee\ \mathbf{s}^t_{(i,j)} \in \left\{\mathbf{r}^t_{(h,j)} - h\right\} \end{array} \quad \forall (i,j) \in \mathrm{R_c} \tag{6}$$

In Figure 4(a), $\mathbf{s}^t_{(i,j)}$ is the route sequence along the route $s^t_{(i,j)}$ of the target content (k-th popular content). Meanwhile, $\mathbf{r}^t_{(h,j)}$ is the route sequence along the route of more popular content (m-th popular content: $m < k$) having the same origin site t as the target content. Furthermore, site h represents the starting site of the route sequence $\mathbf{r}^t_{(h,j)}$ and $\left\{\mathbf{r}^t_{(h,j)} - h\right\}$ represents the subsequence excluding the starting site h from the route sequence $\mathbf{r}^t_{(h,j)}$.

- **Cache location constraint:** The same replicas should not be cached on the designed transmission route of the target content (k-th popular content). This constraint is illustrated in Figure 4(b).

$$s^t_{(i,j)} + u_k \leq 1 \quad \forall (i,j) \in \mathrm{R_c},\ k \in \left\{\mathbf{s}^t_{(i,j)} - i\right\} \tag{7}$$



(a) Content ID: 60

(b) Content ID: 184

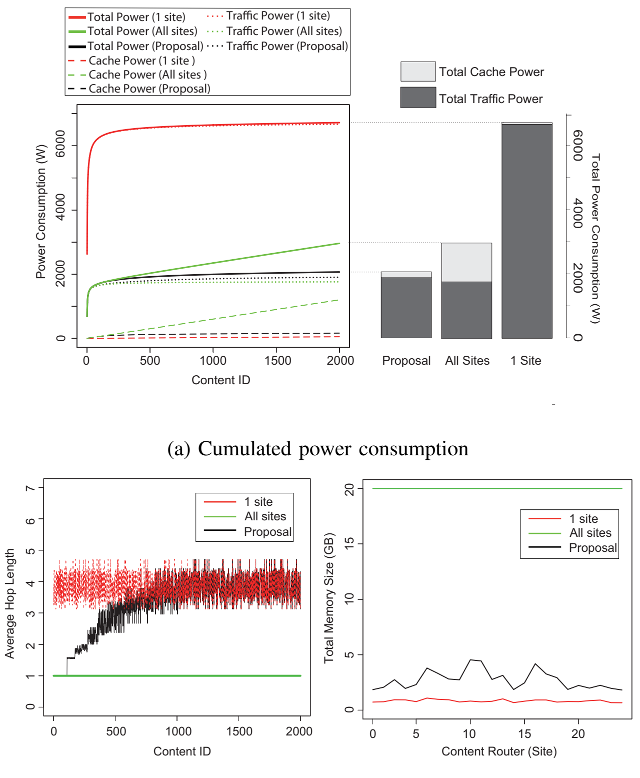
(c) Content ID: 550

Fig. 6. Tradeoff between caching power and traffic transmission power (left) / optimal cache locations (right) in Evaluation 1

V. NUMERICAL EVALUATIONS

We evaluated energy efficiency to realize the optimal cache locations for content using the proposed algorithm. The simulation conditions are set to the following.

- **Test networks:** IP-backbone topology with 24 sites (cf. Figure 5(a)).
- **Content information:** Zipf-distributed requests from each site j for 2000 content items ($K = 2000$) are defined as $R_{k,j} = \lambda \cdot k^{-\alpha}/c$ ($c = \sum_{k=1}^{K} k^{-\alpha}$, $\alpha = 1.5$ [16]), where the total number of requests λ is set to 500. As we further show in Figure 5(b), we set the origin site t of content ID k randomly based on a uniform distribution. The data size D_k of content ID k is set to 10 Mbytes, such as the average size of user generated content [17]. The delivery rates B_k of content ID k are defined as 1, 10, and 20 Mbps.
- **Power density of each device:** The power density of a memory device [W/byte] and a router or WDM node [W/bps] are set to the values given in Table II.



(a) Cumulated power consumption

(b) Average hop-length

(c) Memory usage

Fig. 7. Simulation results in Evaluation 2

Based on the above-mentioned conditions, we evaluated the following points.

- **Evaluation 1:** Tradeoff between caching power and traffic transmission power.
- **Evaluation 2:** Differences of characteristics (such as power consumption, hop-length, memory usage) using three caching policies: caching on the origin site of content, caching on all sites, and the proposed method.
- **Evaluation 3:** Effectiveness of the proposed method when the delivery rate of the content items is changed.

A. *Evaluation 1*

We first verify the tradeoff between the caching allocation power and the traffic transmission power. In these simulations, we assume the delivery rate of content is $B_k = 10\,\text{Mbps}$, and add the following condition in the proposed model, which specifies the number of replicas of the target content.

The number of cache locations constraint:

$$\sum_{i \in \mathrm{V}} u_i = M \tag{8}$$

For the specific content IDs: 60 (more popular content), 184, and 550 (less popular content) having the same origin site 12, the charts in Figure 6 show the changes of the cache allocation power, the traffic transmission power, and the total power, when the number of cache locations M is changed.

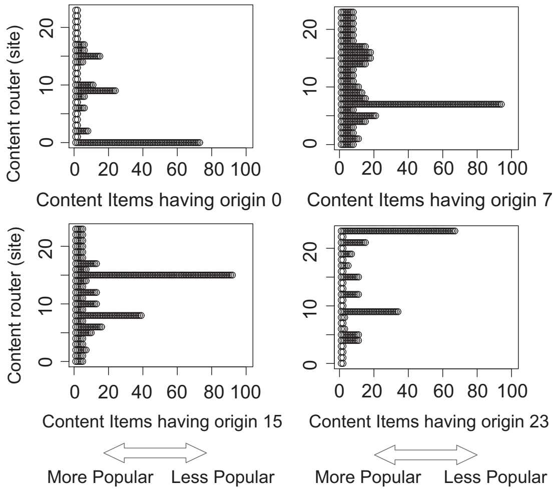


Fig. 8. Caching hierarchy in Evaluation 2

In these results, the cache allocation power increases and the transmission power decreases with locations M. These charts show the tradeoff relationship between the cache allocation power and the transmission power.

On the other hand, the network topologies in Figure 6 show the optimal cache locations of content IDs 60, 184, and 550. Therefore, we see the number of cache locations to minimize the total power consumption changes according to the content popularity.

B. *Evaluation 2*

We now compare the characteristics of power consumption, hop-length and memory usage by the following caching policies when the delivery rate per content is set to $B_k = 10\,\text{Mbps}$.

- **Caching on 1 site:** The replicas of content are stored in a *CR* on its origin site.
- **Caching on all sites:** The replicas of content are stored in all *CR*s.
- **The proposed method:** The replicas of content are stored in *CR*s on the designed sites to minimize the power consumption.

Figure 7(a) shows the cumulative power consumption until the time when the specific content is allocated in the network by applying each considered caching policy to the test topology. These results demonstrate that the proposed method realizes more power-effective cache allocation than the other policies. In the caching policy on a single *CR*, the changes of the traffic transmission power are dominant in the total power consumption. On the other hand, the cache allocation power in the caching policy on all *CR*s is larger than the others.

Figure 7(b) presents the average hop-length from each site to the nearest replicas. If the designed content is less popular, the average hop-length is increased as a result of reducing the number of cache locations. Figure 7(c) shows the memory usage when all content items are allocated. The memory usage for the proposed method is significantly smaller than that for caching on all *CR*s, but only slightly larger than caching on a single *CR*.

In the simulation results, we can see that the proposed algorithm specifies power-effective cache locations by balancing between the cache allocation power and the traffic transmission power. Additionally, it is demonstrated that the proposed algorithm can construct the caching hierarchy as shown in Figure 1. Further results using the proposed algorithm are shown in Figure 8. Figure 8 presents the results of cache locations for content having origin sites 0, 7, 15, and 23. These results show that the caching hierarchy is rooted at the origin site of content.

C. Evaluation 3

We next evaluate the characteristics of power consumption, hop-length, and memory usage when the content delivery rate is changed to $B_k = 1, 10,$ and 20 Mbps. Figure 9(a) shows the results of the cumulative power consumption and Figure 9(b) shows the average hop-length. The memory usage on each *CR* is shown in Figure 9(c).

As the delivery rate of content becomes higher, the traffic transmission power becomes larger, which leads to the following three observations.

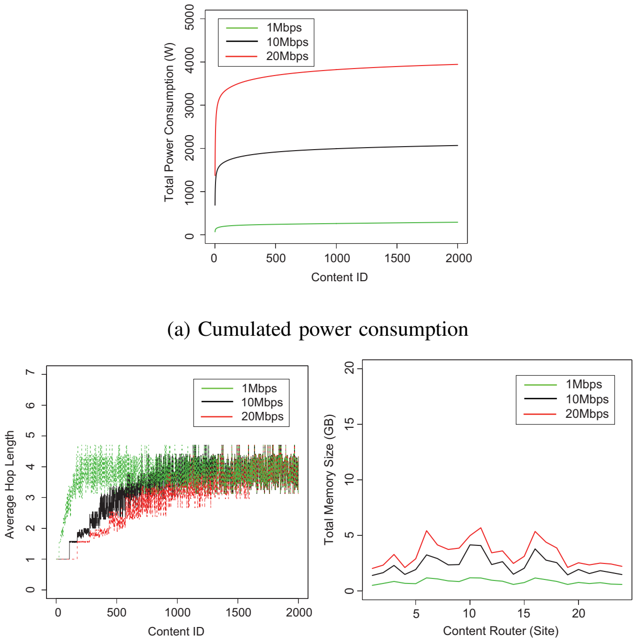
- The cumulative power consumption becomes larger as shown in Figure 9(a).
- The average hop-length is shifted backward to less popular content as shown in Figure 9(b).
- The memory usage becomes larger as shown in Figure 9(c).

Therefore, we can see that our algorithm suggests it is more energy efficient to store content replicas on more locations for content having a higher delivery rate.

VI. Conclusions

We proposed an energy efficient design method to derive the optimal cache locations according to the content popularity. The proposed algorithm can consider the tradeoff between the cache allocation power and traffic transmission power under the constraints of the caching hierarchy.

We showed the energy efficiency of our proposed algorithm by taking into account delivery rate of content and request distribution for each site. Although we only studied uniform traffic, the proposed algorithm is also applicable to heterogeneous conditions. Moreover, our algorithm can provide reference of cache locations for evaluating energy efficiency for cache strategies. In the future, we also plan on adding an adaptive cache management mechanism to realize the optimal locations of content for in-network caching.



(a) Cumulated power consumption

(b) Average hop-length

(c) Memory usage

Fig. 9. Simulation results in Evaluation 3

References

[1] T. Hoshino, “Expectations on Innovative Energy-saving Technologies of Information and Communication Equipment”, Ministry of Economy, Trade and Industry, Green IT Symposium, 2007.

[2] M. Gupta and S. Singh, “Greening of Internet”, in Proc. of ACM SIGCOMM’03, pp. 19–26, Karlsruhe, Germany, 2003.

[3] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, “A Power Benchmarking Framework For Network Devices”, in Proc. of NETWORKING’09, vol. 5550, pp. 795–808, Aachen, Germany, 2009.

[4] T. Harder, V. Hudlet, Y. Ou, and D. Schall, “Energy Efficiency is not Enough, Energy Proportionality is Needed!”, in Proc. of DASFAA’11, pp. 226-239, Hong Kong, China, 2011.

[5] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, “Networking Named Content”, in Proc. of CoNEXT 2009, Rome, 2009.

[6] G. Carofiglio, V. Gehlen, and D. Perino, “Experimental Evaluation of Memory Management in Content-Centric Networking”, in Proc. of IEEE International Conference on Communications (ICC), Kyoto, Japan, 2011.

[7] Z. Li and G. Simon, “Time-Shifted TV in Content Centric Networks: The Case for Cooperative In-Network Caching”, in Proc. of the 4th IEEE International Conference on Communications (ICC) Workshop on Green Communications, Kyoto, Japan, 2011.

[8] S. Borst, V. Gupta, and A. Walid, “Distributed Caching Algorithms for Content Distribution Networks”, in Proc. of INFOCOM 2010.

[9] U. Lee, I. Rimac, D. C. Kilper, and V. Hilt, “Toward energy-efficient content dissemination”, IEEE Network, vol. 25, no. 2, pp. 14–19, 2011.

[10] U. Lee, I. Rimac, and V. Hilt, “Greening the internet with content-centric networking”, in Proc. of e-Energy, pp. 179–182, Passau, Germany, 2010.

[11] K. Guan, G. Atkinson, and D. C. Kilper, “On the Energy Efficiency of Content Delivery Architectures”, in Proc. of the 4th IEEE International Conference on Communications (ICC) Workshop on Green Communications, Kyoto, Japan, 2011.

[12] I. Baev, R. Rajaraman , and C. Swamy,“Approximation Algorithms for Data Placement in Arbitrary Networks”, in Proc. of the 12th ACM-SIAM symposium on Discrete algorithms (SODA), 2001.

[13] L. Qiu, V. N. Padmanabhan, and G. M. Voelker, “On the Placement of Web Server Replicas”, in Proc. of INFOCOM, 2001.

[14] A. Leff, J. L. Wolf, and P. S. Yu, “Replication Algorithms in a Remote Caching Architecture”, IEEE Transactions on Parallel and Distributed Systems, vol. 4, no. 11, 1993.

[15] L. W. Dowdy, and D. V. Foster, “Comparative Models of the File Assignment Problem”, Journal ACM, Volume 14, Issue 2, 1982.

[16] D. Rossi and G. Rossini, “Caching performance of content centric networks under multi-path routing (and more)”, Technical report, Telecom ParisTech, 2011.

[17] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, “Youtube traffic characterization: a view from the edge”, in Proc. of IMC’07, pp. 15–28, San Diego, USA, 2007.