一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NS2023-221(2024-03)

CDNのキャッシュを対象としたDDoS攻撃とCPAの評価と分析

劉　　甲奇[†]　　上山　憲昭[†]

[†] 立命館大学 情報理工学部
〒525–8577 滋賀県草津市野路東 1–1–1
E-mail: [†]is0705vf@ed.ritsumei.ac.jp, [††]kamiaki@fc.ritsumei.ac.jp

あらまし　インターネットでのコンテンツ配信のトラフィック量増加しており，インターネット上でコンテンツを効率的に配信する技術として Content Delivery Network (CDN) が，インターネットの基幹技術として利用が拡大している．CDN を効果的運用するには，CDN のキャッシュサーバを対象とした攻撃方法を解明し，それらを効果的に防ぐ必要がある．そこで本稿では，CDN プロバイダがキャッシュサーバへの攻撃を効率的に防御するための前段階として，キャッシュサーバを対象とした攻撃が効果的となる攻撃戦略を明らかにする．多くのプロバイダは DDoS 攻撃を防ぎサービス品質を向上させるために，複数のキャッシュサーバを配備している．本稿ではキャッシュサーバに対する攻撃として Cache Pollution Attack (CPA) と Distributed Denial of Service (DDoS) 攻撃を想定し，複数のキャッシュサーバの CDN モデルに基づいて，異なるシナリオで CPA と DDoS 攻撃が CDN キャシュに対する攻撃の効果を評価し，攻撃効果に影響を与える要因を分析する．

キーワード　CDN, CPA, DDoS, LRU, キャッシュ

Evaluation and Analysis of Two Types of Attacks CPA and DDoS Targeting CDN caches

Jiaqi LIU[†] and Noriaki KAMIYAMA[†]

[†] College of Information Science and Engineering, Ritsumeikan University
1–1–1, Nojihigashi, Kusatsu, Shiga 525–8577
E-mail: [†]is0705vf@ed.ritsumei.ac.jp, [††]kamiaki@fc.ritsumei.ac.jp

Abstract　The amount of Content Delivery traffic on the Internet has been increasing. The use of Internet technology is expanding. In order to make Content Delivery Network (CDN) effective, it is necessary to clarify the attack method targeting CDN cache servers and prevent them effectively. Therefore, in this paper, we evaluate and investigate the influence of attack pattern against CDN cache servers on the effect of attacks. Many providers deploy multiple cache servers to prevent DDoS attacks and provide better services. In the paper, we investigate the impact of the two types of attacks, the Distributed Denial of Service (DDoS) attack and the Cache Pollution Attack (CPA), against the CDN cache servers. With multiple cache server's CDN model, we evaluate the effect of the CPA targeting CDN caches under different attacking scenarios, and we analyze the factors that affect the attack effect. Finally, we find several factors that affect the effect of the attack, and CDN provider can focus on these factors to prevent the threat of cyber attacks on CDNs.

Key words　CDN, CPA, DDoS, LRU, Cache

1. Introduction

A content delivery network (CDN) consists of geographically distributed servers to cache and efficiently deliver Internet contents, such as HTML pages, images, and videos. CDNs are extremely popular and serve the majority of Web traffic. CDN market value is expected to rise from $11.76 billion in 2019 to $ 49.61 billion in 2025 [1]. CDN also faces the threat of cyber attacks, such as Distributed Denial of Service (DDoS), which affect CDN services and user experience. There is also another type of attack that specifically targets caches, Cache Pollution Attack (CPA), which affect

CDN cache performance and cause high latency. Since the response time of accessing webpages affects user experience, conversion rates, and search engine rankings [4], the response time of content is very important for providers. So, it is necessary to be well protected against these cyberattacks.

Although there are many existing works which investigate the impact of each of these two types of attacks [2] [3], evaluating DDoS and CPA against cache server (CS) in the same way has not been investigated. To effectively allocate resources to defend against these attacks. In this paper, considering that many CDNs consist of multiple CSes with single or multiple layers [9], we build a CDN model with multiple CSes in two layers and evaluate the performance under the DDoS attack and the CPA. We use the queuing model to calculate the load of each CS and link in the multi-layer CSes to derive the total load.

In Section 2., we propose a method to compute the response time of CS. In Section 3., we build the multi-layer CSes model. In Section 4., we evaluate and analyze DDoS and CPA in different scenarios and find the factors that affect the effect of the attack. Finally, we conclude this manuscript in Section 5..

2. Analytical Model

The impact of both DDoS attack and CPA can be measured by the increase of the response time of CSes because the cache miss results in increasing the response time. So we choose response time as a measure of CDN performance.

We use the M/M/1 queue model to derive the average response time of CS. We assume that server requests follow zipf's law, and let M denote the number of contents provided by CDN. Let λ_i denote the Poisson arrival rate of request for content i, and we define $1/\mu$ as the mean of the exponentially distributed service time of CS. Using the M/M/1 queue model, we can obtain$W$, the average service time, by

$$W = \frac{1}{\mu - \sum_{i=1}^{M} \lambda_i}. \quad (1)$$

However, in a network with CSes, there are two situation based on whether the requested content is cached or not, as shown in Fig. 1. The origin server stores all the provided contents, and the CS stores a part of contents. When the content requested by a user does not exist in the CS, which is called *cache miss*, the CS will obtain the content from the origin server, store the content according to the cache replacement policy, and send the content to the user. On the other hand, when the content requested by the user exist in the CS, which is called *cache hit*, the CS will update the cache storage according to the cache replacement policy and deliver the content to the user. Since the LRU (least recently used) is a common cache replacement policy in CDN [6], this paper assumes that all CSes adopt LRU.

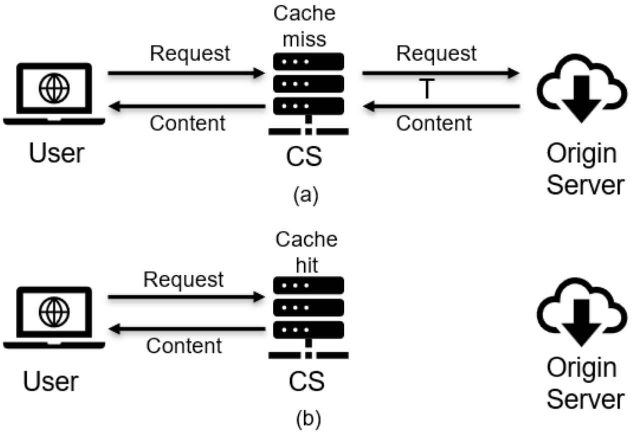


Fig. 1: Workflow of CS

Since the latency between users and CDN CSes depends on the networks between them, and it will not be affected by attacks against CSes, we do not consider the latency between users and CSes. However, we consider the latency between the CSes and the origin server because this will affect the influence of attack strategy on the effect of attack. We define T as the latency between a CS and the origin server which is the latency between the time instance that the CS sends a request to the origin server and the time instance that the CS receives the requested content from the origin server. When the requested content exists in the CS, i.e., cache hit, the average response time is W, whereas the average response time is $W + T$ when the requested content dose not exist in the CS, i.e., cache miss.

Because each content has a different cache hit ratio, let h_i denote the cache hit ratio of content i. We can obtain the average service time of content i, W_i, by

$$W_i = h_i W + (1 - h_i)(W + T). \quad (2)$$

We use the Che-Approximation [7] to predict the hit ratio h_i of each content i on the CS. Let C denote the capacity of the CS, and the maximum number of contents that can be stored in the CS is C. We assume that the request ratio of content i is q_i, and from the Che-Approximation, we can obtain the cache hit ratio of content i, h_i by

$$h_i \approx 1 - e^{-q_i t_c}, \quad (3)$$

where t_c is the characteristic time of the CS, and it is obtained by solving

$$\sum_{i=1}^{M} h_i = C. \quad (4)$$

3. Multilayer CDN Model

Multilayer CDN are designed to provide faster service and to defend against DDoS attacks to some extent, and we focus on this model to evaluate and analyze the CPA and DDoS against CSes [5].

The multilayer CDN model is composed of multiple independent CSes with multiple layers, and we assume CSes of two layers, L1 and L2, as shown in Fig. 2. When a user requests a content, the CS accommodating the requesting user at L1 checks whether the requested content exists or not in its cache storage. If the requested content exists in the cache storage, the CS of L1 sends the requested content to the user. Otherwise, the request is forwarded to the CS of L2 connecting to the CS of L1. If the requested content does not exist in the CS of L2, the origin server which stores all M contents sends the requested content to the user.

We further assume that there are two CSes, α and β, at L2, two CSes, A and B, at layer 1 connecting CS α, and three CSes, C, D, and E, at layer 1 connecting CS β. Let T_1 denote the latency between L1 CSes and L2 CSes, and T_2 denote the latency between L2 CSes and the origin server. Moreover, let W_A, W_α, and W_o denote the average service time of contents at CS A, CS α, and the origin server, respectively.

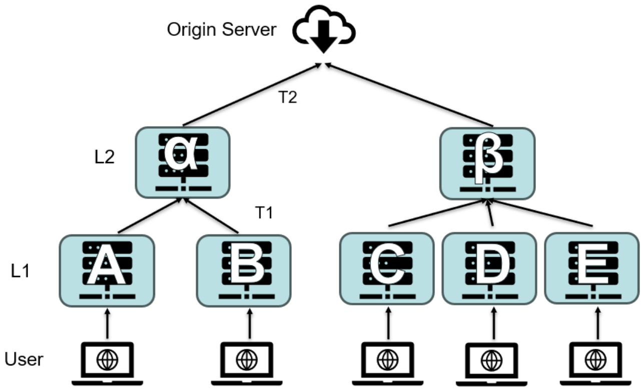


Fig. 2: Multilayer CDN Model

When the requested content exists in the CS A at L1, the average response time, r_A, is W_A. We define r_α as the average response time when the request is cache miss in CS A and forwarded to the CS α at L2. Moreover, we also define r_o as the average response time when the request is forwarded to the origin server. We can obtain r_α and r_o by

$$r_\alpha = W_A + W_\alpha + T_1, \tag{5}$$

$$r_o = W_A + W_\alpha + Wo + T_1 + T_2. \tag{6}$$

The average response time of content i when request arrives at CS A, $R_A(i)$, is obtained by

$$R_A(i) = h_i^A r_A + (1-h_i^A)h_i^\alpha r_\alpha + (1-h_i^A)(1-h_i^\alpha)r_o, \tag{7}$$

where h_i^A and h_i^α is the cache hit ratio of content i, h_i, at CS A and α, respectively. By summing $R_A(i)$ among all i in M, we can obtain the average response time of request when accessing CS A by

$$R_A = \frac{\sum_{i=1}^{M} R_{a_i}}{M}. \tag{8}$$

In the same way, we can also obtain the average response time when accessing each of other CSes.

4. Numerical Evaluation

We perform numerical evaluation through computer simulations and we set the parameters for the simulation based on the convenience of the experiment and the authenticity of the data. We compared the effect of the attack on the average response time in different cases.

The setting parameters of the experiments are shown in Table 1. We assume contents provided by CDN occupy the same amount of space in the cache and have different request rate depending on its popularity. In order to simulate user requests for contents following the zipf's law, we set $\lambda_i = 80, 9, 6, 4, 1$ requests per second at all the five CSes of L1. To keep offered load 50% without attacks, we set the average service time as shown in Table 1. We set the latency based on realistic data [8].

Tab. 1: Simulation parameter settings

Paramater	Value
Content count provided by CDN, M	5
Cache size, C	3
Total requests rate	100 /s
Average service time of L1 CSes	5 ms
Average service time of CS α	5 ms
Average service time of CS β	3.3 ms
Average service time of origin server	3.3 ms
Latency between L1 and L2 CSes, T_1	50 ms
Latency between L2 CSes and origin server, T_2	30 ms

4.1 Attack Definition

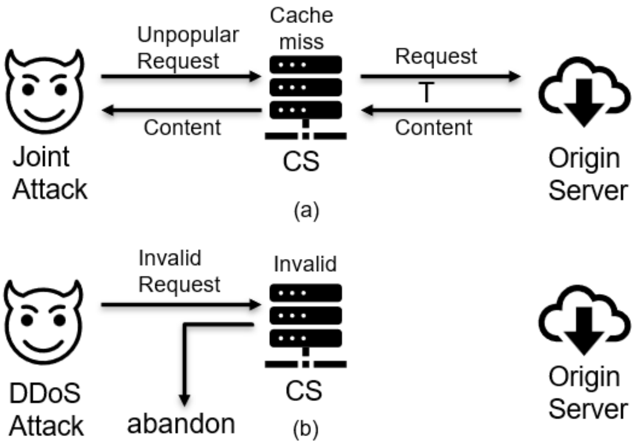


Fig. 3: CPA and DDoS attack

In DDoS attacks, the attackers send many packets from bots to target CSes to increase the processing load of CSes and increase the response time of content delivery. On the other hand, in CPAs, the attackers send request packets for unpopular contents to target CSes to decrease the cache hit ratio of most legitimate users. Because of lower cache hit ratio, most users who request popular contents will have longer response time of content delivery. Therefore, the aim of CPAs is also increasing the response time of content delivery [1], so the purpose of DDoS attack agrees with that of the CPA, and the interest of attackers is how to combine these two types of attacks and select the attacking targets.

In this paper, for simplicity, we define DDoS attack as the cyberattack sending request packets to invalid contents that will increase the processing load of CSes, whereas the request packets are not sent to the CS, i.e., CS at L2, and invalid contents are not stored in the CS. On the other hand, we define CPA as the cyberattack sending request packets to valid contents to decrease the cache hit ratio and increase the processing load.

We define two attack patterns: only DDoS attack and only CPA. In DDoS attack, the attacker sends many requests to invalid contents that will increase the average response time of CSes. The CPA will send many requests to unpopular contents, i.e., content 5, which will spread the effect of the attack throughout the path as shown in Fig. 3.

We define two attack scenarios. The first is that the attacker has limited resources, which means that the attacker has an attack capacity and searches for the optimal attack strategy on the premise that the sum of the resources of the attacks on each server does not exceed the capacity. The other case is that the CSes has a protection mechanism and starts to detect the attack or transfer request when the usage rate reaches the threshold. The attacker aims to maximize the attack without triggering the protection mechanism and compares the effect of CPA and DDoS attack.

4. 2 Attack with Limited Resources

Note that the total attacking capacity of the attacker is 80 requests per second in all the attack patterns. When the attacker send request packets to multiple CSes, it equally sends packets among the CSes. For example, in the AC DDoS attack, the attacker sends request packets to CS A with 40 requests per second, and it sends request packets to CS C with 40 requests per second.

Figure 4 shows the average response time of each CS in each of attack patterns as well as the case without attacks. In the figure, "AC CPA" means the case in which the attacker sends CPA packets to both CS A and CS C, for example. Compared with the case of making DDoS attack only, the CPA largely increased the response time of CSes with the same attacking capacity. Both the DDoS attack and the CPA increased the average response time of the target CSes, whereas the CPA also increased the response time of other L1 CSes connecting to the same L2 CS with the target L1 CS. However, when two or more CSes are attacked at the same time as shown in Fig. 4 (b), the DDoS attack has little effect because the resources are dispersed, whereas the CPA still has significant advantages over the DDoS attack.

Therefore, we confirm that the CPA can improve the effect of the attack compared with the DDoS when the attacker has less attack resources. From the perspective of CDN providers, the threat of CPA is greater than that of DDoS attack, and it is necessary to increase the defense mechanism based on caches.

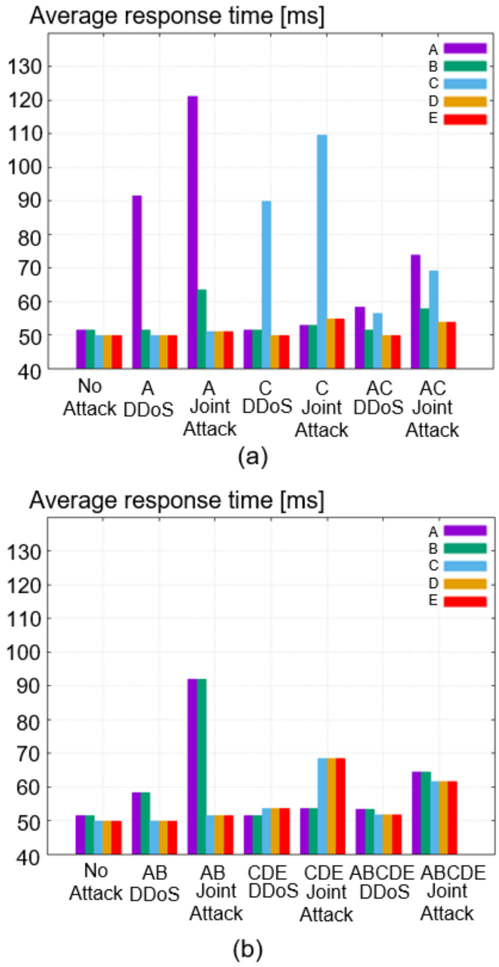


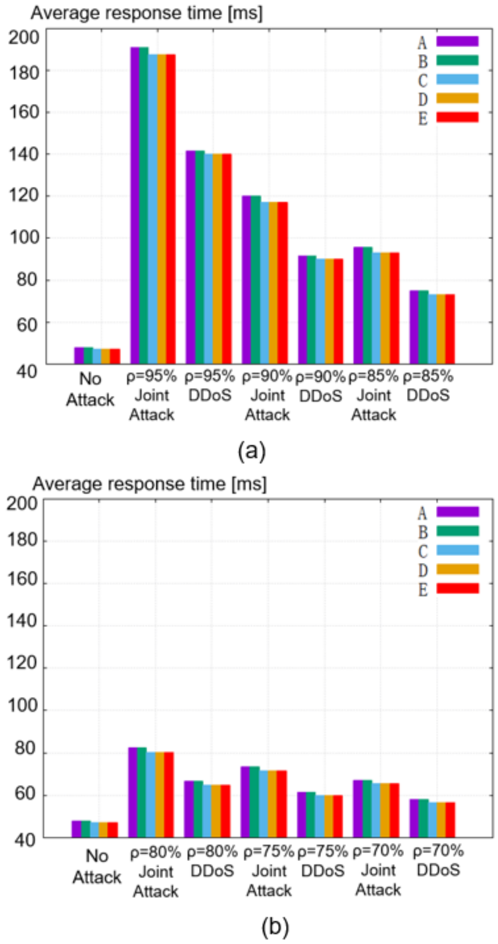
Fig. 4: Average response time of each CS with limited resources

4. 3 Attack under Protection Mechanism

In this section, we assume that CSes have a protection mechanism, and CSes can bound the utilization of processing capacity of CSes below the threshold ρ even when DDoS or CPA occurs. In this case, the attacker will attack as much as possible until the utilization of the attacked CS is close to threshold ρ. Different from the previous case, the attacker will launch an attack on all servers, which will lead to a large increase in the usage of CSes at L2 and the origin server. Considering that when the offered load is greater than 100%, CSes cannot handle the requests, so we set the average service time some CSes as shown in Table 2.

Tab. 2: Average service time of CSes and origin server

CS α	3.3 ms
CS β	2.2 ms
Origin serve	2.2 ms



(a)

(b)

Fig. 5: Average response time of each CS under protection mechanism

Figure 5 shows the average response time of each CS with different threshold ρ as well as the case without attacks. Both the DDoS attack and the CPA largely increased the average response time of the target CSes. Compared with case of DDoS attack, the CPA apparently increased the response time of CSes with the same threshold ρ. When the threshold p becomes low as shown in Fig.4 (b), the effect of both attacks became weak, and the CPA still had apparently advantages over the DDoS attack.

4.4 Factors Affecting Effect of Attacks Targeting CS

We analyze the factors affecting the effect of the attack based on attacking under protection mechanism and find that there are mainly two factors: the latency and the offered load.

Since the tendency of results of all CSes is almost identical under the attack protection mechanism, we focus on the result of CS A. Figure 6 shows the average response time of CS A for various settings of T_1 and T_2. The increase in latency improved the effect of both attacks as shown in Fig. 6. As the latency increased, the average response time increased. Because the CPA can enlarge the effect of the attack over multiple cache layers, it was also more susceptible to latency. As the latency increased, the gap between the CPA and the DDoS attack became larger, indicating that the CPA was sensitive to latency. The CPA was more destructive to the CDN as a delay-sensitive service.

On the other hand, the offered loads of CSes of L2 and origin server are also key factors. Once the load is over 100%, the CS will not be able to handle the request. If the CDN provider only considers the load of L1 CSes, and it ignores L2 CSes and the origin server, L2 CSes and the origin server which are not directly attacked, will be also threatened by the CPA.

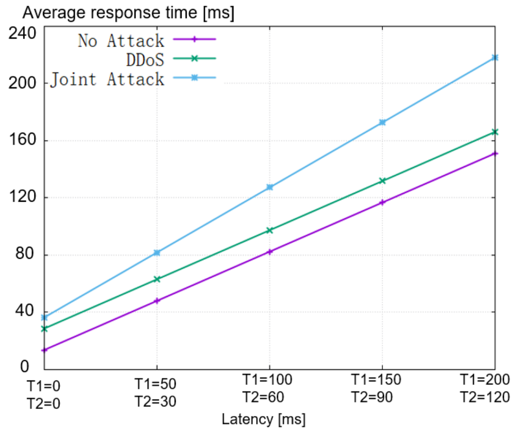


Fig. 6: Average response time of CS A with different latency when $\rho = 80$

In our experiments, we find that the DDoS attack do not have the impact on L2 CSes and the origin server, while in the CPA, the attacker requests will be transferred to L2 CSes or even the origin server. Meanwhile, requests of legitimate users will be also sent to L2 CSes and the origin server due to the decrease of cache hit ratio of the requested content. As a result, L2 CSes and the origin server that would not otherwise be attacked are equally at risk as well as L1 CSes.

When offered Load of origin server without attack is set to 50%, if every CS of L1 is attacked, once the offered load of L1 is over 56%, the origin server will not be able to process the request. We call 56% in this case as the *safe threshold*, and Fig. 7 shows the safe threshold of CS A against the offered load of the origin server without attacks. We find that when the origin server load is less than 39%, the security threshold ρ exceeds 100%, which means that the origin server is not compromised even if the L1 CSes are unable to serve the request. As the load balance of the origin server increases, the security threshold ρ decreases sharply.

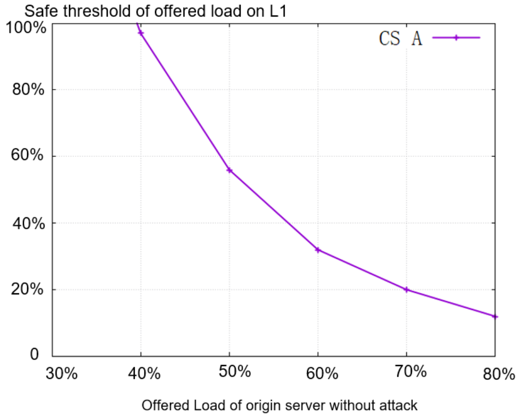


Fig. 7: Maximum allowable offered load on L1 CS A under different offered load of origin server

For CDN providers, they should not only pay attention to the defense of L1 CSes, but also pay attention to the defense of L2 CSes and the origin server. In the case that the attacker has a lot of resources, the requests to L2 CSes and the origin server will be much higher than when the attacker is not attacked. Compared with L1 CSes, L2 CSes and origin server will incur more losses when attacked.

5. Conclusion and Future work

In this paper, we analyzed and evaluated the impact of the two attacks against CDN cache servers, the DDoS attack and the CPA. We used the M/M/1 queue model to derive the response time for CSes in CDN, and we associated the cache hit rate of each content with their request rate through the Che-approximation. We build a multi-layer CS model according to the actual CDNs, and we compared the response time under the CPA and the DDoS attack with that at non-attack. We also set up two different attack scenarios and found that in both scenarios, the CPA was more threatening.

We investigated the factors that affect the effect of the attack, latency and offered load, and we revealed the potential threats in the multi-layer CS model. By analyzing these two factors, we found what CDN providers should pay attention to help them better defend against attacks. The data used in this paper are not close enough to reality. In the future, we plan to make more realistic evaluations to prove our theory with more realistic data. Moreover, we will also use more CDN models, as well as more CDN defense mechanisms, to analyze the effect of attacks. Meanwhile, we will focus on finding the key factors that potentially affect the effectiveness of the attack, such as the cache hit ratio of each content, to further improve the effect of attacks. Then we will propose more efficient ways to defend against attacks.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 21H03436 and 21H03437.

References

[1] M. Ghaznavi, E. Jalalpour, M. A. Salahuddin, R. Boutaba, D. Migault and S. Preda, ”Content Delivery Network Security: A Survey,” IEEE Communications Surveys & Tutorials, vol. 23, no. 4, pp. 2166-2190, Fourth quarter 2021

[2] L. Yao, Z. Fan, J. Deng, X. Fan and G. Wu, ”Detection and Defense of Cache Pollution Attacks Using Clustering in Named Data Networks,” IEEE Transactions on Dependable and Secure Computing, vol. 17, no. 6, pp. 1310-1321, 1 Nov.-Dec. 2020

[3] K. Kim, Y. You, M. Park and K. Lee, ”DDoS Mitigation: Decentralized CDN Using Private Blockchain,” International Conference on Ubiquitous and Future Networks (ICUFN) 2018

[4] Emma Ryan, “Website Load Time Statistics: Why Speed Matters in 2024” , Website Builder Expert, 2021. [Online]. Available: https://www.websitebuilderexpert.com/building-websites/website-speed. [Accessed: Jan- 2024]

[5] Sankalp Basavaraj, “Multi-Cloud, Multi-CDN: The Future of the Internet” , Medium, 2022. [Online]. Available: https://labs.ripe.net/author/sankalp-basavaraj/multi-cloud-multi-cdn-architecture-a-deceptive-future-of-the-internet. [Accessed:Jan- 2024]

[6] Z. Zeng and H. Zhang, ”A Study on Cache Strategy of CDN Stream Media,” 2020 IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC) 2020

[7] H. Che, et al., “Hierarchical Web Caching Systems: Modeling, Design and Experimental Results,” IEEE J. Selected Areas of Commun., vol.20, no.7, Sep. 2002

[8] R. K. Thelagathoti, S. Mastorakis, A. Shah, H. Bedi, and S. Shannigrahi, ”Named Data Networking for Content Delivery Network Workflows,” IEEE International Conference on Cloud Networking (CloudNet) 2020

[9] K. Wang, Y. Wu, J. Chen and H. Yin, ”Reduce Transmission Delay for Caching-Aided Two-Layer Networks,” IEEE International Symposium on Information Theory (ISIT) 2019