A Cost Effective Strategy for Optimizing the Positioning of Replica Servers to Improve Service Delivery in a Cloud-Based Content Delivery Network

Priyanka Dharmapal[1], Channakrishnaraju[2]

[1] *Research Scholar, Department of Computer Science and Engineering, SSAHE University, Tumkuru, Karnataka, India*

[2] *Professor, Department of Computer Science and Engineering, SSIT, Tumkuur, Karnataka, India*

Email: prsh2019@gmail.com

Abstract

One of the most important considerations for a given geographic diversity connected with placement issues in Content Delivery Network (CDN) is the location of Replica Servers. Following a review of the literature, it became clear that research were focused more on traditional CDN placement issues than cloud-based CDN architectures, and the few studies that did focus on replica selection were still in their infancy. Such models are additionally not benchmarked or objectively evaluated to demonstrate their efficacy. Therefore, the suggested study provides a novel computational framework architecture that is linked to a cloud-based CDN and can make replica server management more affordable for better service delivery. The simulated study results, which were implemented utilizing analytical research methods, reveal that the proposed strategy delivers reduced cost, reduced resource dependency, and reduced resource consumption when it is implemented utilizing analytical research approach.

Key words: *Replica server, cost, content delivery network, resource, cloud*

1. INTRODUCTION

There is a need for an efficient networking technology to help in providing the smooth services regardless of traffic condition in a dedicated manner due to the increasing demands for data and service availability. One such solution is a Content Delivery Network (CDN), which uses a surrogate server and content placement to help deliver the content fast and effectively [1] through [5]. Although there are a number of stated benefits to using a CDN system to speed up data delivery, there are also a number of difficult conditions [6]. The main issue with CDN implementation is that it involves a significant financial investment and an extremely complex deployment process [7]. Based on a number of research, it has been determined that CDN enabled technologies currently use cloud computing more [8]. This adoption offers advantageous features for resource allocation, pay-per-use, hosting service supportability, and service provisioning [9]. It should be mentioned that the role of the replica server, which is in charge of copying the original material to the content replica server, is crucial to the overall functioning of the CDN. Despite the several ways being used to improve the content placement algorithm, there is still a significant trade-off between the needs of the user and the replica server's current level of operation. The main problem with replica server placement is related to the optimization problem, which can be either of the constraint or unconstrained kind. In the current system, which considers network topology, network performance metrics, latency, the number of hops, and other factors, there is not much emphasis placed on the deployment cost of the replica server, as well as its associated delivery cost and update cost.

In addition, there are a number of theoretical models that have been given in previous research that examine the location of the facility with both capacitated and incapacitated variants while taking constraints such server capacity into account. Other methods used in previous studies include K-median, minimal K- centre, K-cache location, etc. in order to localize the replica server [10]. A detailed analysis of the current plan reveals that the bulk of methods for placing replica servers are based on traditional CDNs rather than cloud-based CDNs. The primary distinction between these two approaches is that while cloud-based CDN are extremely distributed and still have the potential to be centralized to create a vast network, conventional CDN are somewhat centralized. When a standard CDN is exposed to difficulties with greater traffic conditions as well as the unpredictable behaviour of many services and applications running over it, it will imply that there are numerous unavoidable obstacles to be confronted. In the context of the entire scenario, it should be highlighted that the cloud-based CDN system places less focus on leveraging the virtualized environment to expand the reach of replica servers [11]. Even though it is practically impossible to place the content server in several different locations, the cost can be greatly reduced if the virtualized environment (or machine) that already exists can be divided into local and global forms in order to provide connectivity on both a small- and large-scale. In order to enable better networking in a cloud-based CDN system [12], it is crucial to conduct an inquiry in this direction to understand the impact of employing local virtual machines as well as proxies [13]. Understanding how spatial variables affect the location of the replica server in the cloud-based CDN and their potential effects on large-scale networks is also crucial. Therefore, the suggested study also conducts a deeper analysis of current literatures in order to examine these possibilities. It is discovered that studies are more focused on placing content and less on placing replica servers while installing the virtualized environment [14].

The majority of current studies have focused on the crucial QoS attribute; however, it is also crucial to take into account the user's experience, which is not often reported to be emphasized. ii) Current approaches have deployed a mechanism taking into account different types of files; however, such schemes are not very effective when deployed under a situation of non-uniform distance between content server and replica server, and iii) further, non-utilization of virtual machines Using local virtualized machines and proxies connected to a centralized content server that is further maintained in a distributed fashion, the proposed approach discussed in this research proposes a novel replica server placement. The major goal of this effort is to provide comparable user experience quality and service quality in the face of changing traffic conditions in an economical way.

The discussion of relevant literature is as follows: In our previous review, we explored various approaches and techniques aimed at enhancing the performance of Content Delivery Networks (CDNs) [15]. Currently, there are several ongoing efforts that directly or indirectly contribute to improving CDN performance.

Kalghoum et al. [16] have conducted work focusing on a probability-based model to manage stalling events in streaming services within CDNs. Their model is built upon the concept of limited storage space in caches and efficient allocation of CDN resources to enhance streaming delivery. Additionally, Ghasemi et al. [17] have presented an optimized model for content placement, employing a belief propagation system within a distributed network. This comprehensive model follows a sequential flow, involving a content helper selection mechanism, content caching, and precise content delivery to the intended destinations. Their iterative approach to modeling takes into account the states of communication channels and content popularity, with a particular emphasis on controlling energy consumption in wireless networks. Huang et al. [18] have conducted research on an allocation model for replica servers, claiming it to be both energy-efficient and reliable. Their model facilitates concurrent task processing on the primary server, which is connected to a dedicated storage point.

Cheng et al. [19] have developed a model for data replication with a primary focus on enhancing data availability within a complex, weighted network, taking centrality factors into account. Their model also assesses graph partitioning attributes while storing data replicas in fog devices. In this model, one unit stores fog information and a single data replica, while the other unit stores all file replicas using a greedy-based approach. Goudarzi and Pejman [20] have emphasized the significance of caching mechanisms, particularly in large distributed networks. Their study underscores that there are still unresolved issues related to caching mechanisms that require attention. Fan et al. [22] have highlighted the importance of vertex markers as a means to discuss potential improvements in grid methods. Their work explores possible adjustments to scale within grid-based approaches.

The suggested document is divided into the following sections: Section 1 addresses the introduction, while Section 2 describes the approach being used, Section 3 examines the results being achieved, and Section 4 discusses the paper's conclusion.

2. METHODOLOGY

The main goal of this phase of research is to create a cutting-edge model that can execute resource provisioning to ensure better replica server placement in a dynamic distributed cloud-based CDN system. Utilizing an analytical research technique, the suggested system begins its design by taking into account a multi cloud system employing a communication model, as illustrated in Figure 1.

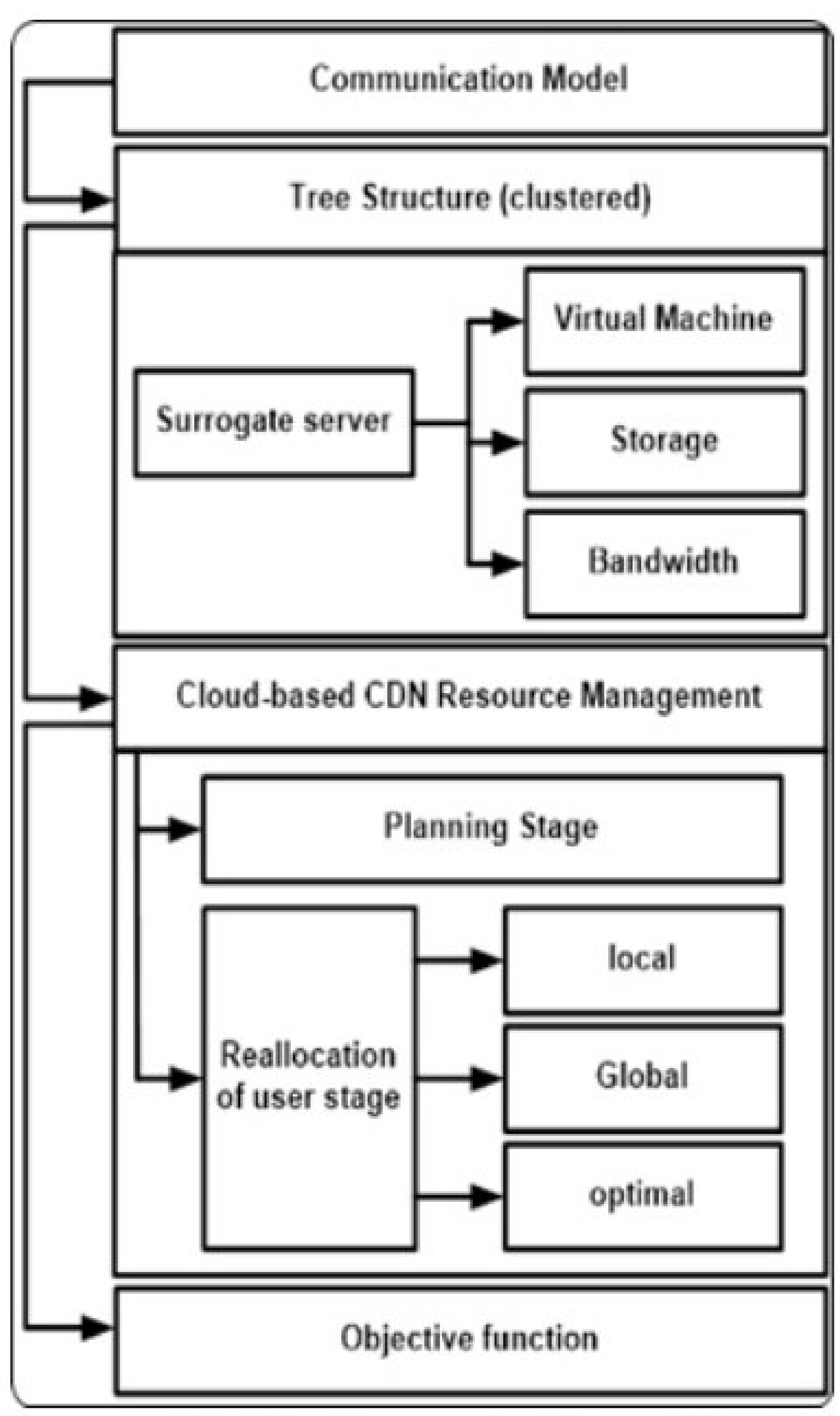

Figure-1: Communication Model

The suggested system uses a highly clustered tree structure to simplify the topology design. According to this tree structure, a dedicated cloud system using surrogate servers that are further integrated with virtual machine, storage, and bandwidth operations serves each user. The creation of a cloud-based CDN resource management system falls under the following phase of implementation, which is further divided into the planning stage and the reallocation of user stage. The cost model will be used by the proposed system to determine how much it would cost to allocate resources under the current situation. The third set of activities are carried out toward resource provisioning in the subsequent stage of user reallocation while taking formulation of realistic restrictions and obtaining maximal optimality in replica server placement into consideration. In order to effectively position replica servers in a cloud-based CDN system, an objective function will be built that can balance the resource provisioning demands with dynamic content delivery system. It is believed that this phase of implementation would result in the cloud-based CDN performing at its best. To evaluate the impact of several network parameters, such as channel capacity, cost, uncertain traffic scenario, etc. on the model, a test environment will be built. The model is expected to display less computational complexity as well. The research approach used for the implementation briefed in regard to system design and algorithm implementation is discussed in the following part.

2.1 SYSTEM DESIGN

In order to provide higher service availability via cloud networks, the main system design of the proposed method is based on the position and closeness of the replica server and the proxy cache. This plan takes into account a Centralized Content Server (CCS), which is situated uniformly close to several local virtual machines (LVM), and which stores data on replica servers and proxy cache.

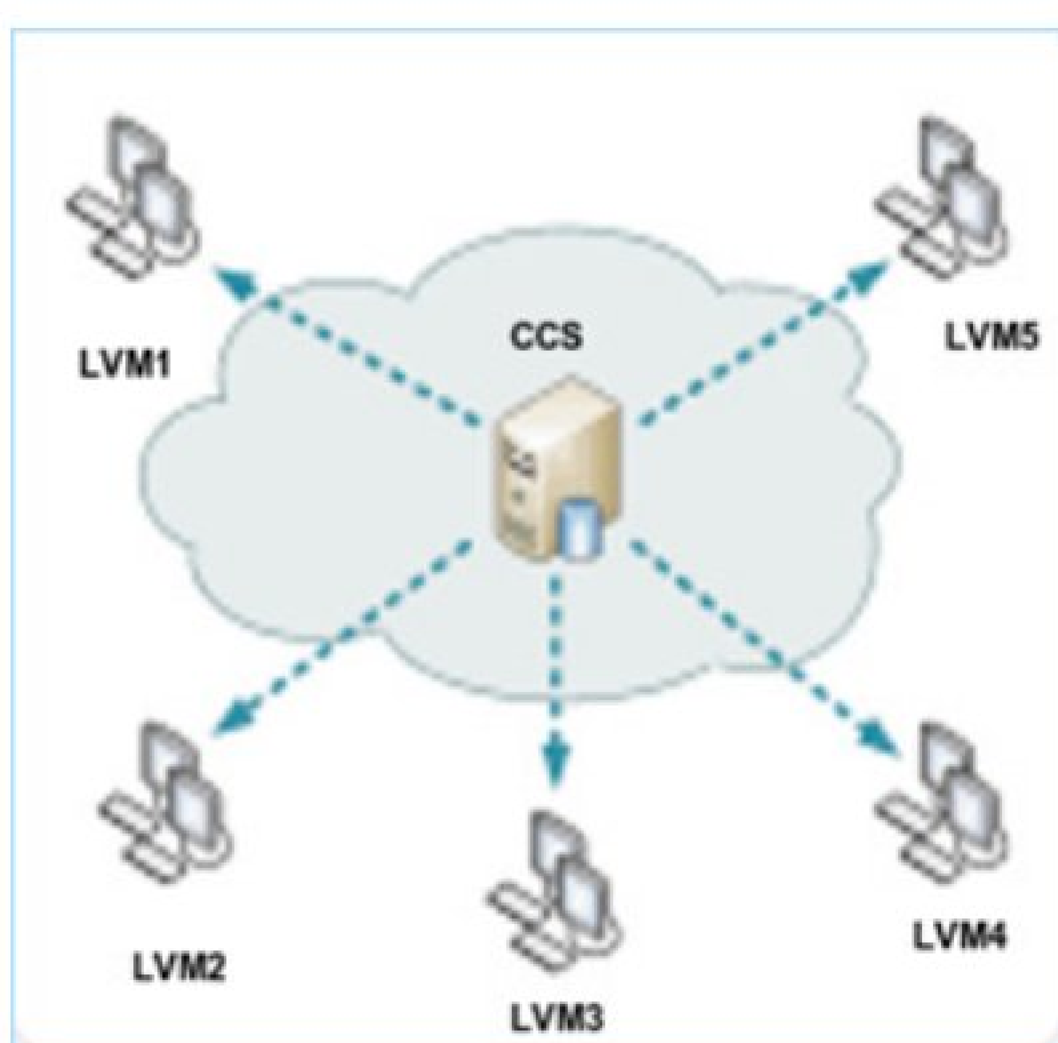

Figure-2: Scheme of Replica Server placement

The placement strategy of the replica server is highlighted in Figure 2 above where ith multiple observation zone Oz exists, i.e., Ozi=[Oz1, Oz2,...Ozi] where the deployment of LVM is done. It will imply that each Oz will contain a particular number of LVM (based on the distribution of Oz). It can be stated mathematically as follows:

$$LVM_i=[P_i, R_i]^{\alpha\beta} \quad (1)$$

Algorithm for Cost-based Replica Server Placement

Input: No, Ne, LVM, i, P
Output: ω
Start
1. Initial network [No, Ne (LVMj), P]
2. G →(No, Ne) P
3. $\mu = \theta$ (G)
4. formulate matrix G=[μ, IL, OL]
5. Extract new tree= Φ (G)
6. Lt=(G, cond(probability))
7. $\omega = \rho(\mu)$

end

In the above algorithm the input of origin node is No,last node is Ne, local virtual machine LVM, number of LVM j, and service given by the content delivery network is P that will give the final result as ω.The initial task of the above algorithm is to initialize the exact number of local virtual machine by taking the placement strategy and its connectivity with origin node and last node(line-1).the next stage of implementation consists of constructing a graphical tree G for connecting all the nodes and links in the form of a network topology in distributed and very large scale(Line-2). With the help of centrally located CCS it constructs a distributed scheme of replica server placement(Line-3).The calculation of the middle location is obtained by the score of each LVM that is divided with respect to available least quantity nodes. The output of this operation gives the score of importance attribute θ the specific function specific function μ(Line-3) After calculation of important attribute the algorithm will generate a two dimensional matrix, incoming link IL, and outgoing link OL (Line-4). This stage helps to build a distributed network links connecting all the LVMs in different places. This assessment is performed for the purpose of assessing the consistency of the new algorithm in order to acquire less delay by taking multiple locations of the LVM links with the replica server in specific Oz. After this task is accomplished, the algorithm deploys another function ω that constructs a diagraph tree structure G over the existing tree (Line-5). The system than generates a local tree Lt by taking the newly generated tree G with the embedded condition (Line-6). When probability prob is statistically significant when the condition cond is formed with Lt .In the last step, an explicit function ρ is formed that performs final estimation of the cost ω which is linked with the placement of surrogate server at the time of transmission of data. (Line-7).

3. RESULTS AND DISCUSSION

In the preceding section, it was observed that the proposed approach introduces an innovative mechanism for placing replica servers to enhance performance within a distributed CDN system hosted in the cloud. Consequently, it is crucial to outline a definitive evaluation strategy to ensure improved data transmission performance in variable test environments within the CDN. The algorithm for the proposed system is implemented in MATLAB, with a focus on achieving a tree-like structure for server deployment, where the cost associated with replica server placement serves as the primary factor for result analysis. Table 1 presents an analysis of the costs associated with individual LVMs situated across various orientations while considering the central location of CCS (Content-Centric Storage). To evaluate the proposed scheme, a bipartite graph is constructed as part of the assessment process.

Table-1: simulation parameters

Simulation Parameter	Values Assigned
Total deployment areas of LVM	9
Total replica server	10
Bandwidth assigned for replica server	5000 mbps
Total proxy (cache)	25
Bandwidth assigned for proxy	3000 mbps

Utilizing the simulation parameters and their respective values presented in Table 1, the proposed system conducts cost estimation for individual LVMs in nine distinct circular positions, with CCS located at its centre. The comprehensive analysis is performed using variable-sized random data packets, each containing 3000 bytes, within a simulation area measuring 1000 x 1000 m^2, and it involves the presence of 500 nodes. The assessment takes into account randomly selected origin and end nodes. The individual cost outcomes for the nine different locations are displayed in Table 2, and this assessment is conducted over 50 iterations.

Table-2: Estimated Cost

Location	LVM No	ω
LVM_2	7	4
LVM_4	6	0
LVM_1	3	28
LVM_3	1	7
$LVM_2 \rightarrow LMV_1$	2	11
$LVM_2 \rightarrow LVM_3$	8	29
$LVM_4 \rightarrow LVM_1$	5	1
$LVM_4 \rightarrow LVM_3$	9	21
CCS	4	25

Figure-3:highlights the outcomes related to the estimated cost of replica server placement, demonstrating an impressive gain of approximately 45.2% compared to existing schemes such as the QoS-aware [20] and consistency-aware algorithms [20], which are commonly employed in replica server placement. The existing QoS-aware approach is known for its predefined computations, where the system determines its network topology to achieve a specified target Quality of Service (QoS) through Load Virtual Machines (LVMs) for processing generated queries. However, this static characteristic of the QoS-aware approach becomes problematic as it encounters increased traffic, leading to higher costs associated with replica server placement allocation.

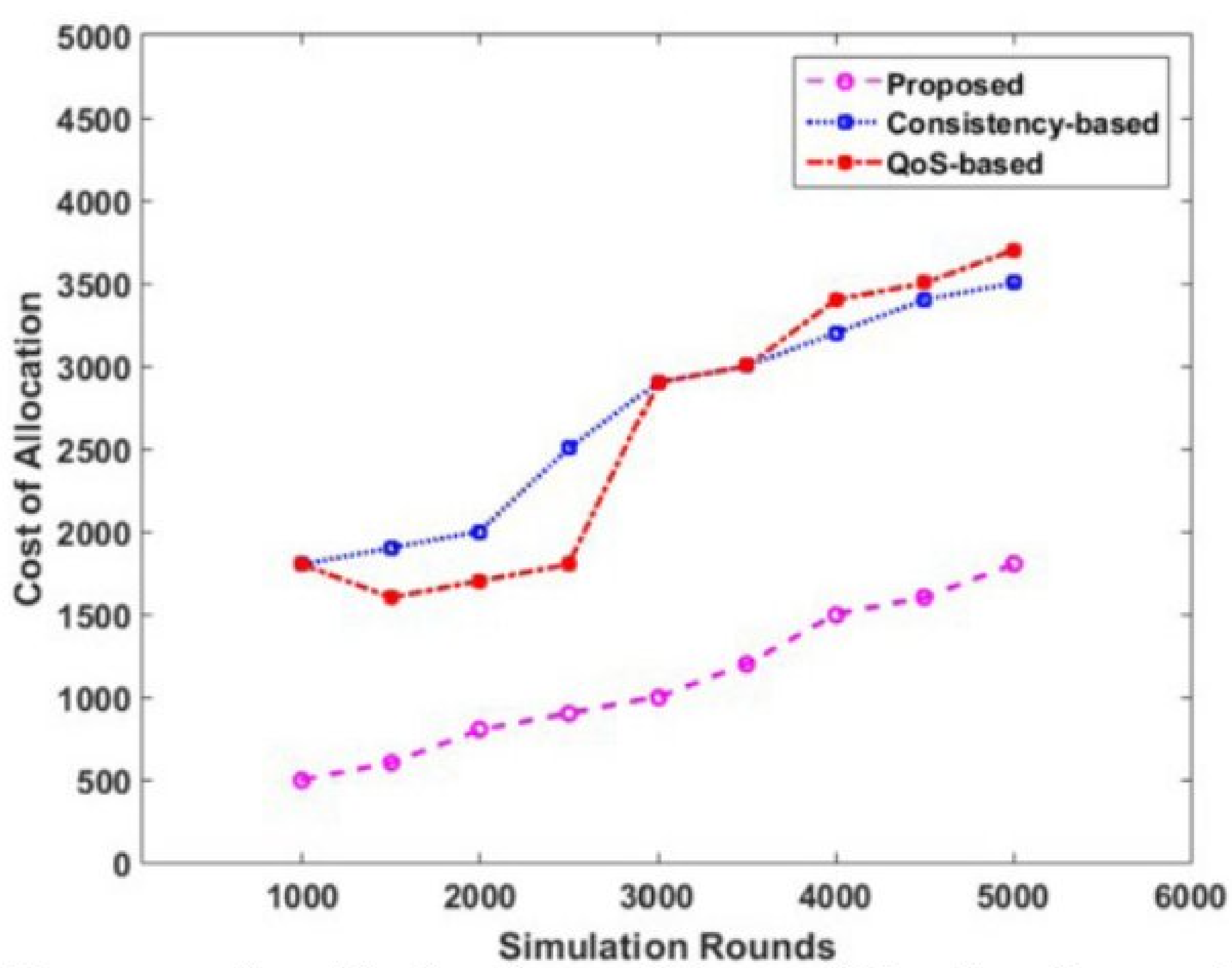

Figure - 3: Comparative Evaluation of Cost of Replica Server Placement

4. CONCLUSION

The primary contribution of the study model being proposed is the development of an innovative approach for cost-effective positioning of replica servers, specifically tailored to the use case of cloud-based CDN systems. The contributions of this proposed model can be summarized as follows: It introduces a unique method for computing costs when constructing the network's record structure, which reduces the reliance on a larger number of replica servers. This reduction in server dependency ultimately leads to cost savings in data transmission. The proposed scheme incorporates a novel concept of utilizing local virtual machines to replace traditional content servers, resulting in cost reduction. This adaptation enhances virtualization efficiency, creating an extensive chain of highly indexed structures that significantly improve data and service availability. The replica server placement strategy outlined in the proposal places equal emphasis on enhancing the user's experience and maintaining service quality. The study's findings demonstrate that the proposed system outperforms existing schemes, showcasing significantly improved performance.

REFERENCES

1. Liu, Xuan, Zhuo Li, Peng Yang, and Yongqiang Dong. "Information-centric mobile ad hoc networks and content routing: a survey." *Ad Hoc Networks* 58 (2017): 255-268

2. Stocker, Volker, Georgios Smaragdakis, William Lehr, and Steven Bauer. "The growing complexity of content delivery networks: Challenges and implications for the Internet ecosystem." *Telecommunications Policy* 41, no. 10 (2017): 1003-1016

3. Xu, Yanli, Xujie Li, and Jun Zhang. "Device-to-device content delivery in cellular networks: Multicast or unicast." *IEEE Transactions on Vehicular Technology* 67, no. 5 (2017): 4401-4414.

4. Xu, Changqiao, Wei Quan, Athanasios V. Vasilakos, Hongke Zhang, and Gabriel-Miro Muntean. "Information-centric cost-efficient optimization for multimedia content delivery in mobile vehicular networks." *Computer Communications* 99 (2017): 93-106.

5. Amadeo, Marica, Claudia Campolo, and Antonella Molinaro. "A novel hybrid forwarding strategy for content delivery in wireless information-centric networks." *Computer Communications* 109 (2017): 104-116.

6. Shuai, Qianjun, Keqin Wang, Fang Miao, and Libiao Jin. "A cost-based distributed algorithm for load balancing in content delivery network." *In 2017 9th International Conference on Intelligent Human-Machine Systems and*

Cybernetics (IHMSC), vol. 1, pp. 11-15. IEEE, 2017.
7. Bajpai, Vaibhav, Saba Ahsan, Jürgen Schönwälder, and Jörg Ott. "Measuring YouTube content delivery over IPv6." *ACM SIGCOMM Computer Communication Review* 47, no. 5 (2017): 2-11.
8. Kyryk, Maryan, Nazar Pleskanka, and Maryana Pleskanka. "Content delivery network usage monitoring." In *2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*, pp. 306-308. IEEE, 2017.
9. Shojafar, Mohammad, Zahra Pooranian, Paola G. Vinueza Naranjo, and Enzo Baccarelli. "FLAPS: bandwidth and delay-efficient distributed data searching in Fog-supported P2P content delivery networks." T*he journal of supercomputing* 73, no. 12 (2017): 5239-5260
10. Chu, Weibo, Mostafa Dehghan, John CS Lui, Don Towsley, and Zhi-Li Zhang. "Joint cache resource allocation and request routing for in-network caching services." *Computer Networks* 131 (2018): 1-14.
11. Poirion, Pierre-Louis, Jérémie Leguay, and Liu Ruosi. "Scalable request routing for VR-ready CDNs." *In 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, pp. 1-6. IEEE, 2018.
12. Yang, Hui, Hewu Li, and Qian Wu. "An integrating unicast and multicast solution by extending content delivery network to satellite." I*n 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pp. 1414-1419. IEEE, 2018.
13. Sarkar, Darothi, Nitin Rakesh, and K. K. Mishra. "Population-based clustering to enhance the utilization of surrogate in Content Delivery Networks." *Intelligent Decision Technologies* 12, no. 4 (2018): 453-460.
14. Vu, Thang X., Symeon Chatzinotas, and Björn Ottersten. "Blockchain-based content delivery networks: Content transparency meets user privacy." *In 2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1-6. IEEE, 2019.
15. Cui, Xin, Xiaohong Huang, Yan Ma, and Qingke Meng. "A load balancing routing mechanism based on SDWSN in smart city." *Electronics* 8, no. 3 (2019): 273.v
16. Kalghoum, Anwar, and Leila Azouz Saidane. "Fcr-ns: a novel caching and forwarding strategy for named data networking based on software-defined networking." *Cluster Computing* 22, no. 3 (2019): 981-994.
17. Ghasemi, Hamid, Mahdi Jafari Siavoshani, and Saeed Hadadan. "A Novel Communication Cost Aware Load Balancing in Content Delivery Networks using Honeybee Algorithm." a*rXiv preprint arXiv*:1902.04463 (2019).
18.Huang, Lemei, Yu Guan, Xinggong Zhang, and Zongming Guo. "Consistent

Dynamic Server Assignment in Content Delivery Network." *CoRR (2019).*
19. Cheng, Yinhang. "Routing-Independent Anycast For IPv6 Content Delivery Networks." PhD diss., Case Western Reserve University, 2019.
20. Goudarzi, Pejman, and Mahmoud Mousavinejad. "A Lyapunov optimization approach to energy-efficient content delivery networks." *Simulation* 96, no. 3 (2020): 269-280
21. Harahap, Erwin, Yusuf Fajar, Dadi Ahmadi, Abdul Kudus, and Rakhmat Ceha. "Modeling of request routing management on router for content delivery network." *International Journal of Scientific & Technology Research* 9 (2020): 308-14.
22. Q. Fan, H. Yin, L. Jiao, Y. Lyu, H. Huang and X. Zhang, "Towards Optimal Request Mapping and Response Routing for Content Delivery Networks," in *IEEE Transactions on Services Computing*, vol. 14, no. 2, pp. 606-613, 1 March-April 2021, doi: 10.1109/TSC.2018.2796567.
23. Omotunde, A. A., S. O. Okolie, Y. A. Adekunle, A. A. Izang, and O. O. Ebiesuwa. "Request-Routing for Content Delivery Networks (CDN)." I*nternational Journal of Advanced Research in Computer Science* 6, no. 6 (2021).
24. Ghaznavi, Milad, Elaheh Jalalpour, Mohammad A. Salahuddin, Raouf Boutaba, Daniel Migault, and Stere Preda. "Content Delivery Network Security: A Survey." *IEEE Communications Surveys & Tutorials* (2021).
25. Varshney, Pulkit, Sanchit Gaur, and Satya Pal. "A novel approach of Load Balancing in Content Delivery Networks by optimizing the surrogate server." In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 790-795. IEEE, 2021.
26 Ghaznavi, Milad, Elaheh Jalalpour, Mohammad A. Salahuddin, Raouf Boutaba, Daniel Migault, and Stere Preda. "Content Delivery Network Security: A Survey." *IEEE Communications Surveys & Tutorials* (2021).
27. H. Hu, Y. Wen, T. Chua, J. Huang, W. Zhu, and X. Li, "Joint Content Replication and Request Routing for Social Video Distribution Over Cloud CDN: A Community Clustering Method," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 7, pp. 1320-1333, July 2022, doi: 10.1109/TCSVT.2015.2455712.
28. Priyanka D, Channakrishnaraju, Chethan B K "Insights on Effectiveness Towards Research Approaches Deployed in Content Delivery Network", in the Springer Series: "Software Engineering Perspectives in Systems" in the *proceedings of 11th Computer Science Online Conference* 2022, Volume 1, (CSOC2022).
29. Priyanka D, Channakrishnaraju, Chethan B K "A Novel Cost-based Replica

Server Placement for optimal service quality in cloud-based Content Delivery Network-"*International Journal of Electrical and Computer Engineering*" Scopus based Q2 Journal- ISSN 2088-8708, VOL-13-2023

30. C. Xu, C. Xu, and B. Li, "Multi-agent deep Q-network based dynamic controller placement for node variable software-defined Mobile Edge-Cloud Computing Networks," *Mathematics*, vol. 11, no. 5, p. 1247, 2023, doi: 10.3390/math11051247.