

Framework for content server placement using integrated learning in content delivery network

Priyanka Dharmapal¹, Channakrishnaraju²

¹Department of Computer Science and Engineering, Sri Siddhartha Academy of Higher Education University, Tumkur, India

²Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumkur, India

Article Info

Article history:

Received Sep 15, 2023

Revised Oct 9, 2023

Accepted Jan 6, 2024

Keywords:

Content delivery network

Content placement

Content server

Machine learning

Predictive

ABSTRACT

Content placement is a significant concern in content delivery networks (CDN), irrespective of various evolving studies. Existing methodologies showcase various significant unaddressed issues concerning content placement approaches' complexities. Therefore, the proposed study presents a novel computational framework towards dynamic content placement strategy using a novel integrated machine learning approach. Simplified mathematical modelling is used to formulate and solve the content placement problem. At the same time, reinforcement learning and the sequential attentional neural network have been utilized to optimize the decision-making towards placement of content servers. Designed and assessed over a Python environment, the proposed scheme is witnessed to exhibit 35% reduced bandwidth utilization, 20% reduced delay, 23% reduced computational resource utilization, and 28% reduced algorithm processing time in contrast to existing predictive content placement schemes.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Priyanka Dharmapal

Department of Computer Science and Engineering, Sri Siddhartha Academy of Higher Education University

Tumkur, Karnataka 572017, India

Email: prsh2019@gmail.com

1. INTRODUCTION

From the perspective of modern internet infrastructure, the content delivery network (CDN) facilitates a dedicated transmission of requested content over many interconnected servers towards content distribution [1]. A CDN environment can optimise the data transfer while reducing the channel capacity cost for both the user and service provider [2]. The contents are replicated to multiple content servers on multiple locations that directly contribute to the efficient availability of content [3]. CDN also mitigates significant downtime risk due to its scalable performance towards handling high-traffic loads [4]. Various current research-based approaches are proven to offer highly efficient modelling of CDN environment [5]–[7]. However, it also suffers various challenges due to network complexities and content adaptation [8]. Apart from this, a CDN system also suffers from content freshness [9], security concerns [10], and integration with edge computing [11]. Out of all these, content placement is one of the primary significant problems encountered in CDN systems [12]. A content placement problem relates to undertaking a strategic decision towards finding an effective content server to store and distribute the content across the network. This should improve overall efficiency, cost reduction, latency reduction, and optimized content distribution performance [13]. This problem defines identifying the specific set of contents that must be replicated or cached over the edge server's location within the CDN infrastructure coverage [14]. To deploy an effective content placement algorithm, it must comply with load balancing, geographical distribution, dynamic content, cache management, and content popularity [15]. Notably, replicating or caching the necessary content over multiple

locations of edge servers is quite expensive. Hence, there is a need to design a cost-effective decision-making system to do this task. Another significant challenge of content placement in existing time is aligning placement strategy with the delivery mechanism. There is a fair possibility of deploying manifold delivery strategies for heterogeneous forms of content, which can impose significant challenges in the CDN environment. It is also necessary for a CDN system to offer an assurance towards a higher degree of consistency of cached contents and coherency over manifold edge servers. However, most of the underlying issues associated with content placement have not been effectively addressed in existing approaches, leaving an open scope for further research [16], [17].

Therefore, the proposed research presents a computational framework for content server placement using a machine-learning approach. The contributions of the model are as follows: i) the scheme presents a replica/content server placement strategy emphasizing both quality of experience and quality of service, ii) an integration machine learning using reinforcement learning and sequential attentional neural network is used for optimizing the decision making towards location of server placement, iii) the study model improves the service delivery by reducing bandwidth consumption and delay while performing the predictive operation, and iv) the model is designed to reduce computational resource utilization as well as faster algorithm processing time thereby offering a cost-effective framework towards replica server placement in CDN environment. The manuscript's organisation is as follows: section 2 discusses the adopted research methodology, while problem formulation is discussed in section 3. System implementation is elaborated in section 4. The result discussion is carried out in section 5, while the conclusion is presented in section 6.

Currently, various studies are being carried out to address content placement issues in CDN. Cao *et al.* [18] presented a bioinspired algorithm to sort out an issue of placing edge servers. The technique deploys an optimization method based on the fruit fly's cognitive behaviour for maximized precision and faster convergence speed. Chandrasekaran *et al.* [19] have presented a scheme associated with placing the storage unit over a cloud environment as a container. This work aims to deploy context-based services to offer location-independent content delivery services. Reali and Femminella [20] have presented a unique network caching mechanism to increase the hit ratio of the content over a 5G network system. The study model presented a distinct overlay system towards caching the content distribution and was proven to be better than the conventional least-frequently used caching scheme. Musa *et al.* [21] have developed a scheme towards optimizing the edge caching principle. This framework's core notion is to offer mobility services over vehicular networks using proactive caching. The scheme uses the Markov process to model the vehicular network infrastructure with a core agenda to enhance the data transmission with a minor transmission delay. The existing system has reported various studies where caching-based methodologies have been adopted towards content delivery over a large channel. Research by Delvadia *et al.* [22] has presented a unique caching mechanism to forward the task request. Although the study model emphasizes an information-centric network (ICN), it presents a mechanism of content placement using the Markov chain principle in the shortest time. Research by Gui and Chen [23] has presented another strategy of content placement using a caching strategy based on a weighted entropy mechanism. The model has presented a cache replacement algorithm that can minimize the redundancy present over the path of content delivery. Wu *et al.* [24] used the cache replacement scheme to minimise content propagation delay using the cooperative caching principle for edge networks. Zhou *et al.* [25] presented a study towards proactive caching, considering the mobility constraint associated with caching strategy over delivery networks.

Länderanta *et al.* [26] have discussed a server placement method to address the issues of the extensive distance between the access point and the content server. The applicability of this model is assessed on both high and low-capacity servers. A study towards the popularity-based content placement method is presented by Li *et al.* [27], where the idea of the model is to address the issues associated with frequent replacement of cache and distribution of contents unreasonably. Research by Liu and Han [28] has presented a distinct allocation scheme towards cache memory to ensure an effective data dissemination process. The study model implements a partitioning technique using a hierarchical scheme to allocate the size of cache information. Research by Silva *et al.* [29] has discussed a performance evaluation mechanism towards content placement problems in association with caching operation. The study has discussed the conventional caching principles and performed a comparative analysis over them using the number of hops upstream, retransmission, delay in retrieving information, network traffic, and ratio of cache hit. The result of the study shows that the least-frequently used scheme is proven to offer better performance compared to other caching approaches.

Sun *et al.* [30] have presented a study towards the CDN environment emphasizing over analysis of request logs. This work aims to facilitate seamless content access to users with variable traffic distribution. The study model has also used machine learning, deep learning, and statistical models to develop this tool. It has been noted that the advent of federated learning approaches is adopted towards mechanizing collaborative cloud computing [31]. However, it is still in its infancy stage. Adopting deep learning is witnessed in Kang and Chung [32], where an optimized content placement strategy is formulated using long

short-term memory (LSTM). The study model is designed using a conventional content popularity-based placement method with a bioinspired algorithm. Adopting a generative adversarial network (GAN) is proven beneficial for high dimensional content delivery as reported in [33].

The core idea of this study model is to deliver the highest possible quality in content propagation. The adoption of machine learning is witnessed in [34], where the predictive modelling mainly emphasises energy efficiency. The model uses an extreme learning method with an online sequential methodology to determine the optimal form of network for content propagation. The adoption of machine learning and artificial intelligence is witnessed in [35], where the authors present a discussion towards beneficial characteristics of these concerning processing edge-based information. Research by Shi *et al.* [36] has presented the usage of a deep reinforcement learning approach for facilitating an efficient caching process associated with multimedia content delivery over a cloud system. The model has used Markov's decision to address the agent's operation towards learning the strategy of local cache. It also targets to reduce the number of transmissions over multi-cloud systems characterized by restricted storage capacity. The idea of this model is also to minimize the service latency and enhance the backhaul network system's computational efficiency. Tang *et al.* [37] have used reinforcement learning to stream multimedia content to enhance the quality of experience. Therefore, there are various studies towards content placement in the existing era. At the same time, the discussion of identified research problems after reviewing the above literature is carried out in the next section.

After reviewing the methodologies of existing solutions towards addressing problems of content placement in CDN systems, it is noted that caching-based strategies are frequently adopted. At the same time, there is a progressive gain in adopting machine/deep learning approaches. Irrespective of beneficial outcomes stated in existing literature, the following are certain loopholes identified:

- Issues in dynamic placement of content: various popularity-based caching strategies are implemented; however, these schemes have apriori definitions of popularity scores without considering the dynamic possibilities. For this reason, adopting such schemes doesn't let CDN adapt to dynamic user demands.
- Fewer studies towards updating strategies: existing schemes towards content placement don't offer a seamless connection between the content server, CDN, and user. A one-directional task request is always being formulated, which restricts the determination capabilities of either updating or removing the contents over the cache to assure users of fresh content.
- Issues in the selection of server: at present, cloud network extensively utilises either fog or edge servers and data centres. Hence, existing studies don't offer a predictive or direct determination of an eligible content server that can effectively process the dynamic task request over a service chain in CDN. The majority of existing studies address these issues considering geographical proximity, network condition, and server load, and they are identified not to consider content placement function effectiveness over dynamic networks as well as lack of resource utilization. Such schemes sound effective from a research viewpoint; however, they will encounter issues when deployed in the practical scenario of CDN.
- Complex learning implementation: existing predictive approaches have increasingly used machine and deep learning approaches. However, deep learning is gaining a progressive pace where it is noted that higher predictive accuracy is obtained at the cost of computational complexity. Further, the reinforcement learning scheme is proven a better alternative, but its applicability towards content placement is still seen.

From the existing literature review, it has been noticed that there is less work towards replica management. The mechanism of implementation of machine learning is also required to be revised to offer better performance in terms of service delivery. Apart from this, it is also observed that conventional CDN is limited to facilitating delivery paths with the aid of local decisions. Not much inclusion is carried out towards global dynamics towards peak traffic conditions in cloud-based CDN. Hence, the problem statement is "to generate a cost-efficient data transmission scheme in the presence of peak traffic conditions for better replica management in cloud-based CDN is highly challenging." The following section discusses about adopted research methodology.

2. RESEARCH METHOD

The prime aim of this framework is to design a computational framework of cloud-based CDN that can facilitate cost-effective replica server management for enhanced service delivery. The proposed system will adopt an analytical research methodology where the problems of conventional CDN are improved upon by obtaining local and global bottleneck information based on learned data packets. The first part of the implementation will be developing a cloud-based CDN communication model. The study will define local bottleneck information as the duration of forwarding data to neighbouring nodes from cloud proxy servers. The global bottleneck information will be obtained from the bottleneck condition of the existing cloud proxy server to the end proxy servers. This information is stored as a matrix, updated using learned packets. This is

a significant contribution where the proposed model offers better adaptability towards dynamic topology and selection of communication links in cloud-based CDN based on reduced cost computed. Figure 1 highlights the proposed architecture.

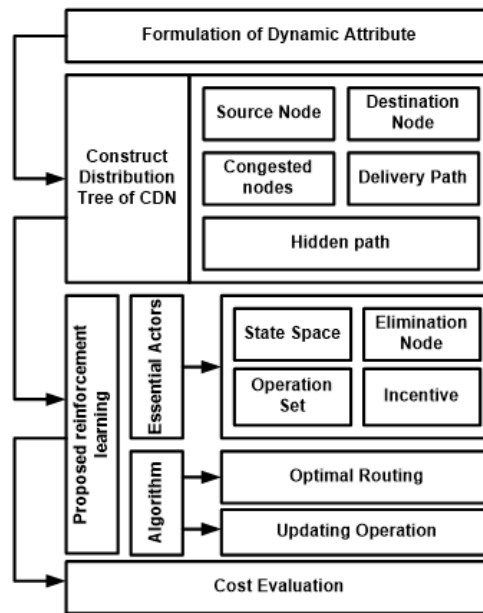


Figure 1. The architecture of 1st module of implementation

According to Figure 1, the proposed system considers three types of nodes, i.e., source server, cloud proxy server, and user in a cloud-based CDN system. The first operation block is about formulating dynamic attributes to facilitate modelling peak traffic conditions in a cloud environment. A tree-based concept is used for developing this topology, which mainly has nodes as vertices in the network and communication links as the edges. Further, this tree structure will formulate roles for nodes (source node, destination node, congested nodes) and links (delivery path, hidden path). The next part of the implementation will be focused on implementing a learning mechanism using a reinforcement learning scheme. This is because the reinforcement learning scheme facilitates solving problems associated with global optimization compared to other machine learning schemes. The proposed learning scheme will consist of formulating four different actors. The first actor is state space, which consists of the node's saturation state concerning its neighbouring node to transmit the data using specifically formulated conditions for making the data transmission eligible between two adjacent nodes. The second actor is the elimination node, a node with a copy of data from source nodes with a higher correlation. The third actor is the operation set, which selects the route from the current node to its adjacent nodes. The fourth actor is an incentive, allocating an incentive to a specific node to reach its packet to a neighbouring node through a defined path.

The proposed system will also present two specific algorithm formulations, where the first algorithm will be focused on optimal routing. In contrast, the second algorithm will focus on updating the learned data matrix. With the aid of an objective function, the algorithm ensures the link selection with lower congestion cost. Finally, the proposed system performs cost computation associated with the bottleneck condition, where an objective function will be defined for data transmission during the bottleneck condition. The idea is to obtain the cumulative cost of all the transmission probabilities and then sort it to select the reduced cost. The proposed system's novelty is that it can replica server placement considering the dynamic characteristics of a cloud-based CDN system, unlike any existing approaches in literature. Another novelty of the proposed system is its capability to make an efficient routing decision using the proposed reinforcement learning scheme in peak traffic conditions.

The novelty and contribution of the proposed system is towards developing an optimized solution towards addressing the content placement problem. The presented study model can distribute and store essential demanded contents subjected to delivery to end users over the CDN environment. Various novelty attributes introduced in the proposed research methodology are as follows:

- Unlike existing approaches towards content placement in a CDN environment that mainly use spatial distance and traffic load, the proposed system mechanism is a novel decision-making approach that allows an appropriate placement of point-of-presence (PoP) over various locations worldwide. Such geographical distribution lets the virtualized contents be highly available for its end users, minimizing the latency.
- The proposed system carries out an anycast routing mechanism that is responsible for automatically directing the request generated by the user to the distributed algorithm running over an edge server that can make decisions to find the nearest available content server. Hence, the proposed scheme can offer significant latency control and reduce network hop-based propagation.
- One of the prominent contributions of the proposed system is its capability to analyse traffic demands using an attention-based neural network. The structure of the learning-based scheme is designed to analyze the captured network information required to carry out predictive analysis associated with the degree of user demands. This is a highly essential contribution as this learning process also assists in dynamically finetuning the content placement by the dynamic form of demands.
- A unique and revised form of caching principle is introduced in the proposed system, different from existing approaches reported in prior sections. Most existing caching principles are designed for content caching, object caching, full-page caching, session-based caching, time-based caching, and mobile caching. However, the proposed scheme introduces a unique prefetch and predictive caching. According to this unique research methodology, the scheme uses a learning approach to expect specific set of information that a user might generate a request followed by proactively performing caching to reduce the latency associated with the process. Further, reinforcement learning and attention-based neural networks optimize these caching strategies based entirely on user behaviour patterns.
- The study model prioritizes catering to the user's demand in the CDN environment and considers the content servers' sustainable operation. The proposed research methodology also offers its formulation where energy dissipation in each content server and resources demanded by them are considered prime constraints. For this purpose, the study model uses content placement functions and considers the path among content servers to reduce resource consumption owing to algorithmic decisions of content placement, storage, maintaining server infrastructure, and propagating requested content over an optimized route with reduced latency.
- Finally, the ultimate contribution of the proposed study model is to deploy a simplified and computationally efficient machine learning method that considers network conditions, user dynamic behaviour, and dynamic generation of task requests over service chains of contents servers. The proposed study uses a reinforcement learning approach to optimize the policy of content placement. At the same time, the attention-based neural network is adopted to optimize further, considering the practical constraints associated with content placement problems in CDN.

Based on the novelties mentioned above and their contribution, it can be stated that the proposed scheme offers a novel framework where integrated machine learning has been used to optimise the solution towards addressing content placement problems in CDN. The next section further elaborates on the problem formulation followed by system implementation. Before illustrating the system design, it is necessary to look into the actual problem formulation for the proposed study. For this purpose, consider several content servers available in the CDN system such that $a \in A$ that are equipped with c ($c \in C$) computational resources. All the content servers c is connected with p path ($p \in P$) considering star topology. This problem formulation aims to realize the need for an optimal routing leading due to the proper content placement process over a distributed CDN system. Therefore, the problem formulation of the proposed study is expressed via an objective function O_{fun} as in (1):

$$O_{fun} = m_{fun}\{(X_1 + X_2) + X_3\} \quad (1)$$

In (1), the objective function is exhibited to be formulated using three dependable parameters X_1 , X_2 , and X_3 that are further subjected to minimization function m_{fun} . The first parameter X_1 is a product of energy dissipation and resource demanded by the content placement function, which is mathematically represented as in (2):

$$X_1 = (E_a \cdot C_\theta \cdot V_{\alpha,a})^\alpha \quad (2)$$

In (2), the computation of parameter X_1 is carried out considering i) E_a energy dissipated in each content server a , ii) C_θ summed value of resources demanded by content placement function θ , and iii) $V_{\alpha,a}$ variable

for content server placement for α functions in a content server, such that variable $\alpha \in \theta$. On the other hand, the variable X_2 is mathematically represented as in (3):

$$X_2 = (E_{a(\min)}, d_a)^a \quad (3)$$

In (3), $E_{a(\min)}$ represents passive energy dissipation of a content server a while d_a represents the decision attribute for a content server a . Finally, the last variable X_3 is mathematically represented as in (4):

$$X_3 = E_p \cdot \sigma_\theta^t \cdot V_{\alpha,a} \quad (4)$$

In (4), the computation of final variable X_3 for objective function is carried out using E_p energy dissipated in each path of propagation and σ_θ^t channel capacity required by content placement function θ in servicing chains of task t . Further, it is required to be noted that the objective function O_{fun} can be stated to be subjected to optimization only when it caters up to three following empirical conditions:

$$\text{Condition}_1: X_1 \cdot X_2 < d_a \cdot q_{ca} \quad (5)$$

$$\text{Condition}_2: \sigma_\theta^t \cdot V_{\alpha,a} < d_p \cdot \sigma_p \quad (6)$$

$$\text{Condition}_3: [\gamma_\theta \cdot V_{\alpha,a} + \gamma_p \cdot V_{\alpha,a}] < \gamma^t \quad (7)$$

In (5) to (7), the new variable q_{ca} , d_p , σ_p , γ_θ , and γ_p represent the quantity of available resources in a content server, decision attribute for p path of propagation, channel capacity of p path of propagation, delay incurred on servicing θ content placement function, and delay incurred in servicing towards p path of propagation. Therefore, the prime research challenge is to explore the optimal location for content placement represented as PL, equivalent to V_α , which represents a decision variable for content placement using Boolean order. The idea is to find if the function α is a part of the content placement function θ or not.

The proposed study model deploys the reinforcement learning approach to perform an optimized content placement in the CDN system. Figure 2 highlights the mechanism of the proposed learning scheme where a feedback loop is constructed for action and reward. According to the presented learning mechanism exhibited in Figure 2, the model acquires the input in the form of task $t=(\alpha_1, \alpha_2, \dots)$ with an outcome of content placement attribute $\beta_t=(\beta_1, \beta_2, \dots)$ that represents the assignment of each content placement function θ . The learning approach introduces a content placement scheme ρ_w concerning content placement attribute β_t and task t , where w represents the weight of the neural network. The scheme restricts the learning architecture to a sequence concerning its outcome to control the computational efficiency score. As per this methodology, when the system receives the request for a specific task from any user in CDN, the agent in the reinforcement learning algorithm constructs a state vector. It further incorporates the environment state in the form of tasks and specific tasks requested by clients globally. This operation is followed up by generating an associated content placement vector in the form of action in reinforcement learning that is used for determining the need to substitute the new task. The computation of the decision towards content placement is carried out. It is further used for assessing the signal for determining the signal quality in the form of reward in reinforcement learning. Therefore, the proposed scheme introduces a robust interface between the system environment and the agent of the proposed reinforcement learning model. This phenomenon assists in the relaxation of the constraints to accomplish a higher degree of positive rewards from the reinforcement learning model. The algorithmic steps towards optimized content placement are as follows:

The Algorithm 1 takes the input of η (nodes) and α (content placement function) that, after processing, yields an outcome of H (solution matrix for content placement). According to this algorithm, the system considers all the nodes requesting tasks in the CDN environment where the overall deployment of the proposed reinforcement learning architecture is presented (line-1). The agent performs sequential operation where a novel encoder and decoder design is formed. The core idea is to realize and retain information about the extended dependencies in the network to carry out sequential predictive operations. The proposed study model considers all the content placement functions θ in sequence to act as the input to the predictive learning model. This model consists of various tasks associated with CDN services t , which further consists of individual functions of content placement α present in the content server of n numbers (line-2). One of the prominent contributions of this operation is that it can process various multi-dimensional input tasks of the CDN traffic system without any dependencies towards its internal operation. For this purpose, an attentional network is constructed to increase the performance of the presented learning approach. The proposed scheme uses an attentional-based neural network to form its decoding operation with equivalent steps to a sequence of input of content placement function. With every progress in learning operation, the scheme lets the decoder

generate a content server to place its requested contents incorporated by the encoder module. For this purpose, the algorithm forms a latent state of the attention network associated with the decoder as μ_t (line-4) that is a content placement function for each content server α with revised elements, i.e., μ_{s-1} , μI_{s-1} , l_s (line-4).

Algorithm 1: Optimized content placement

Input: η , α

Output: H

Start

1. For $i=1:\eta$
2. $t=(\alpha_1, \alpha_1, \dots, \alpha_n)$
3. $l_s=f_1(\psi_{t,a}, \mu_a)$
4. $\mu_t=\alpha(\mu_{s-1}, \mu I_{s-1}, l_s)$
5. $h=\text{rank}(\mu_s, \mu I_a)=f_2(\lambda_1+\lambda_2)$
6. If $h<T$
7. $H=h(\theta, n)^a$
8. End
9. End
- End

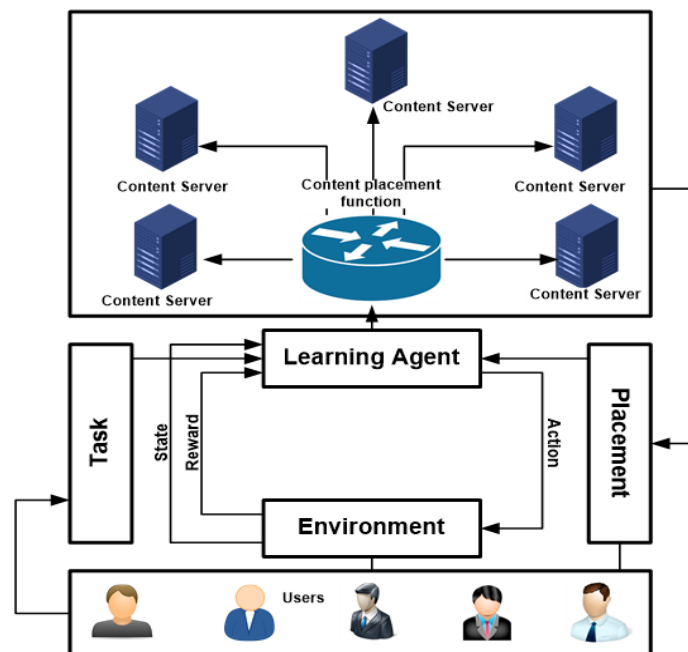


Figure 2. Mechanism of reinforcement learning

It should be noted that this latent state is a function linked with a prior state integrated with an attention network connected to the hidden states of the encoder. The variable μ_{s-1} and μI_{s-1} represents primary and secondary hidden states associated with the attention network. In contrast, the third variable l_s represents integrated latent states associated with the sequential input elements of this algorithm. This variable of latent state l_s is empirically expressed further as a function $f_1(x)$ with an input argument of $\psi_{t,a}$ and μ_a (Line-3). The function $f_1(x)$ performs a series of summation of values obtained by the product of $\psi_{t,a}$ and μ_a that represents the weight attribute of the latent state of the network concerning task t and content server a and the latent state of decoder in a network concerning individual function α of content placement function (Line-3). To reduce the computational complexities associated with the learning operation, the proposed scheme uses an equivalent number of involved steps for the alignment variable size and sequence of sources. The computation of this variable $\psi_{t,a}$ is carried out by applying an activation function to scale up numbers in the form of probabilities, thereby offering a more significant number of candidate solutions of the position of content servers per demand of generated task in CDN.

Further, the system computes the rank values of this placement considering the present latent target of decoder μ_s with all individual states of source, i.e., μl_a (line-5). The algorithm further defines the ranking order value using a function $f_2(x)$ considering two input arguments λ_1 and λ_2 (line-5). The function considers the hyperbolic tangent function to play the role of activation function considering content function θ . At the same time, the empirical definition of newly incorporated variables λ_1 and λ_2 are represented as in (8) (line-5):

$$\lambda_1 = E_1 \cdot \mu_s \text{ and } \lambda_2 = E_2 \cdot \mu l_s \quad (8)$$

In (8), the variables E_1 and E_2 represent target weight attributes subjected to learning while exploring an optimal position of the content server. After the computation towards acquiring all the candidate solutions of content server placement is accomplished, the presented algorithm assigns a threshold T , which defines a maximum cut-off of resources required to perform this task of content placement along with a sequential learning process. This is quite a challenging task as such a threshold value can eventually differ from one geographical location based on the fluctuating demands of users in CDN. Hence, to simplify this process, the proposed scheme let its agent carry out the search towards multiple candidate solution evaluated along with energy being dissipated on each route and towards each content server location. A route with the least cost of both energy and distance is considered as the finalized value of threshold T , which is compared with generated ranking scores h (line-6). Although the system generates multiple ranking scores of content placement in CDN, the best position will be characterized by any ranking score within the value of threshold T (line-6). Finally, the algorithm generates the confirmed content server location H by revising the ranking score matrix h concerning content placement function θ and several content placement function n considering several content servers a (line-7). Hence, the prime novelty of this algorithm is that it carries out optimized content placement operations considering the practical demands of resources and all dependencies for better reliable implementation.

3. RESULTS AND DISCUSSION

This section presents the outcome accomplished after implementing the proposed study model discussed in the prior section. The implementation has been done in a Python environment with a target area of $1,000 \times 1,000 \text{ m}^2$. The study considers 50 nodes and 60 number of edge nodes. The initialized energy for each node is taken as 0.5 J. To effectively analyze the outcome, the proposed study has been compared with related predictive learning approaches of deep reinforcement learning [38], deep Q-learning [39], and Q-learning [40]. The assessment is based on the performance metric scale of bandwidth utilization, delay, computational resources, and algorithm processing time. The results obtained from the simulation study are shown in Figure 3.

Figure 3(a) showcases that the proposed scheme offers approximately 35% reduced bandwidth utilization in contrast to its existing approaches of deep reinforcement learning, deep Q learning, and Q-learning. Although the Q-learning strategy can solve issues about actions space and discrete state, its extensive exploration strategy towards confirming efficient content placement server is relatively high, leading to increased bandwidth utilization [41]. Deep Q-learning somewhat solves this problem by offering a function approximation capable of learning complex policies. However, it suffers from significant instability issues while performing training, causing a slightly increased bandwidth utilization (although slightly lower than the Q-learning approach). Finally, the deep reinforcement learning algorithm has more flexibility towards complex CDN architecture with dynamic content placement issues [42].

Figure 3(b) exhibits that the proposed scheme offers approximately 20% delay reduction compared to existing predictive approaches. Similar characteristics of existing approaches can be attributed to this outcome concerning Q-learning and deep Q-learning. However, deep reinforcement learning offers optimal policy and feature representation, yet they depend more on substantial data during training operations. This is quite frequent when it encounters dynamic requests for tasks from users.

The outcome in Figure 3(c) showcases that the proposed scheme offers approximately 23% lower computational resources than the existing system. There are two reasons for this outcome, viz. i) the proposed scheme has an inclusion of energy as the prime constraint, which the complete decision towards the selection of content server is based on the energy budget on the optimal route, and ii) proposed scheme involves the decision attribute of content server considering the quantity of available resource in them. This cause the model to generate more candidate solutions while the threshold converges to the final location of the content server, causing less involvement of computational resources [43]. Unfortunately, these properties cannot be witnessed within existing predictive approaches. Figure 3(d) showcases that the processing time of the proposed scheme is approximately 28% reduced compared to existing approaches. The prime reason for this is proposed scheme offers less iterative operation, causing reduced processing time [44].

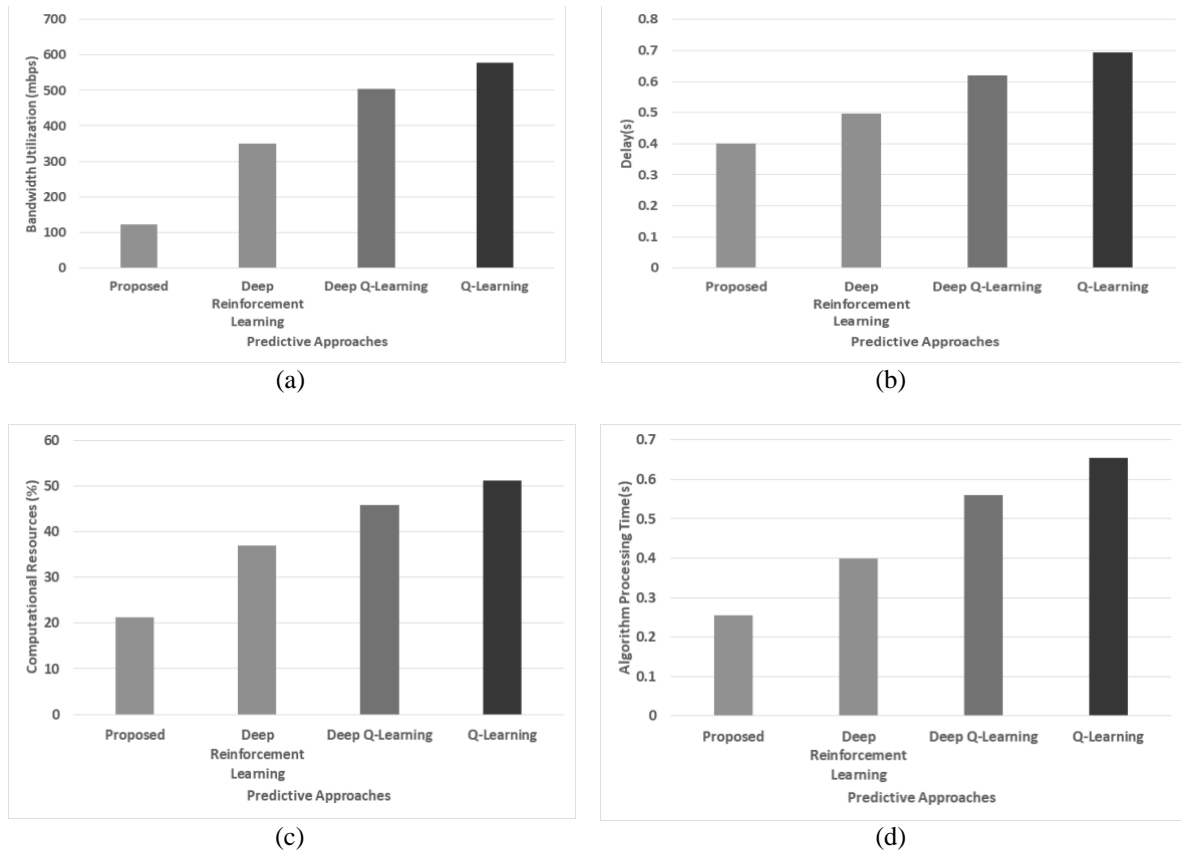


Figure 3. Comparative analysis of (a) bandwidth utilization, (b) delay, (c) computational resources, and (d) processing time

4. CONCLUSION

The proposed scheme contributes to introducing a simplified yet distinct solution for content placement in the CDN environment. The principle of optimization has been carried out using reinforcement learning and attentional neural network sequential model. The contribution of the proposed scheme is as follows: a simplified mathematical problem formulation and problem solution have been presented where the multi-objective criterion is formulated to accomplish the optimal placement of the content server. The proposed scheme uses a reinforcement learning approach and sequential attentional neural network as an integrated machine learning approach to incorporate the optimization principle to address dynamic task processing in the CDN environment. A unique ranking mechanism is presented towards all the candidate solutions obtained by an integrated machine learning approach, capable of prioritizing the demands of traffic distribution and catering to multiple dynamic tasks in the service chain. The study outcome of the proposed scheme is found to offer better performance in all the considered performance metrics in contrast to existing predictive approaches towards solving content placement problems in CDN.

REFERENCES





- [1] I. L. -Mayorga *et al.*, "Network-coded cooperation and multi-connectivity for massive content delivery," *IEEE Access*, vol. 8, pp. 15656–15672, 2020, doi: 10.1109/ACCESS.2020.2967278.
- [2] Y. Fan, B. Yang, D. Hu, X. Yuan, and X. Xu, "Social- and content-aware prediction for video content delivery," *IEEE Access*, vol. 8, pp. 29219–29227, 2020, doi: 10.1109/ACCESS.2020.2972920.
- [3] M. S. Al-Abiad, A. Douik, and M. J. Hossain, "Coalition formation game for cooperative content delivery in network coding assisted D2D communications," *IEEE Access*, vol. 8, pp. 158152–158168, 2020, doi: 10.1109/ACCESS.2020.3020472.
- [4] J. Chuan, B. Bai, X. Wu, and H. Zhang, "Optimizing content placement and delivery in wireless distributed cache systems through belief propagation," *IEEE Access*, vol. 8, pp. 100684–100701, 2020, doi: 10.1109/ACCESS.2020.2996222.
- [5] H. Yang, H. Pan, and L. Ma, "A review on software defined content delivery network: a novel combination of CDN and SDN," *IEEE Access*, vol. 11, pp. 43822–43843, 2023, doi: 10.1109/ACCESS.2023.3267737.
- [6] C. T. E. R. Hewage, A. Ahmad, T. Mallikarachchi, N. Barman, and M. G. Martini, "Measuring, modeling and integrating time-varying video quality in end-to-end multimedia service delivery: a review and open challenges," *IEEE Access*, vol. 10, pp. 60267–60293, 2022, doi: 10.1109/ACCESS.2022.3180491.

- [7] M. I. A. Zahed, I. Ahmad, D. Habibi, Q. V. Phung, M. M. Mowla, and M. Waqas, "A review on green caching strategies for next generation communication networks," *IEEE Access*, vol. 8, pp. 212709–212737, 2020, doi: 10.1109/ACCESS.2020.3040958.
- [8] Z. H. Wang and A. L. Dong, "Research on load balancing and caching strategy for central network," *Journal of Electrical and Computer Engineering*, vol. 2022, pp. 1–12, Apr. 2022, doi: 10.1155/2022/6601965.
- [9] M. Karaata, A. Al-Mutairi, and S. Alsubaihi, "Multipath routing over star overlays for quality of service enhancement in hybrid content distribution peer-to-peer networks," *IEEE Access*, vol. 10, pp. 7042–7058, 2022, doi: 10.1109/ACCESS.2021.3139936.
- [10] J. Zhang, S. Li, and C. Wang, "A secure dynamic content delivery scheme in named data networking," *Security and Communication Networks*, vol. 2022, pp. 1–15, 2022, doi: 10.1155/2022/6304927.
- [11] C. Fang *et al.*, "Energy-efficient hierarchical collaborative scheme for content delivery in mobile edge computing," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–10, Apr. 2022, doi: 10.1155/2022/8392511.
- [12] C. Natalino, A. D'Sousa, L. Wosinska, and M. Furdek, "Content placement in 5G-enabled edge/core data center networks resilient to link cut attacks," *Networks*, vol. 75, no. 4, pp. 392–404, 2020, doi: 10.1002/net.21930.
- [13] C. Ding, A. Zhou, J. Huang, Y. Liu, and S. Wang, "ECDU: an edge content delivery and update framework in mobile edge computing," *Eurasip Journal on Wireless Communications and Networking*, vol. 2019, no. 1, 2019, doi: 10.1186/s13638-019-1590-2.
- [14] H. Wu, Y. Fan, Y. Wang, H. Ma, and L. Xing, "A comprehensive review on edge caching from the perspective of total process: placement, policy and delivery," *Sensors*, vol. 21, no. 15, 2021, doi: 10.3390/s21155033.
- [15] A. Adel, "Utilizing technologies of fog computing in educational IoT systems: privacy, security, and agility perspective," *Journal of Big Data*, vol. 7, no. 1, Nov. 2020, doi: 10.1186/s40537-020-00372-z.
- [16] B. Zolfaghari *et al.*, "Content delivery networks: state of the art, trends, and future roadmap," *ACM Computing Surveys*, vol. 53, no. 2, pp. 1–34, 2021, doi: 10.1145/3380613.
- [17] M. Ghaznavi, E. Jalalpour, M. A. Salahuddin, R. Boutaba, D. Migault, and S. Preda, "Content delivery network security: a survey," *IEEE Communications Surveys and Tutorials*, vol. 23, no. 4, pp. 2166–2190, 2021, doi: 10.1109/COMST.2021.3093492.
- [18] Q. Cao, B. Liu, and Y. Jin, "Locality sensitive hashing-aware fruit fly optimization algorithm and its application in edge server placement," *Journal of Cloud Computing*, vol. 11, no. 1, 2022, doi: 10.1186/s13677-022-00313-6.
- [19] G. Chandrasekaran, B. Ramasamy, P. Dhondi, P. B. Thorat, and R. Challa, "CASE: a context-aware storage placement and retrieval ecosystem," *IEEE Access*, vol. 10, pp. 1956–1967, 2022, doi: 10.1109/ACCESS.2021.3139339.
- [20] G. Reali and M. Femminella, "Two-layer network caching for different service requirements," *Future Internet*, vol. 13, no. 4, 2021, doi: 10.3390/fi13040085.
- [21] S. S. Musa, M. Zennaro, M. Libsies, and E. Pietrosemoli, "Mobility-aware proactive edge caching optimization scheme in information-centric IoV networks," *Sensors*, vol. 22, no. 4, 2022, doi: 10.3390/s22041387.
- [22] K. Delvadia, N. Dutta, and R. Jadeja, "CCJRF-ICN: a novel mechanism for coadjutant caching joint request forwarding in information centric networks," *IEEE Access*, vol. 9, pp. 84134–84155, 2021, doi: 10.1109/ACCESS.2021.3087558.
- [23] Y. Gui and Y. Chen, "A cache placement strategy based on entropy weighting method and TOPSIS in named data networking," *IEEE Access*, vol. 9, pp. 56240–56252, 2021, doi: 10.1109/ACCESS.2021.3071427.
- [24] J. Wu, J. Zhang, and Y. Ji, "DCEC: D2D-enabled cost-aware cooperative caching in MEC networks," *Electronics*, vol. 12, no. 9, Apr. 2023, doi: 10.3390/electronics12091974.
- [25] T. Zhou, P. Sun, and R. Han, "An active path-associated cache scheme for mobile scenes," *Future Internet*, vol. 14, no. 2, 2022, doi: 10.3390/fi14020033.
- [26] T. Lähderanta *et al.*, "Edge computing server placement with capacitated location allocation," *Journal of Parallel and Distributed Computing*, vol. 153, pp. 130–149, 2021, doi: 10.1016/j.jpdc.2021.03.007.
- [27] Y. Li, J. Wang, and R. Han, "PB-NCC: a popularity-based caching strategy with number-of-copies control in information-centric networks," *Applied Sciences*, vol. 12, no. 2, 2022, doi: 10.3390/app12020653.
- [28] H. Liu and R. Han, "A hierarchical cache size allocation scheme based on content dissemination in information-centric networks," *Future Internet*, vol. 13, no. 5, 2021, doi: 10.3390/fi13050131.
- [29] E. T. D. Silva, J. M. H. de Macedo, and A. L. D. Costa, "NDN content store and caching policies: performance evaluation," *Computers*, vol. 11, no. 3, 2022, doi: 10.3390/computers11030037.
- [30] Z. Sun *et al.*, "DNS request log analysis of universities in Shanghai: A CDN service provider's perspective," *Information*, vol. 13, no. 11, 2022, doi: 10.3390/info13110542.
- [31] G. Bao and P. Guo, "Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges," *Journal of Cloud Computing*, vol. 11, no. 1, 2022, doi: 10.1186/s13677-022-00377-4.
- [32] M. W. Kang and Y. W. Chung, "Content caching based on popularity and priority of content using seq2seq LSTM in ICN," *IEEE Access*, vol. 11, pp. 16831–16842, 2023, doi: 10.1109/ACCESS.2023.3245803.
- [33] J. D. M. Liborio, C. Melo, and M. Silva, "Internet video delivery improved by super-resolution with GAN," *Future Internet*, vol. 14, no. 12, 2022, doi: 10.3390/fi14120364.
- [34] F. Rau *et al.*, "A novel traffic prediction method using machine learning for energy efficiency in service provider networks," *Sensors*, vol. 23, no. 11, 2023, doi: 10.3390/s23114997.
- [35] J. K. P. Seng, K. L. M. Ang, E. Peter, and A. Mmonyi, "Artificial intelligence (AI) and machine learning for multimedia and edge information processing," *Electronics*, vol. 11, no. 14, 2022, doi: 10.3390/electronics11142239.
- [36] R. Shi, Q. Fan, S. Fu, X. Zhang, X. Li, and M. Chen, "COCAM: a cooperative video edge caching and multicasting approach based on multi-agent deep reinforcement learning in multi-clouds environment," *Journal of Cloud Computing*, vol. 12, no. 1, 2023, doi: 10.1186/s13677-023-00510-x.
- [37] X. Tang, F. Chen, and Y. He, "Intelligent video streaming at network edge: an attention-based multiagent reinforcement learning solution," *Future Internet*, vol. 15, no. 7, 2023, doi: 10.3390/fi15070234.
- [38] C. Wu, S. Shi, S. Gu, L. Zhang, and X. Gu, "Deep reinforcement learning-based content placement and trajectory design in urban cache-enabled UAV networks," *Wirel. Commun. Mob. Comput.*, vol. 2020, pp. 1–11, 2020, doi: 10.1155/2020/8842694.
- [39] 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, 2023, doi: 10.3390/math11051247.
- [40] Y. Liu, D. Lu, G. Zhang, J. Tian, and W. Xu, "Q-learning based content placement method for dynamic cloud content delivery networks," *IEEE Access*, vol. 7, pp. 66384–66394, 2019, doi: 10.1109/access.2019.2917564.
- [41] P. Dharmapal, Channakrishnaraju, and C. B. Krishnamurthy, "A novel cost-based replica server placement for optimal service quality in cloud-based content delivery network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 5, pp. 5588–5598, 2023, doi: 10.11591/ijece.v13i5.pp5588-5598.





- [42] B. K. Chethan, M. Siddappa, and H. S. Jayanna, "Trust correlation of mobile agent nodes with a regular node in a Adhoc network using decision-making strategy," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 2, pp. 1561-1569, 2020, doi: 10.11591/ijece.v10i2.pp1561-1569.
- [43] B. K. Chethan, M. Siddappa, and H. S. Jayanna, "Novel framework using dynamic passphrase towards secure and energy-efficient communication in MANET," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 2, pp. 1552-1560, 2020, doi: 10.11591/ijece.v10i2.pp1552-1560.
- [44] D. Priyanka, Channakrishnaraju, and B. K. Chethan, "Insights on effectiveness towards research approaches deployed in content delivery network," in *Software Engineering Perspectives in Systems*, Cham: Springer International Publishing, 2022, pp. 224-243, doi: 10.1007/978-3-031-09070-7_20.

BIOGRAPHIES OF AUTHORS



Priyanka Dharmapal     has 10 years of teaching experience for UG and PG courses in Computer Science and Engineering and is presently doing her research in the Department of Computer Science and Engineering, Sri Siddhartha Academy of Higher Education. She obtained B.E. degree from Visveswaraiah Technological University in the year 2009 and PG in Computer Science and Engineering in the year 2012 from Visveswaraiah Technological University. Her research interests are in the areas of content delivery networks, network security, mobile computing, and cloud computing. She is currently pursuing doctoral degree in Sri Siddhartha Academy of Higher Education, Tumkur, Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumkur. She can be contacted at email: prsh2019@gmail.com.



Channakrishnaraju     has 26 years of teaching experience for UG and PG courses in Computer Science and Engineering, and is presently working as professor in the Department of Computer Science and Engineering, in Sri Siddhartha Institute of Technology, Tumkur. He obtained B.E. from Bangalore University in the year 1995 and PG in software systems in the year 2000 from BITS Pilani and doctoral degree in Sri Siddhartha Academy of higher Education, Tumkur. His research interests are in the areas of wireless sensor networks, network security, and artificial intelligence. He can be contacted at email: rajuck@ssit.edu.in.