VIRTUAL CACHE & VIRTUAL WAN ACCELERATOR FUNCTION PLACEMENT FOR COST-EFFECTIVE CONTENT DELIVERY SERVICES

Shin-ichi Kuribayashi

Department of Computer and Information Science, Seikei University, Japan

ABSTRACT

The algorithm to determine the place where network functions are located and how much capacities of network function flexibly are required is essential for economical NFV (Network Functions Virtualization)-based network design. The authors proposed a placement algorithm of virtual routing function and virtual firewall function in the NFV-based network for minimizing the total network cost and developed the effective allocation guidelines for these virtual functions.

This paper proposes an NFV-based virtual cache placement algorithm for cost-effective content delivery service such as video streaming, which judges the optimal placement of the cache per content, not on a virtual machine (VM) like the conventional CDN (virtual CDN). Moreover, the content is dynamically cached at the time of first content delivery like ICN (Information-Centric Networking) without placing the cache in advance like CDN .Our evaluation results revealed that the proposed algorithm could reduce total network costs by about 15% compared with CDN .Even if the content cache is deployed economically, performance will degrade if the latency between the content cache and user terminals is long. In order to prevent such a case, this paper also propose a cost-effective placement method of NFV-based WAN accelerator function.

KEYWORDS

NFV, resource allocation, virtual cache, virtual WAN accelerator, content delivery services, minimum network cost

1.INTRODUCTION

In an NFV-based network [1]-[3], a variety of network functions is implemented in software on general-purpose servers. This makes it possible to select the capacity and location of each function without any physical constraints. It is essential to optimize the location and capacity of each network function for economical NFV-based network design. A new placement method should be required as the existing placement method that has limitations on capacity and placement location cannot be used as it is. The authors proposed a virtual function allocation algorithm that minimizes network cost, focusing on two important network functions: routing and firewall, and developed the effective functional allocation guidelines [4],[5]. The authors also clarified the influence of QoS conditions such as the maximum allowable network delay on its placement guidelines [6].

DOI: 10.5121/ijcnc.2020.12204

The main objective of References [4]-[6] and of this paper is not to solve the VNE (Virtual Network Embedding) problem, which tries to allocate the virtual network function efficiently as a short-term perspective [7]-[8]. The network design guidelines show the trend for the capacity required for each network function and the optimal location of each function and provide network operators with critical information needed in designing and building an NFV-based network. It is very important to use quantitative evaluation results in identifying practical network design guidelines, even when rough design policy can be presumed in advance. For example, whether the total network cost can be reduced by 2% or by 20% can significantly affect the business decision [6].

A form in which CDN functions such as video streaming is virtually provided as vCDN (virtual CDN) has been studied in [9]-[15]. The vCDN is virtually realized as a CDN controller, a CDN cache node or a cache server. In vCDN for each content provider, it is being examined mainly in which area the virtual machine (VM) realizing the cache server is optimally deployed. Also, it is assumed that the content cache is placed in the virtual cache server in advance.

This paper explores the approach to reduce the network cost by maximizing the features of NFV, considering the optimal placement of caches not on a VM like vCDN but a content (or group of multiple contents) basis. In other words, we consider the possibility of placing the content cache more economically in a place closer to the user. Moreover, in order to allow the flexible and economical allocation, the content is dynamically cached at the first time of content delivery like ICN (Information-Centric Networking) [16],[17] without placing the cache preliminarily like vCDN. When it is cached, it is stored in a virtual cache server shared with other services and other venders beforehand. Even if the content cache is deployed economically, performance will degrade if the latency between the content cache and user terminals is long. In order to prevent such a case, this paper also proposes a cost-effective placement algorithm for the NFV-based WAN accelerator function [18]-[22].

The rest of this paper is organized as follows. Section 2 introduces the basic concept of virtual cache placement and proposes a virtual cache placement algorithm that minimizes the network cost. Section 3 evaluates the proposed cache placement algorithm and shows the effectiveness of the proposed algorithm through simulation evaluations. The effectiveness of virtual cache placement against DDoS [23] attacks is also evaluated. Section 4 proposes a placement algorithm of the NFV-based virtual WAN accelerator, which prevents the degradation in performance when the latency between the content cache and user terminals is long. Section 5 provides the conclusions. This paper is an extension of the study in Reference [23].

2. Virtual Cache Placement Algorithm for Content Delivery Services

2.1 Basic Concept Of Virtual Cache Placement

1) To simplify the processing, content caching is judged for each content (or group of multiple contents) not on a virtual machine (VM) like vCDN.

2) The virtual cache placement is carried out on a per-area basis, without considering a plurality of areas together. An area is a unit for allocating a cache and corresponds to a city or edge-region in vCDN.

3) It is assumed that a virtual cache server shared with other services and other vendors is deployed in each area and a content cache (replica) is stored in the virtual cache server.

4) The content caching is predetermined based on demand forecast including content popularity etc., and the content is dynamically cached at the first time of content delivery like ICN (Information-Centric Networking) [16],[17] without placing the cache preliminarily like vCDN.

5) The dynamic adaptive streaming that changes distribution speed according to the situation is not assumed in this paper.

Figure 1 shows the main differences between the proposed approach and the vCDN approach, and an example of a content delivery image, based on the above description. Arrows indicate each content delivery. Each content cache is placed in each area in the proposed approach although virtual machine (VM) in which multiple contents are cached preliminarily in CDN.

If it is cached in advance like vCDN, it will be wasted if the prediction goes wrong. For this reason, it is not arranged in advance but is cached at the time of first content distribution, and subsequent distribution is performed from the cache. In figure 1, the content requested for the first time from a terminal in Area 1 (first delivery request) is delivered from the original server in Area 3. The content is cached in the virtual cache server that is connected to a router on the transmission route within Area 1. When another terminal in Area 1 requests the same content (Second and subsequent delivery requests), the content is delivered from the virtual cache server in Area 1.

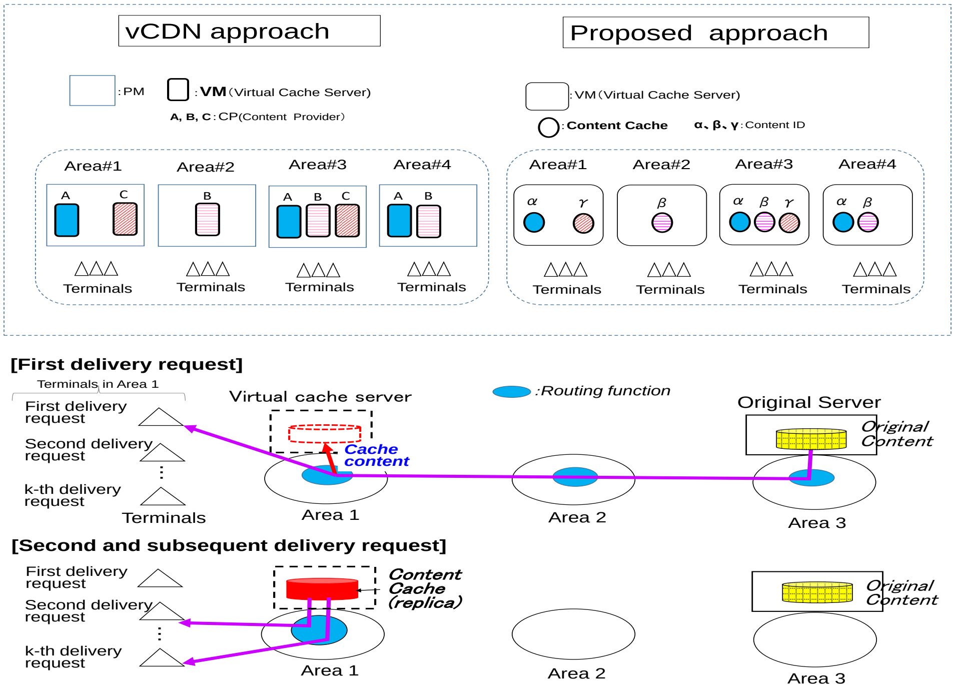


Figure 1. Main difference between the proposed approach and vCDN approach, and content delivery image of the proposed approach

2.2 Proposed Content Cache Placement Algorithm

The network costs are calculated in the following three Modes based on the demand forecast, and the mode with the lowest cost is adopted:

- Mode 1: Content is cached in the area and the requested content is delivered from the content cache in the area.

- Mode 2: No virtual content cache is placed in the area, and
  Mode 2-1: The requested content is delivered from the original server
  Mode 2-2: The requested content is delivered from the content cache stored in another area

Figure 2 shows a network cost comparison between Mode 1 and Mode 2-1. The original server is in Area 3. A terminal in Area 1 requests a content delivery. As explained in Section 2.1, the network cost for the first delivery of content is the same for both modes. Therefore, Figure 2 shows the comparison of costs for only the subsequent delivery request. If the following equation is satisfied, Mode 1 (placing content cache in Area 1) will be more cost-effective.

$$(n_i-1)* \{2*M*(\alpha+\beta) > S*H*\delta \qquad (1)$$

where n_i is the number of terminals requesting the same content in Area i. α is the routing function cost per packet, β the circuit bandwidth cost per packet per 100 km. δ is the cache cost per MB per second. M is the number of packets required for the delivery of content. S is the memory size[MB] of content cache and H is the cache holding time. The circuit bandwidth cost of the virtual cache server is not included in the above equation (1) as it is also required for the original server.

There are the following three methods to determine the order of the areas to be subjected to the cache placement judgment:

-Method a: Select an area with the longest circuit distance from the original server.
-Method b: Select an area with the shortest circuit distance from the original server.
-Method c: Select an area with the maximum n_i

Based on the existing evaluation results in References [4]-[6], it is proposed to implement as follows:

<When n_i is uniform in all areas> Since the possibility that a farther area from the original server is more likely to place content cache is high, Method a is adopted.

<When $n_i$is non-uniform in all areas> Method c is adopted, as there is a high possibility of placing content cache for areas with more terminals requesting the same content.

2.3 Management of Content Cache Judgement in The First Delivery Request

There are the following three methods to judge content cache in the first delivery request:

-Method x: A server in each area manages a table that indicates the placement of a cache for each content.

-Method y: The original server manages a table that indicates the placement of a cache for each content in each area.

Since it is complicated to manage such a table in each area for Method x, this paper adopts Method y. When caching is advantageous for the content requested, the original server instructs the virtual cache server in the relevant area to store the requested content before it delivers the content. It starts delivering the content after it has received confirmation from the corresponding virtual cache server. Then, the virtual cache server in the relevant area caches the content being delivered.

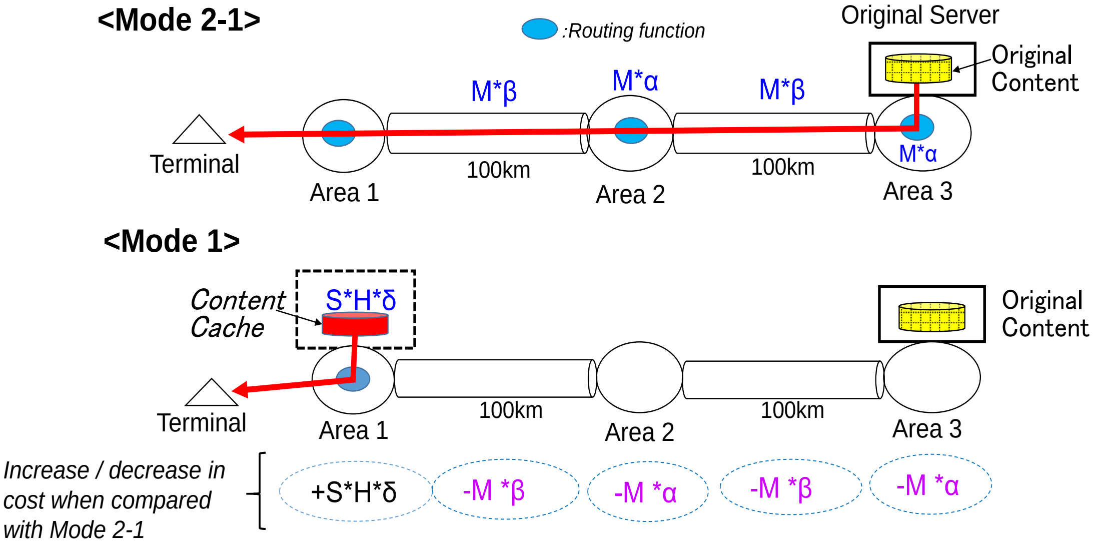


Figure 2. Cost comparative example of Mode 1 and Mode 2-1

3. Evaluation of the Proposed Virtual Cache Placement Algorithm

3.1 Evaluation Conditions

1) Simulation program

If the number of areas is 10, the number of function placement forms is 2^{10} and there are multiple routes for each communication flow per each placement form. We have developed a simulation program in the C language that executes the proposed algorithm proposed in Section 2 and adopts a heuristic approach.

The virtual cache placement is performed as follows with the simulation program:

Step 1: Select an area according to Method a or Method c proposed in Section 2. Move to step2.

Step2: Compare the costs of Mode 1 and Mode 2 proposed in Section 2, and place a virtual cache in the corresponding area if the former is small. Return to step 1 until the judgment of virtual cache placement for all areas is completed.

2) Network configuration

The ladder-shaped model illustrated in Figure 3 which simulates the shape of Japan, is used as in References [4]-[6].

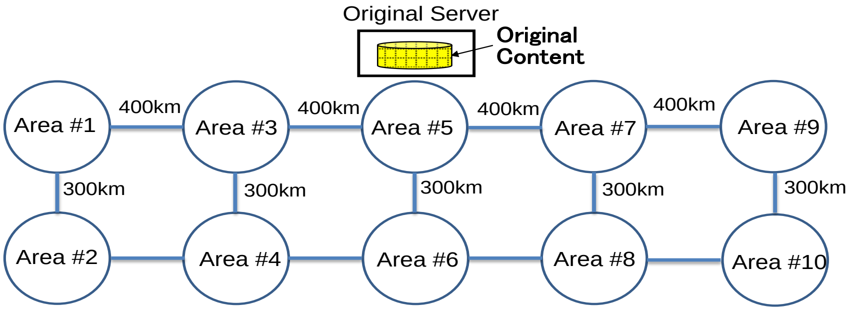


Figure 3. Ladder-shaped network for evaluation which simulates the shape of Japan

3) Request generation model

Figure 4 illustrates a generation model of content delivery requests. The horizontal axis indicates time and the vertical axis video transmission speed, V. In this model, terminals in Area i request the same content in sequence. P is video viewing time, and Hi the time from the start of viewing on the i-th terminal to the start of viewing on the i+1-th terminal. It is assumed here that V and P are constant here.

4) Network cost parameters

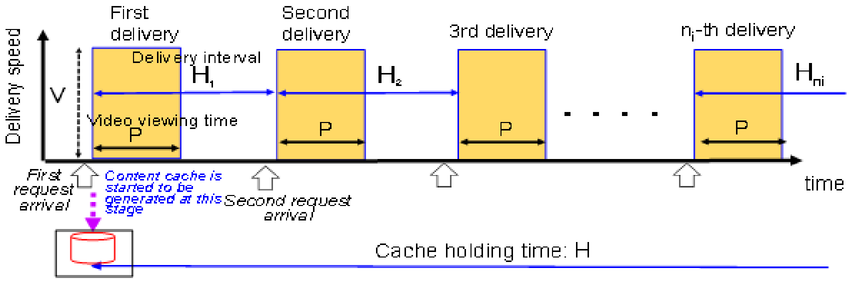


Figure 4. Request generation model

Two cost ratios, Z_C, which is a ratio of the routing function cost to the circuit bandwidth cost, and Y, which is a ratio of the content cache cost to the circuit bandwidth cost, significantly affect content cache placement, are defined as:

$$Z_C = \alpha/\beta \tag{2}$$

$$Y = \delta/\beta \tag{3}$$

The fixed costs for installing content cache, routing functions, and circuit bandwidths are not considered.

5) Effect of content caching against DDoS attacks

Caching in a CDN has an effect of dispersing DDoS attacks [22]. As with the virtual firewall function proposed in Reference [5], restricting invalid traffic near the transmission source can prevent wasteful use of network resources. The cost reduction is roughly calculated in such a way that the virtual cache placement reduces the circuit bandwidth cost and relay processing cost by a factor of Cr (restriction coefficient; its value is equal to or greater than 1.0) although it is not a detailed evaluation.

3.2 Evaluation Results and Discussions

Evaluation results are shown in Figures 5 to 9. Except for Figure 6, a uniform model in which n_i is the same in all areas is supposed. Figure 6 supposes a non-uniform model in which $n_i$in Areas 1 and 9 are R times (peak coefficient) greater than those in other areas. The vertical axis in Figures 5 to 9 shows a normalized network cost which is the actual network cost divided by the network cost in the case where requested contents are not cached in any area and delivered all from the original server.

Figure 5 shows the impacts of n_i and Y on the network cost. It shows the cost breakdown for a case where n_i is 20 and indicates how the value of Y changes the locations of the content cache. Figures 5 also shows the network cost of ‘minimum cost solution’. All possible content cache placement patterns are checked and the pattern with the minimum cost is selected as ‘minimum cost solution’. Figure 6 shows the impact of n_i for the non-uniform model.Figure7 shows the comparison between vCDN and the proposed method and how the value of n_i changes the locations of the content cache. Figures 8 and 9 respectively show the effect of the video viewing time, P, and the impact of the restriction coefficient, Cr, on the network cost. Figure 9 also shows how the value of Cr changes the locations of cache functions.

The following points are clear from these Figures:

1) Placing content caches appropriately can reduce network cost by more than 30%, compared to delivering all from the original server. The larger the number of terminals requesting the same content in Area i, $n_{i,}$ and the lower the cache cost relative to the circuit bandwidth cost (i.e., the smaller the value of Y), the smaller the network cost.

<Reason> In a case where Y = 6 in Figure 5, for example, the placement of content cache reduces the network cost by about 30%. The video delivery to the first requesting terminal requires access to the original server and thus incurs network cost, but this cost is shared by the subsequent (n_i-1) terminals. So, the larger n_i is, the lower the normalized network cost. In a case where Y=10 in Figure 5, the content caches are placed when n_i is more than 20 and the network cost is reduced by about 20%.

The lower the value of Y, the lower the relative cache cost, and the more advantageous it is to place content caches. Placing a virtual cache function can reduce the cost of the circuit bandwidth to the original server and the routing function cost.

2) The result 1) above is effective even if n_i is uniform or non-uniform in all areas.

<Reason> As shown in Figure 6, the larger the peak coefficient, R, the more areas with a large number of terminals place a content cache. These content caches are used when contents are delivered to terminals in adjacent areas that have a smaller number of terminals.

3) The proposed algorithm could reduce network cost by as much as 95% of the cost reduction of ‘minimum cost solution’. As a result, the effective cache placement could be achieved even with an algorithm that easily determines whether or not the content can be cached for each content and each area. Note that the relative network cost does not vary even if the video transmission speed, V, changes.

4) The longer the video viewing time, P, the higher the cache costs and thus the network cost. The longer P has the same effect as the larger Y. Note that the relative network cost does not vary even if the video transmission speed, V, changes.

5) The proposed algorithm could reduce network cost by about 15% compared with vCD. This is because content caches could be placed in a location closer to the user more precisely than vCDN.

6) Impact of network topology on the proposed virtual cache placement

The above evaluations are based on a ladder-shaped network model with 10 areas (Figure 3). Even if the number of areas is increased to 100, for example, or a star-shaped or a loop-shaped model is used, the proposed cache placement algorithm can be applied basically. Therefore, the main trends discussed above and the proposed network design guidelines can be also applicable to such cases. However, additional evaluations are necessary to decide where and how many virtual caches should be placed.

7) The greater the DDoS countermeasure effect (i.e., the greater the restriction coefficient, Cr), the lower the network cost.

<Reason> This is due to an increased reduction in wasteful use of bandwidth and routing processing and increases number of areas where content caches are placed. More detailed evaluation is required in the future.

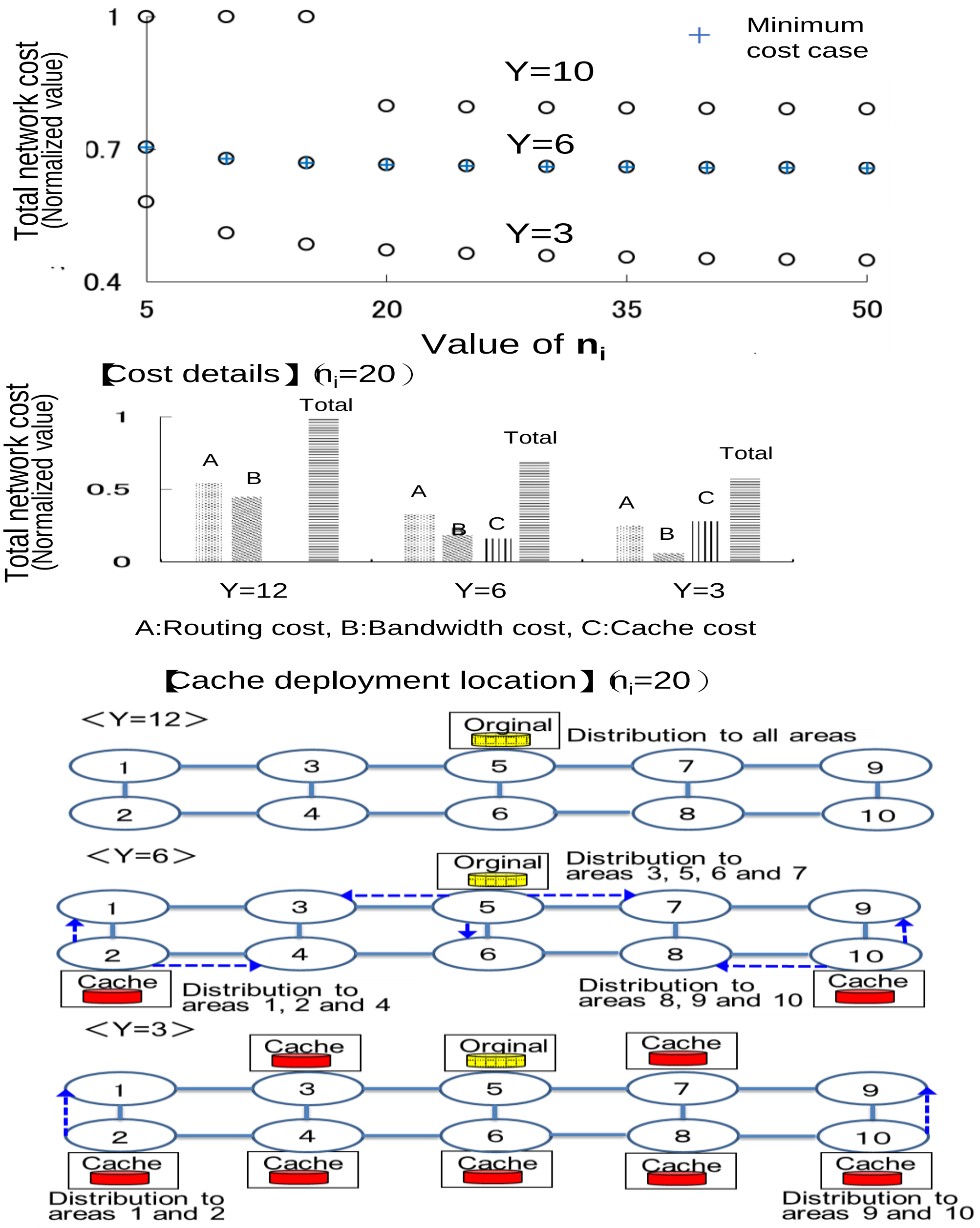
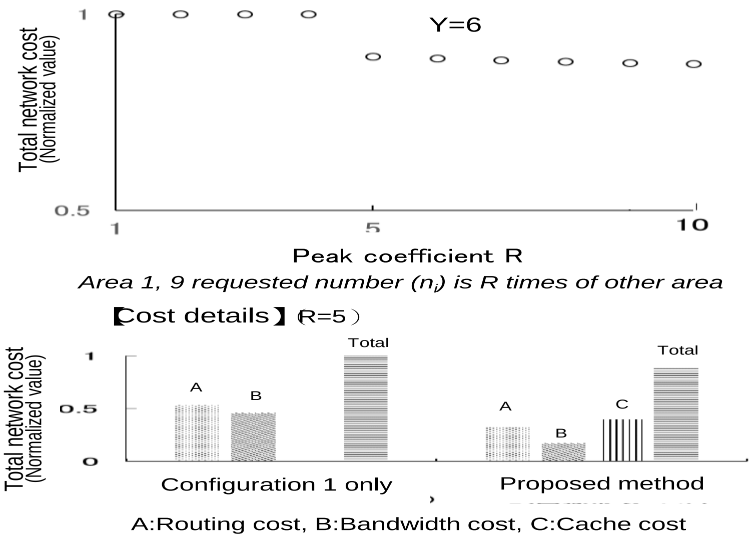


Figure 5. Impact of n_i and Y on total network cost (Uniform model)

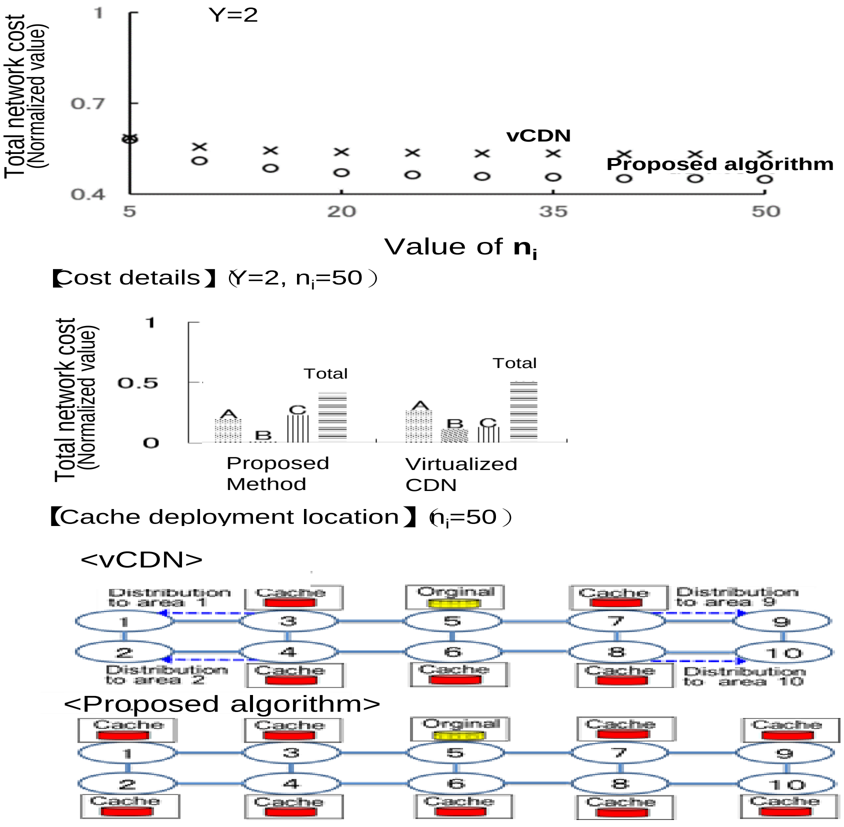


Area 1, 9 requested number (n_i) is R times of other area

【Cost details】(R=5)

A:Routing cost, B:Bandwidth cost, C:Cache cost

Figure 6. Impact of n_i on total network cost
(Non-uniform model)



【Cost details】(Y=2, n_i=50)

【Cache deployment location】(n_i=50)

Figure 7. Comparison between vCDN and proposed algorithm

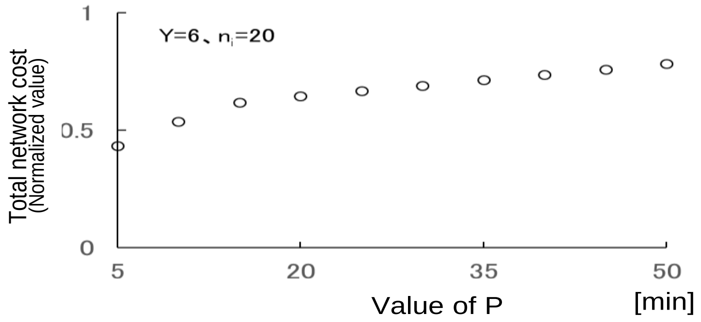


Figure 8. Impact of P on total network cost
(Uniform model)

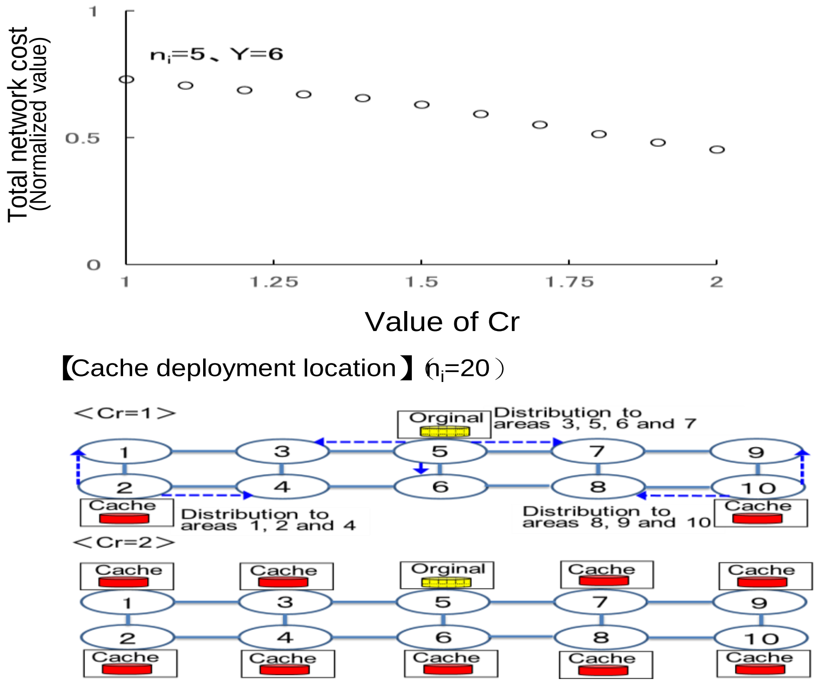


Figure 9. Impact of Cr on total network cost
(Uniform model)

4. Virtual WAN Accelerator Placement Algorithm for Content Delivery Services

Even if the content cache is deployed economically according to the proposed algorithm, performance will degrade if the latency between the content cache and user terminals is long. In order to prevent such a case, this section proposes a cost-effective placement algorithm for the NFV-based WAN accelerator function.

The WAN-accelerator establishes three TCP connections between the user and the server with content cache [20].

-The client-side WAN-accelerator completes the TCP connection setup with the client as if it were the server.

-The WAN-accelerators complete the TCP connection between each other. TCP window size of this connection will be set large, to cope with the long latency in WAN.

-The server-side WAN-accelerator completes the TCP connection with the server as if it were the client.

After the three TCP connections are established, the data sent between the client and server for this specific connection is optimized and carried on its TCP connection between the WAN-accelerator

<Proposed placement algorithm of WAN accelerator function>

The minimum number of relay areas that cannot achieve the required throughput without deploying the WAN accelerator function is evaluated as N. For example, the use of the WAN accelerator function is essential if the content cache is not deployed except in the area when N = 0 and between adjacent areas when N = 1.

The system cost of placing NFV-based WAN accelerator function to each node is compared with that of additionally placing content cache near the terminal to reduce the latency, and if the former is cheaper, WAN accelerator functions will be deployed. This judgment is performed per area with mathematical formulas. The placement configurations are shown in Figure 10 when the network configuration used in the evaluation of Section 2 (Figure 3) is assumed. There is no need to place the WAN acceleration function when N is 3 or more in this network configuration.

It can be assumed that the form of deploying the WAN accelerator function is advantageous in cost under the following conditions:

- The cost of the WAN accelerator function is relatively low.

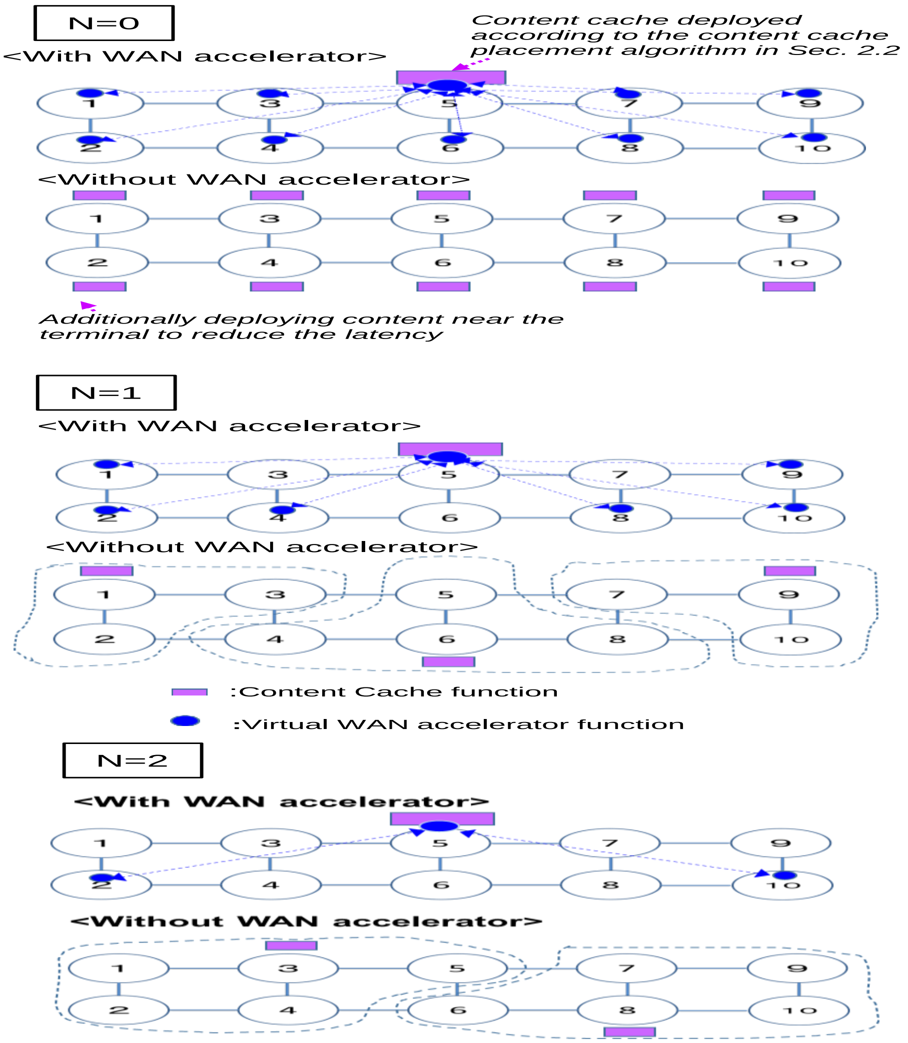

N=0
<With WAN accelerator>
Content cache deployed according to the content cache placement algorithm in Sec. 2.2
1
2
3
4
5
6
7
8
9
10
<Without WAN accelerator>
Additionally deploying content near the terminal to reduce the latency
N=1
<With WAN accelerator>
<Without WAN accelerator>
:Content Cache function
:Virtual WAN accelerator function
N=2
<With WAN accelerator>
<Without WAN accelerator>

Figure 10. Evaluation model for deployment of virtual WAN accelerator

- The Server fixed cost is relatively high.

- The number of terminals n_i in the area is small.

- N is small.

5. CONCLUSIONS

This paper has proposed the virtual cache placement algorithm which judges the optimal placement of the cache not on a virtual machine (VM) like vCDN, but a content (or group of multiple contents) basis. Moreover, the content is dynamically cached at the time of first content delivery without placing the cache preliminary like vCDN. The effectiveness of the proposed algorithm has been identified, based on the simulation evaluation results. Next, this paper has proposed a cost-effective placement algorithm of NFV-based virtual WAN accelerator, which prevents the degradation in performance when the latency between the content cache and user terminals is long exceeded. The effectiveness of virtual cache placement against DDoS attacks has been also evaluated.

It will be necessary to evaluate the algorithm with the original server placed in different areas, network models of different topologies, and different demand generation models.

ACKNOWLEDGMENT

We would like to thank Mr. Kakeru SHIBUYA for his help with the evaluation.

REFERENCES

[1] M. Chiosi et al., “Network Functions Virtualization - An Introduction, Benefits, Enablers, Challenges and Call for Action,” ETSI NFV, Oct. 2012.

[2] “Network Functions Virtualisation (NFV); Architectual Framework,” ETSI GS NFV 002 v1.2.1, Dec. 2014.file:///C:/Users/Kuribayashi/Downloads/gs_NFV002v010p.pdf

[3] R.Mijumbi, J.Serrat, J.Gorricho, N.Bouten, F.D.Trurck and R.Boutaba, “Network Function Virtualization: State-of-the-art and Research Challenges,” IEEE Communications Surveys & Tutorials, Vol. 18, Issue 1, pp.236-262, 2016.

[4] K.Hida and S.Kuribayashi, “Virtual Routing Function Allocation Method for Minimizing Total Network Power Consumption,” 18th International Conference on Information and Communication Systems (ICICS Venice 2016), Aug. 2016.

[5] K.Hida and S.Kuribayashi, “Joint Placement of Virtual Routing Function and Virtual Firewall Function in NFV-Based Network with Minimum Network,” Advances in Network-Based Information Systems-2019, pp.333-345, Aug. 2018, Springer.

[6] S.Kuribayashi, “Virtual Routing Function Placement in NFV-based Networks Under Network Delay Constraints”, International Journal of Computer Networks & Communications (IJCNC) Vol.10, No.1, pp.35-44 , Jan. 2018.

[7] A.Fischer, J.F.Botero, M.T.Beck, H.Meer, and X.Hesselbach,” Virtual Network Embedding: A Survey”, IEEE Communications Surveys & Tutorials, Vol. 15, No.4, 2013.

[8] X.Wei, S.Hu, H.Li, F.Yang and Y.Jin, “A Survey on Virtual Network Embedding in Cloud Computing Centers”, The Open Automation and Control Systems Journal, Vol. 6, pp.414-425, 2014.

[9] “Network functions virtualization (NFV); Use cases (Use case #8 vCDN”, ETSI GS NFV 001 v1.1.1, Oct. 2013.

[10] D.King, M.Broadbent and D.Hutchison, "Evolution of OpenCache: an OpenSource Virtual Content DistributionNetwork (vCDN) Platform," https://www.cambridgewireless.co.uk/media/uploads/resources/Virtual%20Networks%20Group/07.05.15/VirtualNetworks-07.05.15-Uni_Lancaster-Daniel_King.pdf

[11] "The Case for a Virtualized CDN (vCDN) for Delivering Operator OTT Video," Akamai https://www.akamai.com/kr/ko/multimedia/documents/white-paper/the-case-for-a-virtualized-cdn-vcdn-for-delivering-operator-ott-video.pdf

[12] Amazon CloudFront,https://aws.amazon.com/cloudfront/

[13] H.I.Khedher, E.A.Elrahman, A.E.Kamal, and H.AFIFI, "OPAC: An optimal placement algorithm for virtual CDN, " Computer Networks 120, pp.12-27, 2017.

[14] S.Hasan, S.Gorinsky, C.Dovrolis, and R.Sitaraman, "Trad-offs in Optimizing the Cache Placements of CDNs," IEEE INFOCOM'14, pp.460-468, 2014.

[15] P.Marchetta, J.Llorca, A.M.Tulino, and A.Pescape, "MC3:A Cloud Caching Strategy for Next Generation Virtual Content Distribution Networks," IFIP Networking 2016.

[16] G. Xylomenos et al., "A Survey of Information-Centric Networking Research," IEEE Communications Surveys & Tutorials, Vo.16, Issue 2, pp.1024-1049, 2014.

[17] A. Ioannou and S. Weber, "A taxonomy of caching approaches in information-centric network architectures," Technical report, School of Computer Science and Statistics, Trinity College Dublin, Jan. 2015.

[18] Y. Zhang, N. Ansari, M. Wu and H. Yu, "On Wide Area Network Optimization", IEEE Communications Surveys & Tutorials, Vol.14, No.4, Fourth Quarter 2012.

[19] J. Lee, P. Sharma, J. Tourrilhes, R. McGeer, J. Brassil and A. Bavier, "Network Integrated Transparent TCP Accelerator", AINA2010.

[20] Riverbed, "Optimization for the Public Cloud" http://www.riverbed.com/us/products/cloud_products/cloud_steelhead.php

[21] S. Kuribayashi,"Improving Quality of Service and Reducing Power Consumption with WAN accelerator in Cloud Computing Environments", International journal of Computer Networks & Communications (IJCNC), Vol.5, No.1, pp.41-52, Jan. 2013.

[22] Akamai DDoS CDN https://www.akamai.com/us/en/resources/ddos-cdn.jsp

[23] S. Kuribayashi,"Allocation of Virtual Cache & Virtual WAN Accelerator Functions for Cost-Effective Content Delivery Services", The 27th International Conference on Information, Communication and Automation Technologies(ICAT 2019), Oct. 2019.

AUTHOR

Shin-ichi Kuribayashi received the B.E., M.E., and D.E. degrees from Tohoku University, Japan, in 1978, 1980, and 1988 respectively. He joined NTT Electrical Communications Labs in 1980. He has been engaged in the design and development of DDX and ISDN packet switching, ATM, PHS, and IMT 2000 and IP-VPN systems. He researched distributed communication systems at Stanford University from December 1988 through December 1989. He participated in international standardization on ATM signaling and IMT2000 signaling protocols at ITU-T SG11 from 1990 through 2000. Since April 2004, he has been a Professor in the Department of Computer and Information Science, Faculty of Science and Technology, Seikei University. His research interests include resource management for NFV and SDN-based networks, QoS control, traffic control for cloud computing environments, IoT traffic management and green network. He is a member of IEEE and IEICE.