ComEX

IEICE Communications Express, Vol.8, No.8, 318–322

Optimally designing virtualized CDN

Noriaki Kamiyama[a)] **and Yutaro Hosokawa**

Faculty of Engineering, Fukuoka University,
Nanakuma, Jonan-ku, Fukuoka 814–0180, Japan
a) *kamiyama@fukuoka-u.ac.jp*

Abstract: Using virtualized CDN such as Amazon CloudFront in which cache servers are provided on virtual machines, content providers can flexibly select the regions of using cache servers according to the geographical demand pattern of their users, so they can expect to reduce the cost of using CDN. However, to maximize the profit of content providers, they need to carefully select the locations of cache servers which affect the quality perceived by users as well as the total cost. We propose a method of optimally selecting the geographical regions to use cache servers maximizing the profit of content providers.

Keywords: virtualized CDN, optimum design, profit
Classification: Network

References

[1] N. Herbaut, D. Negru, Y. Chen, P. A. Frangoudis, and A. Ksentini, "Content delivery networks as a virtual network function: A Win-Win ISP-CDN collaboration," GLOBECOM, 2016. DOI:10.1109/GLOCOM.2016.7841689
[2] N. Kamiyama and Y. Hosokawa, "Optimally designing virtualized CDN maximizing profit of content providers," IEEE CCNC, 2019. DOI:10.1109/CCNC.2019.8651800
[3] R. Ma, J. Wang, and D. M. Chiu, "Paid prioritization and its impact on net neutrality," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 2, pp. 367–379, Feb. 2017. DOI:10.1109/JSAC.2017.2659020
[4] H. Chang, S. Jamin, and W. Willinger, "To peer or not to peer: Modeling the evolution of the internet's AS-level topology," IEEE INFOCOM, 2006. DOI:10.1109/INFOCOM.2006.236
[5] A. Dhamdhere and C. Dovrolis, "Can ISPs be profitable without violating network neutrality," ACM NetEcon, 2008. DOI:10.1145/1403027.1403031
[6] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305–1314, Sept. 2002. DOI:10.1109/JSAC.2002.801752
[7] Amazon CloudFront Pricing, On-Demand Pricing, https://aws.amazon.com/cloudfront/pricing/.

1 Introduction

Content provider (CP), e.g., YouTube, has its special requirement for the design of Content delivery network (CDN) depending on the geographical demand pattern

DOI: 10.1587/comex.2019XBL0063
Received March 28, 2019
Accepted May 8, 2019
Publicized May 24, 2019
Copyedited August 1, 2019

against the content items. However, CDN providers do not distinguish CPs when inserting and removing content items in cache servers, i.e., applying the identical cache-insertion and cache-replacement policy to content items of all CPs, and CPs cannot limit the locations where content items are cached to reduce the cost paid to the CDN providers. Therefore, the virtualized CDN or cloud-based CDN in which cache servers are provided by VMs on public cloud have gathered wide attention and investigated [1], and commercial virtualized CDN services, e.g., Amazon CloudFront, have been already provided by major cloud providers. In the virtualized CDN, CPs directly contract with the cloud providers, and each CP can flexibly construct its own CDN by itself. In other words, each CP becomes a CDN provider serving just its content items.

Before starting to use the virtualized CDN, CPs will face the problem of optimally design its CDN, i.e., selecting the regions of using cache servers. CPs can expect to minimize the cost of using CDN by limiting the regions for using cache servers. However, the user-perceived quality, e.g., delay and smoothness in playback of video, will be degraded for users in these regions without cache servers. The number of views of users with degraded service quality will decrease, and the revenue of CPs will also decrease. Therefore, to maximize the profit of CPs, CPs should carefully select the regions to use cache servers considering both the fee paid to the cloud provider and the user-perceived quality. In this letter, we propose a method of optimally designing the virtual CDN maximizing the profit of CPs considering both the user-perceived quality and the cost of using cache servers when geographical demand distribution is given[1].

2 Assumptions

2.1 Virtualized CDN service

We assume one virtualized CDN service which provides cache servers at N regions, and we consider just a single CP. Let R_n denote region n, and we define C_n as the cache servers provided in R_n. Let x_n denote a binary variable which takes unity if CP decided to use C_n or takes zero if CP decided not to use C_n. Moreover, we define $\boldsymbol{X}$ as the set of R_n with $x_n = 1$. In other words, $\boldsymbol{X}$ is the set of regions where CP decided to use cache servers. The origin servers are provided in just a single region R_o, and CP does not use C_o, i.e., $x_o = 0$, because content can be delivered to users of R_o from origin servers. For requests from users of R_n, content is delivered from C_n if requested content exists in C_n, i.e., cache hit, and content is delivered from R_o if requested content does not exist in C_n, i.e., cache miss. When users of R_n with $x_n = 0$ request content, content is delivered from the nearest cache servers of $\boldsymbol{X}$ or R_o giving the minimum average RTT to the requesting users.

2.2 User demand

We assume that each user of CP views content W times on average within a month with totally satisfying the quality of video streaming service. It is anticipated that the number of views of users will decrease as the quality of video streaming is

[1]A previous version of this manuscript was presented at [2]. This letter extended [2] by reflecting the cache hit ratio in the optimization problem.

DOI: 10.1587/comex.2019XBL0063
Received March 28, 2019
Accepted May 8, 2019
Publicized May 24, 2019
Copyedited August 1, 2019

degraded. Therefore, we assume that $w(d)$, the average view count of each user when the RTT is d, is given by

$$w(d) = W\mathrm{e}^{-\alpha d}, \tag{1}$$

where α is a sensitivity parameter of users against delay [3]. As α increases, $w(d)$ decreases more sharply with increase of d.

2.3 Cost of delivering content from cache servers

The unit price of delivering content from cache servers to users depends on both the region of cache servers and the total amount of data transmitted from the cache servers within one month. We define $p_n(v)$ as the unit price when delivering content from C_n to users when the total amount of data transmitted from C_n is v, and the total cost of delivering content from cache servers is given by $\sum_{i\in X} p_i(v_i)v_i$.

2.4 Cost of delivering content from origin servers

Many ISPs charge transit fee to CPs based on the total amount of data transmitted on transit links within one month. According to the analysis of transit fee of 20 ISPs in USA in 2004, monthly transit fee $T(v)$ can be approximated by $T(v) = \epsilon \mathrm{v}^{0.75}$ where v is the monthly traffic volume transmitted on transit links [4]. Many ISPs use the 95 percent value of data transmission rate in each five-minutes bin as v, and we assume that v is three times of average data transmission rate [5]. We set $\epsilon = 0.165$ USD which is the average transit fee for 1 Mbps in 2018.

3 Optimum design method of virtualized CDN

Let z_u denote the region from which content items are delivered to users of R_u, and z_u is given by

$$z_u = \arg\min_{z\in \boldsymbol{X}\cup R_o} d_{z,u}, \tag{2}$$

where $d_{i,j}$ is the average RTT between R_i and R_j. Let M denote the number of content items, B denote the average size of content, q_n denote the number of subscribers of CP in R_n, E_n denote the storage capacity of C_n, and r_m denote the request ratio for content m. From (1), v_i, the average amount of data transmitted from C_i of $x_i = 1$ to users of CP within one month, is obtained by

$$v_i = \sum_{u\in X\cap\{u;\, z_u=i\}} \sum_{m=1}^{M} q_u r_m W\mathrm{e}^{-\alpha d_{i,u}} B h_i(m), \tag{3}$$

where $h_i(m)$ is the cache hit probability of content m at C_i. Using Che's equation [6], $h_i(m)$ is obtained by

$$h_i(m) \cong 1 - \mathrm{e}^{-r_m t_c} \tag{4}$$

where t_c is the solution of t in equation $\sum_{m=1}^{M}(1-\mathrm{e}^{-r_m t}) = E_n$. v_o, the amount of data transmitted from origin servers within one month, is

DOI: 10.1587/comex.2019XBL0063
Received March 28, 2019
Accepted May 8, 2019
Publicized May 24, 2019
Copyedited August 1, 2019

$$v_o = \sum_{u \in \boldsymbol{X} \cap \{u;\, z_u = o\}} \sum_{m=1}^{M} q_u r_m W \mathrm{e}^{-\alpha d_{o,u}} B + \sum_{u \in \boldsymbol{X} \cap \{u;\, z_u \neq o\}} \sum_{m=1}^{M} q_u r_m W \mathrm{e}^{-\alpha d_{i,u}} B(1 - h_{z_u}(m)). \quad (5)$$

We assume that CP obtains r fee for each view, and $R(\boldsymbol{X})$, the revenue of CP, is given by

$$R(\boldsymbol{X}) = \sum_{u=1}^{N} q_u r W \sum_{m=1}^{M} \{\mathrm{e}^{-\alpha d_{z_u,u}} r_m h_{z_u}(m) + \mathrm{e}^{-\alpha d_{o,u}} r_m (1 - h_{z_u}(m))\}. \quad (6)$$

$P(\boldsymbol{X})$, the average profit of CP within one month, is obtained by

$$P(\boldsymbol{X}) = R(\boldsymbol{X}) - \sum_{i \in \boldsymbol{X}} p_i(v_i) v_i - T(v_o). \quad (7)$$

We define the optimization problem maximizing $P(\boldsymbol{X})$:

$$\max \quad P(\boldsymbol{X}) \quad (8)$$

$$s.t. \quad x_n = \{0, 1\}, \quad 1 \leq n \leq N, \quad (9)$$

$$x_o = 0. \quad (10)$$

The computational complexity to solve this optimization problem is $O(N^2)$, and CP can obtain the optimum $\boldsymbol{X}$ in quite short time.

4 Numerical evaluation

4.1 Evaluation conditions

As the virtualized CDN service, we assume Amazon CloudFront which provides cache servers at eight regions: North America, Europe, South Africa, Japan, Australia, Singapore, India, and South America. To obtain the setting values of $d_{i,j}$, we measured the average RTT using *ping* command on the virtual machines in these regions of Amazon EC2. However, we cannot setup virtual machines in Africa on Amazon EC2, we selected the seven regions excluding Africa. We assume that delay was zero when delivering servers and users existed in the same region. We set $p_n(v)$ according to the price list of Amazon CloudFront [7]. The ratio of users in Netflix was about 0.03, so we set q_n to the population in each region multiplied by 0.03. We set B to 3 GB, W to 15, and r to 0.6 USD which was obtained by monthly flat-fee of Netflix, 10 USD, divided by W. Moreover, we set the content count M to 1,000, and we set the identical size E to E_n, the storage size of C_n.

4.2 Monthly profit of CP

We compare P, the average monthly profit of the CP, in the proposed design method as well as the two cases: *No Cache* and *All Cache*. No Cache is the case without using any cache servers, and All Cache is the case using cache servers in all the regions excluding R_o. Fig. 1 plots P in each of the three methods against the cache size E for two values of α when placing the origin servers in North America. The following tendencies were also observed when placing the origin servers in other areas as well. In the proposed method, as the result of optimally selecting the

DOI: 10.1587/comex.2019XBL0063
Received March 28, 2019
Accepted May 8, 2019
Publicized May 24, 2019
Copyedited August 1, 2019

locations to use cache servers, P was the largest among the three methods in the entire range of E and α. For example, when $\alpha = 3$ and $E = 1{,}000$, the proposed method increased P about 6% and about 13% compared with All Cache and No Cache, respectively.

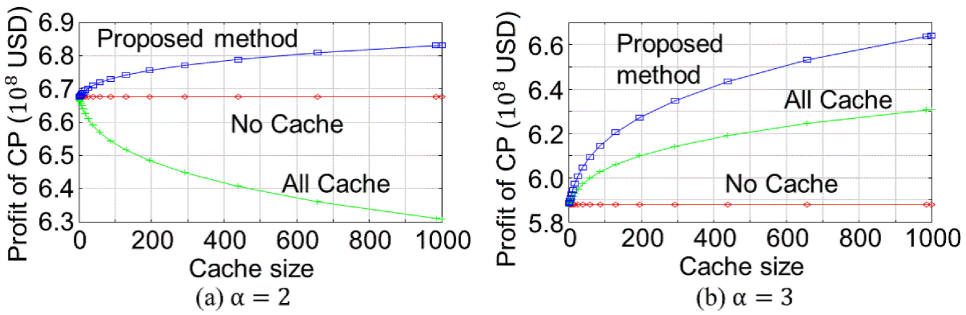

Fig. 1. Average monthly profit of CP against cache size when placing origin servers in North America

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 18K11283.

DOI: 10.1587/comex.2019XBL0063
Received March 28, 2019
Accepted May 8, 2019
Publicized May 24, 2019
Copyedited August 1, 2019