Design of a Novel Architecture for Cost-Effective Cloud-based Content Delivery Network

Suman Jayakumar[1], Prakash .S[2], C.B Akki[3]
Research Scholar, Department of CSE, VTU, Belgaum, India[1]
Professor and Executive Director, University Institute of Engineering, Chandigarh University[2]
Professor and Registrar, Department of Computer Science and Engineering[3]
Indian Institute of Information Technology (IIIT) Dharwad, Dharwad, India[3]

***Abstract*—Content Delivery Network (CDN) offers faster transmission of massive content from content providers to users using servers that are distributed geographically to offer seamless relay of service. However, conventional CDN is not capable of catering to the larger scope of demand for data delivery, and hence cloud-based CDN evolves as a solution. In a real-world scenario, each requested content has different popularity for different users. The problem arises with deciding which content objects should be placed in each content server to minimize delivery delays and storage costs. A review of existing approaches in cloud-based CDN shows that yet the problem associated with content placement is not solved. In this regard, a precise strategy is required to select the contents objects to be placed in a content server to achieve higher efficiency without affecting the CCDN performance. Therefore, the proposed system introduces a novel architecture that addresses this practical problem of content placement. The study considers placement problem as optimization problem with the ultimate purpose of maximizing the user content requests served and reducing the overall cost associated with content and data delivery. With an inclusion of a bucket-based concept for cache proxy and content provider, a novel topology is constructed where an optimal algorithm for placement of content is implemented using matrix operation of row reduction and column reduction. Simulation outcome shows that the proposed system excels better performance in contrast to the existing content placement strategy for cloud-based CDN.**

Keywords*—Content delivery network; content placement; cloud; optimization; data delivery; cost

I. INTRODUCTION

In the area of Content Delivery Network (CDN), the prime target is to offer a seamless relay of data and services associated with the delivery of contents by the content provider to a destined user [1]. There is various research being carried out towards this purpose while it was seen that it is challenging to offer this service of content delivery to large scale deployment regions by the conventional CDN [2]. Apart from constructing an appropriate CDN system and carrying out explicit maintenance of the distributed storage, followed by the delivery of appropriate content is heavily expensive from content providers' perspective. There is always a dependency of appropriate resources to perform maintenance of such servers [3]. To sort out this problem, the content providers are now seeking an alternative option of hosting the CDN over the cloud environment with much cost-effective solution [4]. Adopting the cloud environment is feasible to offer on-demand delivery of an appropriate content in much reduced time and at a cheaper cost [5]. At present, data are evolving exponentially with respect to size and complexity, while processing such data is now feasible in cloud computing owing to its capability to offer distributed storage and analytical options more effectively [6]. Existing trends of research also showcase that cloud-based CDN has become a pivotal topic when it comes to content delivery [7]. It is also noticed that the majority of the research work is focused on performing optimization approaches towards solving the problems associated with effective placement of contents in the presence of low usage of resources [8]. Basically, the term resource in this domain of discussion pertains to the quantity of storage that is demanded to be used by the content providers for the Cloud. Out of this, the essential problem is to find out the mechanism of positioning the informative contents over the incorrect server location [9]. The idea is to accomplish the optimal cost of the content delivery system as well as to ensure the minimal consumption of cumulative cloud storage [10]. Irrespective of the availability of various forms of literature towards improving this issue, it is found that the majority of the existing approaches emphasize a specific set of problems with narrowed usage of parameters [11]. However, problems associated with optimizing the storage as the resource are not much addressed in the existing solution. Existing approaches also don't offer a discussion of the inclusion of any user or computing devices and its related connection with the cloud terminals. Therefore, there is a need to carry out an investigation in order to offer a cost-effective solution in terms of modelling content placement approach with a target to offer a higher degree of performance of content delivery in cloud-based CDN. It is also necessary to ensure that the modelling is carried out considering practical constraints that are normally connected to the incoming and outgoing stream of data. Therefore, this paper presents a novel architecture in the form of a framework that is meant for cloud-based CDN with a single target of achieving optimal cost of allocation of resources. The core goal of this study is to ensure optimal performance achievement. The significant contribution of this paper is highlighted as follows:

- A bucket concept is considered which have a caching proxy and content server, with specific storage capacity.
- A novel topology construction is performed using graph theory for the bucket placement that keeps the content server close to the users for faster content access and

cost-effective task allocation under peak traffic conditions. This not only reduces request latency, but also balances the load between content servers.

- The study utilizes node centrality and computes sparsity towards analyzing higher probability of the request and determining efficient localization of content servers hosted overcloud.
- Content placement is considered as optimal assignment problem, which is solved using an explicit function constructed based on the matrix operation with minimum cost.

The organization of the proposed manuscript is as follows: Section II discusses the explicit problem, and their corresponding solution evolved in present times. Briefing of identified issues in existing solutions towards content placement is carried out in Section III. The highlights of the proposed architecture are carried out in Section IV, followed by an elaborated discussion of adopted research methodology in Section V. Discussion of obtained simulation outcome is done in Section VI, while a conclusive summary of paper with respect to its contribution is carried out in Section VII.

II. Related Work

At present, there have been various works being carried out towards the content-delivery network. Existing approaches have addressed various forms of problems associated with the content delivery network. The most recent work of Qazi et al. [12] has addressed the problem associated with unnecessary caching, leading to cost maximization of various network resources in a content centric network. The work has introduced an optimization concept that minimizes network resources targeting to control the latency and channel capacity. Problems associated with excessive channel capacity usage are also one topic of investigation in existing schemes, which affects the content delivery process over cloud-hosted applications. Research in this direction has been carried out by Khabbiza et al. [13], where the case study of multimedia streaming has been considered. According to this solution, the traffic is directed towards the adjacent node instead of the central server, thereby controlling the servers' load.

The advanced variant of cloud usage, i.e., fog computing, was also used to enhance the content delivery network's operation. The problem associated with cost connected with the content server's placement has been discussed by Liu et al. [14]. According to the author, the existing scheme is not capable of better decision-making considering global dynamics. The authors have used the Q-learning approach to facilitate a significant decision for routing operation over a tree structure. The model performs a selection of paths based on the low cost associated with it for effective content delivery. However, this approach can still not offer much information about the topology, which will affect any form of the pricing scheme.

Moreover, such an approach is not suitable for small-scale content providers. This absence of topological information is discussed by Duan et al. [15] has used a software-defined network where the infrastructure provider hosts the cache servers. This strategy maintains a balance between the content provider and infrastructure provider.

Further work towards content placement is carried out by Qu et al. [16], where problems associated with backhaul congestion are addressed. The study presented the solution to reduce the delay attribute associated with content delivery where mixed-integer linear programming has been used. Apart from content placement, existing studies were also carried out towards virtual network function, a part of the content delivery system. This completely depends upon the resource availability and its quantity. As per the discussion stated by Benkacem et al. [17], this problem is reported to be solved using their mathematical approach for cost minimization and upgrading quality of experience. The work carried out by Alghamdi et al. [18] has addressed the problem associated with the availability of content by using an improved version of the optimized link-state routing protocol. The study has a joint implementation of caching based on popularity and routing scheme over a cloud-based content delivery network. Similar problems of dedicated transmission of contents have been addressed by Asheralieva and Niyato [19] using game theory to model stochastic network control. The study has also used the Lyapunov optimization approach that emphasizes mobile nodes' activity connected with the operator. Another study carried out by Bosunia, and Jeong [20] addresses the challenges associated with the growing mobile internet market that affects seamless content delivery. The study has presented the usage of content-centric networking to carry out content delivery in the presence of a converged network. A case study of heterogeneous networks with radio access over Cloud is also seen in the literature concerning content delivery network investigation. The importance of using both qualities of wireless channels and their respective connection with the mobile station plays a significant role in improving the content delivery network's performance. The work carried out by Liu et al. [21] has addressed the problem associated with increasing the system's utilization using the belief propagation method. The study also presents a solution towards interference among the cells to resist the mobile station's overload at the remote radio unit. However, the adoption of a radio access network offers a significant issue over the delivery and caching of the contents and the capacity of processing. This problem has been considered in the work of Wang et al. [22], where a zone-based approach has been used for content caching cooperatively. The study has used a heuristic-based cooperation policy to better availability, and the transmission of more massive content is possible.

The existing content delivery network uses cloud radio access to support transmission in a faster network like 5G. However, the conventional caching principle offers degradation in network traffic. This problem is solved in a unique study carried out by Lau et al. [23] that has used content distribution based on humans' mobility patterns—the study aimed for the spatial allocation of radio resources and targets for resource efficiency. The resource provision methods always challenge balancing the war between the under and over-provisioning to handle the trade-off between an uncertain pattern of the user demand and their level of experiences as feedback. The authors, Haghighi et al., 2018, have designed an

optimization model for the resource assignment using Markov principles suitable for the C-CDN[24][25].The problems associated with responsive factor in content-centric delivery are addressed in the work of Sinky et al. [26], which discusses the importance of using multiple cloudlets with contents and caching policies in a heterogeneous network.

Apart from the content mentioned above, placement approaches and existing schemes have reported various alternative schemes for content placement viz. Approaches for web content delivery for internet architecture of future (Siracusano et al. [27]), caching using the push-based strategy (Fan et al. [28]), hierarchical modelling (Papagianni et al. [29]), preemptive hierarchical approach (Salahuddin et al. [30]), bee-colony optimization algorithm (Ghalehtaki et al. [31]), cache placement approach (Ha and Kim [32]), and network-slicing (Retal et al. [33]). The next section discusses the problems associated with the existing content placement approaches, followed by a proposed solution to thwart this problem.

III. Problem Statement

After reviewing the existing approaches, the following are the issues explored:

- Less emphasis on Quality of Experience (QoE): Existing studies have implemented a sophisticated mechanism of different types to address the content placement problem. However, these mechanisms do not consider the user's computing and communication device to exercise delivery services. Existing approaches do not offer much scalability concerning QoE regarding peak traffic conditions in both access and core networks.
- Impediment towards heavy file delivery: It is well known that CDN is mainly meant to handle more massive data delivery. However, theories in the paper differ from real-time demands owing to a lack of benchmarking approach. To deal with communication channels with inferior Quality of Service, the transmission rate is reduced to work at a specific limit. However, such approaches are not applicable when it comes to stream real-time multimedia content. Another practical problem is that users are usually considered connected with only one technology of access and hence, the issues shoot up.
- Scheduling of Resources: A specific amount of resources are required to carry out the problem of content placement. Management of resources can be optimized by using resource allocation for controlling cost factors. However, this is not a simplified form of the task as it demands precise information on the topology of the nodes and information by the content providers. Unfortunately, it is not there in the existing system and without which it cannot be deployed over a much complex environment of Cloud.
- Caching Related Issues: While transmitting more massive files or streaming content over the internet, it is essential to reduce the latency/delay. It is also known that the conventional content delivery network offers varied caching algorithms; however, it is not much considered that this algorithm must be run over edges of the delivery network. This increases the complexity of multi-fold when it comes to cloud computing-based content delivery network. As the scale of deploying content servers extends exponentially when hosted over Cloud, developing an algorithm for caching management over edges is highly a complicated task.
- Ambiguity in cost modelling: There have been various models in existing approaches where cost-related modelling has been carried out. However, considered parameters in cost modelling and its relationship with the content placement problem have not yet been built. At present, the term cost is associated with the financial terms connected with the deployment of services. For effective cost modelling, it is necessary to consider all latent parameters that are the indicators of resource allocation and maintain a good streamline with user and content provider. Such consideration is missing in existing approaches.

All the problems mentioned above are addressed in the proposed system discussed in the next section.

IV. Proposed System

The proposed study's primary goal is to develop an analytical framework that could carry out a content placement in a cloud-based content delivery network. The secondary goal is to develop a cost-effective allocation of tasks from the user to the content servers over peak traffic conditions.

Fig. 1 highlights the proposed system's architecture, which shows that the model initiates by taking an input of several requests, bucket formation, and area of deployment. Each bucket is considered to possess a caching proxy and content server with an explicit storage capacity allocated for both of them. The proposed study uses graph modelling for constructing the topology by applying a directional graph where each vertex corresponds to a bucket. Geographical modelling is further carried out considering the domain-based CDN nodes where an orthogonal and symmetric directionality of the node placement is carried out. This directionality attribute plays a vital role in searching for optimal content placement over the buckets. The next step is to construct a proximity priority module where the node centrality is incorporated to understand the node's significance, considering the probability of the request. The traffic stream is judged based on the weight factor associated with the structure of in-degree and out degree nodes. The sparsity computation is carried out to find the best links out of many that lead to the efficient location of content servers hosted overcloud. After this process is carried out, the proximity priority model is executed further, followed by matrix-based operation development to carry out optimal placement using a function for performing allocation and cost estimation. This is based on the allocation of tasks to find an efficient node for content placement in cloud-based CDN. This process's outcome leads to the allocation of the task and the estimated cost of the evaluation. The next section discusses the adopted methodology.

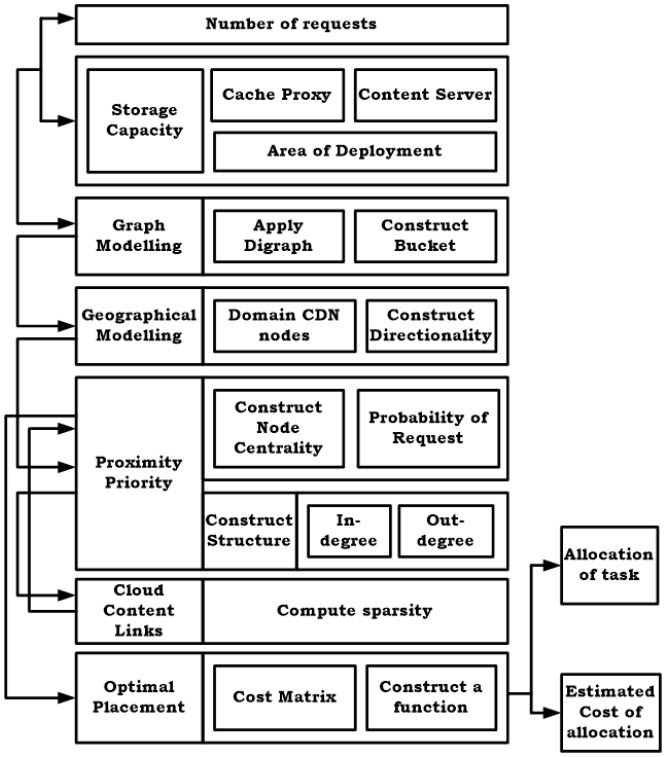

Fig. 1. Architecture of Proposed System.

V. Research Methodology

The proposed research work's prime objective is to ensure that end user is facilitated with a higher degree of service quality using the cloud-based content delivery network's proposed model. The proposed system's solution is based on the appropriate placement of the distributed cloud environment's contents. The idea of the proposed logic is that if the content placement is done accurately in a shared manner, it is feasible to minimize the cost of content maintenance over a server. The proposed system applies a novel optimization approach that can facilitate a better form of content server update and dynamic updating of contents replication. Hence, an analytical research methodology is constructed for this purpose, which can finally ensure a better form of content delivery with a controlled reduction in latency in the content transmission process. This section offers details about the comprehensive process that is adopted for appropriate, cost-effective content placement.

A. Topology for Content Placement

According to the novel concept of content placement, the prime logic is to ensure a better symmetry in the node's geographical distribution. In existing times, the content placement is carried out based on the location of users, which is highly a dynamic event. Hence, toggling the location of contents based on user location (mobility) will demand more cost consumption owing to the non-symmetrical locus of the content server. Hence, better symmetric localization of content servers will lead to better delivery performance and better service availability. Therefore, the proposed topology considers the asymmetric distribution of buckets B, as shown in Fig. 2. Buckets are the nodes, which bear all the information and are directly synched with the cache proxy and content server. The topology also considers a centralized server CN which is equidistant from all the buckets.

The study considers a test region R, which is further classified into *i* number of regions where the placement of the buckets is carried out. It can be empirically expressed as,

$R_i = \{R_1, R_2, \ldots..R_i\}$

Each region is assumed to consist of buckets B, which will mean that $B=B_1, B_2, \ldots.B_N$, where N=i. It is considered that each bucket B has cache proxy C1 and content server C2, which will mean that,

$B_N=\{(C_{1N})_{ij} \mid (C_{2N})_{ik}\}$

In the above expression, *N* represents the total number of buckets, *i* represents several regions, and *j* represents the maximum number of the proxy server while *k* represents the content server's highest capacity. The proposed concept of placement of content on multi-cloud architecture takes '**N**' buckets(**B**) in geographically distributed Data Center (shown as CN in Fig. 1) acts both as Cache-Proxy as well as Content Server. Both of them are interconnected bi-directionally to each other under with a weight(**w**). The weight (w) is considered a set of the properties {caching, cost, latency, dynamicity/ambiguity, and interoperability}. The context of the N=4 as $\{B_1, B_2, B_3, B_4\}$ with the connectivity possibility of pair of :$\{[(B_1\text{-}B_2),(B_1\text{-}B_3),(B_1\text{-}B_4)],[(B_2\text{-}B_4),(B_2\text{-}B_1)],[(B_3\text{-}B_2],[(B_4\text{-}B_3),(B_4\text{-}B_1),(B_4\text{-}B_2)]\}$as shown in the Fig. 2.

The respective weight for different capacities of the connection network, W= ['B1/3' 'B1/3' 'B1/3' 'B2/2' 'B2/2' 'B3' 'B4/3' 'B4/3' 'B4/3'}]. The proposed system performs modeling using graphical constructs, i.e., G(V, E), where the vertices 'V' represents the bucket *B*→C1/C2 and the edges' *E*' represents connecting links among the respective nodes, is represented in Fig. 3. These associating weights mechanism assists in the proper identification of appropriate links connecting to various buckets that have reduced cost factors involved.

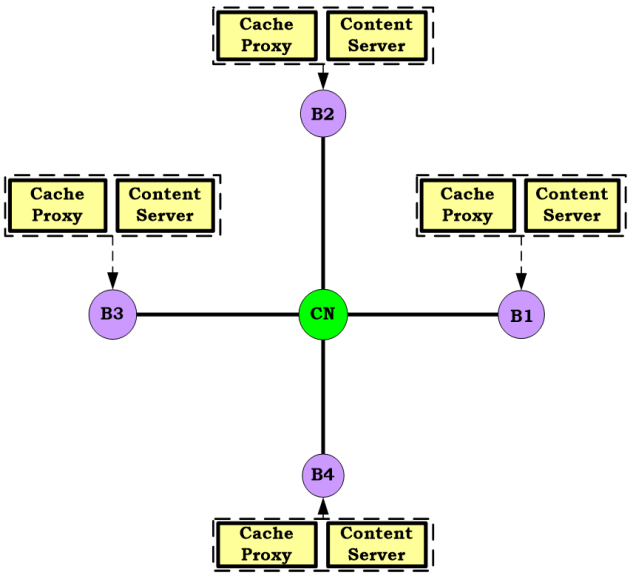

Fig. 2. Proposed Topology of Content Placement.

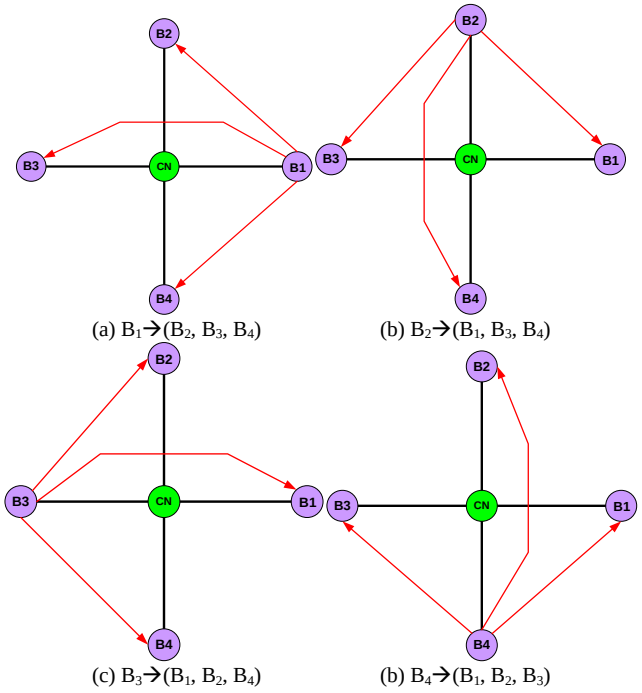

Fig. 3. Connectivity among the Buckets.

All these forms the shape of the matrix, and the entries of the adjacency matrix take either the complete or a sparse of the numeric data with the input elements of the connections of the network as edges among the nodes as a non-zero element. The value represents the weight of the edge connection and if it is a logical adjacency that results in an unweighted graph. If there are non-zero values in the diagonal representing a self-loop, the nodes are connected to themselves with an edge.

The standard topology consists of bucket placement in 4 different directions, which are the right angles. However, for better connectivity, more symmetrical placement is required for effective data transmission. The content bucket modeling for the consistency in the geographical distribution to achieve balanced latency is modeled as locational mapping with the placement of the proxy/content buckets in the location of the L={North (N), South (S), East (E), West (W), North-East (NE), North-West (NW), South-East (SE), South-West (SW)} as shown the Fig. 4.

Fig. 5 highlights the 8 nodes placement of the bucket in a highly symmetric fashion. One of the advantages of this topology is that it offers complete supportability of sharing content in any of the buckets during the dynamic traffic scenario. Hence, the proposed topology is supportive of users with dynamic mobility. One case study of the data center CN's connectivity with all the 8 respective bucket positions is shown in Fig. 6. A graphical direction is given from the data center to the respective buckets.

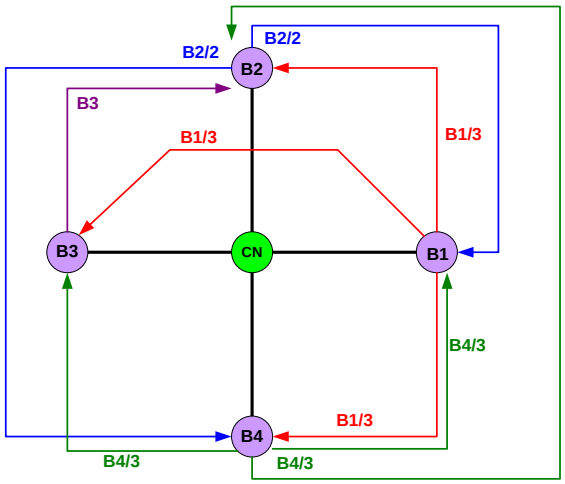

Fig. 4. Constructed Graph with Weights.

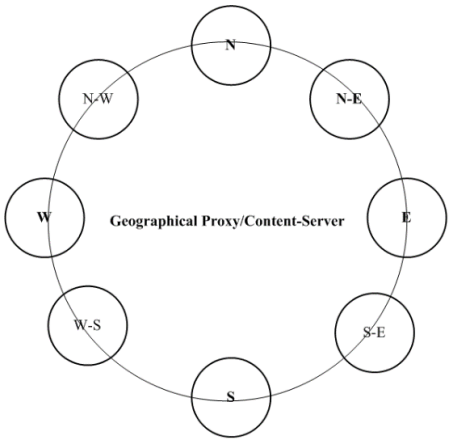

Fig. 5. Geographical Distribution of Cache/Content Server.

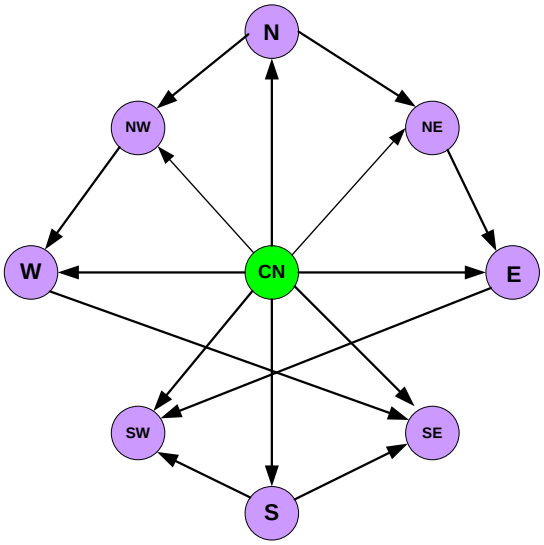

Fig. 6. Directed Graph for Optimal Connectivity.

The information associated with the degrees is captured from each node, assuming that the datacenter hosts a node with a domain http://cdnserviceprovider.com/Central. Therefore, the extracted features will be:

- CN: node http://cdnserviceprovider.com/Central
  - In-degree: 0, Out-degree: 8
- NW: node http://cdnserviceprovider.com/NW
  - In-degree: 2, Out-degree: 1
- NE: node http://cdnserviceprovider.com/NE
  - In-degree: 2, Out-degree: 1
- SE: node http://cdnserviceprovider.com/SE
  - In-degree: 3, Out-degree: 0
- SW: node http://cdnserviceprovider.com/SW
  - In-degree: 3, Out-degree: 0
- W: node http://cdnserviceprovider.com/West
  - In-degree: 2, Out-degree: 1
- N:node http://cdnserviceprovider.com/North
  - In-degree: 1, Out-degree: 2
- S: node http://cdnserviceprovider.com/S
  - In-degree: 1, Out-degree: 2
- E: node http://cdnserviceprovider.com/East
  - In-degree: 2, Out-degree: 1

All the above features are used for computing the cost factor involved in the proposed cloud-based content delivery network. The proposed system enables the connectivity with the cache proxy with the nearest bucket available in the topology. Apart from this, all the buckets are connected and synced with each other, as shown in Fig. 5. This interconnection of the graph edges facilitates the proposed cloud-based content delivery network to carry out the delivery of the contents in a dynamic pattern. The study also considers that datacenter CN consists of all the source content, and it is also directly linked with all the buckets in the proposed cloud-based content delivery network. In the conventional data delivery mechanism, the user-based contents required to be shared with users are duplicated over various variants of the cache proxies. However, this mechanism calls for a significant imbalance between content servers and the data transmission cost as there are a maximized number of increasing replicated files. The novelty of the proposed system is that it can eliminate the duplicated file from cache proxies and save them over different buckets in their respective buckets, unlike existing approaches. This mechanism can significantly minimize the quantity of the replicated file and potentially control the cost of content server placement. The algorithm developed for optimal cost computation is as follows:

Algorithm for Optimal Cost Computation

Input:s, t, B, N, D
Output:c
Start
1. Define s, t (B_N), D
2. Apply graph, $G(s, t)^D$
3. pr→f_1(G)
4. struct G=[*pr*, *id*, *od*]
5. Apply f_2(G)
6. obtain $G_{sub}(G, sig_f>p)$
7. *c*→Apply f_3(*pr*)
End

The above algorithm is responsible for computing the optimal cost in the proposed cloud-based content delivery network, which takes the input of *s* (source), *t* (destination), *B* (bucket), *N* (number of buckets), and *D* (domain hosted) that after processing yields an outcome of *c* (cost). The proposed algorithm's initial step is defining the particular N number of buckets concerning source and destination (Line-1). A digraph structure *G*is used for this purpose in order to give a shape of connected buckets in topology (Line-2). The next part of the implementation is about computing the centrality of graph G (Line-3). It is computed by dividing each bucket's value by one a smaller number of nodes that essentially represents the number of edges connected to the buckets. This operation results in a priority factor *pr* (Line-3). Once the priority factor is computed, the proposed system constructs a graphical structure *G*, which consists of priority factor *pr*, in-degree *id*, and outdegree *od* (Line-4). Finally, the proposed system constructs a sparsity pattern for the given buckets as variable test cases of different placement of the buckets. This is carried out to testify the sustainability of the algorithm towards lower latency over the various position of the bucket in a defined area. This operation is further followed by applying a digraph structure over *G* (Line-5). Finally, the sub-graph G_{sub} is obtained, and only those buckets are selected, whose priority factor is found to be statistically significant (>0.005) (Line-6). Finally, the algorithm constructs an explicit function $f_3(x)$, which is responsible for obtaining the optimal cost of the content delivery placement. This mechanism is carried out using matrix-based operation where the buckets and their region-specific information are considered priority factor *pr*. The formation of this function is carried out in the following manner:

Problem Formulation: A case study using matrix-based operation is considered to understand the proposed study's problem formulation. As the proposed system uses information associated with buckets and priority factors associated with the content placement, it is easier to represent this fact using a matrix. The proposed system constructs a squared matrix of n x n, which exhibits the associated cost for all n buckets to obtain buckets' optimal placement concerning the data center. The complete goal is to reduce the overall cost. As one bucket can be utilized for carrying out one set of job processing and all jobs should be allocated uniquely to each bucket in its respective position. Therefore, this allocation will formulate an independent set of matrix M as below:

$$M(i,j)= \begin{array}{c} \\ p \\ q \\ r \\ s \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{bmatrix} \end{array}$$

In such a situation, a random allocation is used where the bucket *p* is allocated job *b*, bucket *q* is allocated job *d*, and it goes on following such a pattern. The cost factor involved in such a case of job allocation will be 23, while the problem will be to search for allocating much minimal cost value. The constraint will be to perform all the individual allocation has to be relatively discrete and different overall the given rows and columns in matrix *M*. In order to find a solution to this problem, a brute-force approach can be used that can lead to a yield of a different independent set of *M* matrix. It should result in overall cost for all the content placements and explore the smaller set. However, it has to be noted that the complete computational complexity factor is associated with the size of the squared matrix and its associated allocation of buckets. It will mean that n choices will be the primary allocation while (*n*-1) will be the second allocation that finally leads to factorial n feasible sets of allocation. Hence, a significant complexity associated with computational run time is associated with this process. While selecting the assignment, the respective rows and columns must be eliminated; therefore, a problem will be to find the optimality of this reduction process. This leads to higher computational complexity. In this regard, the study presents optimal strategy based on cost matrix formulation as shown in Fig. 7.

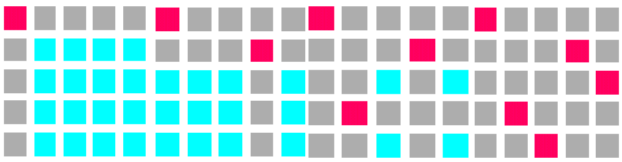

Fig. 7. Visual Representation of Cost Matrix.

Solution: The solution to this problem is carried out by constructing a function f3(x). The matrix operation carried out in this function is showcased in Fig. 8. There are six steps of operation that are carried out to solve this problem viz. i) a non-squared matrix of n x m is constructed where the elements depict cost factor associated with the allocation of one of n bucket to one unique m job. The matrix M is rotated in such a way that there is always a minimum number of rows and columns and considers k→argmin(n, m). ii) the next step is to search for the minimal element over all the rows in the M matrix and subtract it from all the elements present in the row, iii) the consecutive step is to look for the presence of zero in the outcome matrix. The absence of any zero in the row and the column calls for flagging that zero. It is iterated for all the matrix elements. iv) all the zero elements that are starred are covered concerning the column of its position. If K number of columns is covered, then the flagged zeros represent a cumulative set of non-repeating allocation. In such a case, the operation is completed, or else the next step is processed. v) All the zero elements that have not been covered up are identified and then are primed. In case of such primed zero, absence of any flagged zero over the row, the function performs next step, or else this particular row elements are covered, and all the columns consisting of flagged zero are uncovered. This process is repeated until and unless all the zero elements are covered. The minimal value of the uncovered element is saved, vi) a series of alternating flagged zero elements and primed zero elements is constructed, vii) the value of the result obtained in step 5 is added to all the elements over rows with covered elements, followed by subtracting it from all the values of a column that are uncovered. Without performing any alteration over covered lines, primes, and flags, the process returns to the 5th step. viii) The final step indicates the pairs of values to be allocated, considering the flagged zero elements' position over the M cost matrix. Therefore, if M(i,j) is flagged zero, then the values connected with this ith row are allocated to the jth column values.

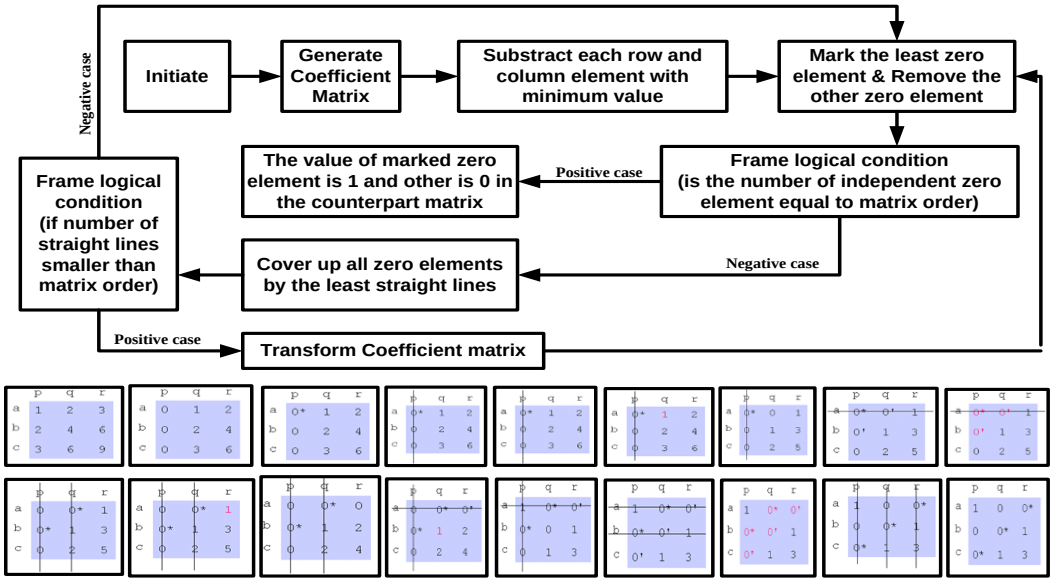

Fig. 8. Matrix Operation Carried Out in Function $f_3(x)$.

The function f3(x) outcome is an optimal allocation with highly reduced cost based on the input argument cost matrix. The complete operation of $f_3(x)$ is discussed concerning the proposed cloud-based CDN as follows:

Consider that there is Φ number of jobs required to be accomplished based on the query generated by the α number of computing devices of the user. Assuming that Φ≤ α, there is a possibility that any computing devices can be allocated in order to accomplish this task over a cloud environment, where each device incurs a cost in the form of resources as well as time to accomplish the task. Hence, the proposed cloud-based CDN system's objective function will be to carry out the complete task without the inclusion of maximized cost while performing a selection of the best resources and computing device for this purpose. Therefore, the objective function developed for this purpose is:

$$arg_{min} \sum_{i=1}^{\emptyset} \sum_{j=1}^{\alpha} m_{ij}.\beta_{ij} | \sum_{i=1}^{\emptyset} \beta_{ij} = 1$$

$$0 \leq \sum_{j=1}^{\alpha} \beta_{ij} \leq 1, \beta_{ij} \in [0,1] \quad (1)$$

The expression (1) represents a cost matrix m_{ij} that essentially depicts the cost incurred by a computing device of user *i* to carry out the j^{th} task. The variable β represents a binary matrix whose value is considered 1 if a specific ith computing device of the user is allocated a specific task of j or else it takes 0. The proposed algorithm of allocating the task offers by considering a bipartite graphical structure G=(V, E) where the vertices *V* is a union of all source node *s* and destination node *d*. All the vertex of the bipartite graph is labeled γ, and the condition for labeling is that all the labels are anticipated to map with the matching constraint in the distributed cloud environment. The system considers the possible labelling as a function that satisfies a criterion of γ(x)+ γ(y)≥λ(x,y), where the variable λ represents the weight factor. The study considers that avertex α is only tagged as matching if this vertex is a part of the main vertex α_m. The subgraph of G is represented by G_λ that consists of information about the edges. The study considers this subgraph G_λ to be a spanning tree of main graph bipartite G. Further; it amalgamates the complete available vertices from the main graph G. The feasibility of the allocation is ensured in this process by ensuring the inclusion of only those communication links (edges) from the core bipartite matched graph. Considering that α' to be precisely matching with the spanning subgraph G_λ then α' is considered to match with core graph G with the highest weight exhibiting the better allocation process and minimal cost. The implementation intends to exhibit that α'is the only perfectly matching matrix where the weight computation for allocation is carried out as follows:

$$\lambda(\alpha) = \sum_{x,y\in\alpha} \lambda(x,y) \leq \sum_{x,y\in\alpha} \{g[\gamma(x,y)]\}$$

$$= \sum_x \gamma(x) + \sum_y \gamma(y) = \sum_{x,y\in\alpha} \{g[\gamma(x,y)]\}$$

$$= \sum_{x,y\in\alpha} \{\lambda(x,y)\}$$

$$= \lambda[\alpha'] \quad (2)$$

In the above mathematical expression (2), the variable *g* represents the summation operator while *α*' signifies the highest form of matching for cost-effective allocation in cloud-based CDN, and thereby a global optimization is ensured. It should be noted that the complete process of cost-effective allocation in cloud-based CDN is applied over a dense matrix, which signifies that the proposed system is capable of withstanding peak traffic conditions. The study finds that such operation is capable of transforming the row and column by harnessing the potential advantage of the proposed data structure that is crossed linked. This structure is utilized to reposit the data while the links are manipulated to forward the data from the server to the users. There is another reason for the efficiency of allocation in the proposed cloud-based CDN. The proposed system uses the function $f_3(x)$, where the difference between the weight is reformulated in the form of a structure matrix. The structured matrix is then subjected to classification concerning its complete columnar elements into a specific number of blocks uniformly. A similar operation of classification is also carried out for row-wise elements. All the sub-problems are now solved in parallel fashion in the proposed system, ensuring the capability to perform job query processing from an incoming stream of traffic. All the tasks that are found to be unique are then checked. This ensures that if the same task resides within multiple computing devices of a user, then only the task with higher profit is only accepted and considered for allocation. With sparsity property aid, the proposed system offers better efficiency of allocation of job cost-effectively irrespective of the choice of operation either in a row or in column-wise. Therefore, the proposed system performs content in cloud-based CDN by formulating a novel topology using a bipartite graph. The solution offered by the.

VI. Result Analysis

The proposed system's implementation is carried out in an analytical fashion, where MATLAB was used for scripting. The analysis is carried out considering the graphical structures where cost and latency play a significant contributory role. Table I highlights the cost associated with each bucket in different position of North (N), East (E), West (W), South (S), North East (NE), North West (NW), South East (SE), South West (SW), and central node of the data center (CN). A bipartite graph is initially constructed for topology creation. The study assumes the simulation parameters as 25 cache proxy with 3000 MB of maximum capacity for each, 10 content servers with 5000 MB of maximum capacity for each, 9 areas of deployment of buckets. Table I shows the position of the buckets vertical wise while its respective cost is highlighted horizontally. The study's complete implementation has been carried out considering 9 buckets that have possession of both cache proxy and content server. The selection of the simulation parameters is highly flexible, and it is considered in such a way that there is an assignment of reduced capacity for the proxy server in contrast to the content server in the bucket. This is carried out in order to map with the practical environment of cloud-based CDN. Table I highlights the instance of the cost matrix for all location ={ P/C-S-N (1), P/C-S-S (2), P/C-S-E (3), P/C-S-W (4), P/C-S-NE (5), P/C-S-NW (6), P/C-S-SE (7), P/C-S-SW(8), P/C-S-CN(9)}. Here, the variable P isan optimal position, and {S, N, W, E} represents 4 orthogonal directions

of south, north, west, and east for the considered topology. The variable Ct represents the cost of all 9 locations. The table highlights the nodes (buckets) under the different possible conditions of 4 orthogonal directions. Table II highlights the accomplished cost associated with all the nodes in 9 different positions in multiple nodes' directions. Table III represents the elements allocated in binary matrix β where the numerical value of 0 will represent zero task allocation, while 1 will represent the allocated task from the user device to the bucket. Table IV highlights the estimated cost for all the nodes in 9 different positions. The estimation of cost is carried out by the proposed algorithm explicitly by applying the function for optimal placement $f_3(x)$.

TABLE I. COST AS LATENCY ALGORITHM

	Ct-1	Ct-2	Ct-3	Ct-4	Ct-5	Ct-6	Ct-7	Ct-8	Ct-9
P/C-S-N (1)	Co-1	Co-2	Co-3	Co-4	Co-5	Co-6	Co-7	Co-8	Co-9
P/C-S-S (2)	Co-10	Co-11	Co-12	Co-13	Co-14	Co-15	Co-16	Co-17	Co-18
P/C-S-E (3)	Co-19	Co-20	Co-21	Co-22	Co-23	Co-24	Co-25	Co-26	Co-27
P/C-S-W (4)	Co-28	Co-29	Co-30	Co-31	Co-32	Co-33	Co-34	Co-35	Co-36
P/C-S-NE (5)	Co-37	Co-38	Co-39	Co-40	Co-41	Co-42	Co-43	Co-44	Co-45
P/C-S-NW (6)	Co-46	Co-47	Co-48	Co-49	Co-50	Co-51	Co-52	Co-53	Co-54
P/C-S-SE (7)	Co-55	Co-56	Co-57	Co-58	Co-59	Co-60	Co-61	Co-62	Co-63
P/C-S-SW(8)	Co-64	Co-65	Co-66	Co-67	Co-68	Co-69	Co-70	Co-71	Co-72
P/C-S-CN(9)	Co-73	Co-74	Co-75	Co-76	Co-77	Co-78	Co-79	Co-80	Co-81

TABLE II. ACCOMPLISHED COST

Position	**Ct-1**	**Ct-2**	**Ct-3**	**Ct-4**	**Ct-5**	**Ct-6**	**Ct-7**	**Ct-8**	**Ct-9**
P/C-S-N (1)	70	72	27	64	22	92	4	24	55
P/C-S-S (2)	64	97	15	96	67	0	18	92	43
P/C-S-E (3)	3	53	28	24	84	46	72	27	64
P/C-S-W (4)	7	33	44	68	34	42	47	77	65
P/C-S-NE (5)	32	11	53	29	78	46	15	19	68
P/C-S-NW (6)	53	61	46	67	68	77	34	29	64
P/C-S-SE (7)	65	78	88	70	1	32	61	9	95
P/C-S-SW(8)	41	42	52	7	60	78	19	58	21
P/C-S-CN(9)	82	9	94	25	39	47	74	68	71

TABLE III. ELEMENTS OF BINARY MATRIX

Position	**Ct-1**	**Ct-2**	**Ct-3**	**Ct-4**	**Ct-5**	**Ct-6**	**Ct-7**	**Ct-8**	**Ct-9**
P/C-S-N (1)	0	0	0	0	0	0	1	0	0
P/C-S-S (2)	0	0	0	0	0	1	0	0	0
P/C-S-E (3)	0	0	1	0	0	0	0	0	0
P/C-S-W (4)	1	0	0	0	0	0	0	0	0
P/C-S-NE (5)	0	1	0	0	0	0	0	0	0
P/C-S-NW (6)	0	0	0	0	0	0	0	1	0
P/C-S-SE (7)	0	0	0	0	1	0	0	0	0
P/C-S-SW(8)	0	0	0	0	0	0	0	0	1
P/C-S-CN(9)	0	0	0	1	0	0	0	0	0

TABLE IV. ESTIMATED COST

Position	**Bucket No**	**Cost**
P/C-S-N (1)	Ct-7	4
P/C-S-S (2)	Ct-6	0
P/C-S-E (3)	Ct-3	28
P/C-S-W (4)	Ct-1	7
P/C-S-NE (5)	Ct-2	11
P/C-S-NW (6)	Ct-8	29
P/C-S-SE (7)	Ct-5	1
P/C-S-SW (8)	Ct-9	21
P/C-S-CN (9)	Ct-4	25

At present, the content placement approaches are mainly of two types, i.e., deterministic and randomized. The deterministic approach selects a static topology and does not offer any changes in the user's due course of query processing. The randomized process involves selecting randomized buckets in due course of task processing query from the user. However, the proposed system is optimal in its approach as it can offer a dynamic update and dynamic change of topology by adapting the new cost associated with the link leading to the content matrix. All the calculations are carried out in a similar environment concerning increasing simulation rounds. Each simulation rounds are incorporated with random allocation of the task from the user towards the edge's bucket. The idea is to assess the performance of the proposed content placement approach concerning multiple performance parameters, e.g., Algorithm processing time, latency, cost of allocation, and the probability of resource allocation.

Fig. 9 shows that the proposed system offers a 20% improvement in faster processing than the existing one. The prime reason behind this is the faster update exchange within the topology, where it consumes less time to find the efficient bucket for content placement than the existing approach.

Fig. 10 highlights that the proposed system offers a significantly higher reduction in latency, approximately 60% compared to the existing system. A closer look at the graphical outcome shows no significant difference in randomized and deterministic approach much. The reason behind it is that, in the existing approach, the problem is solved from local problem space, which consumes more time and hence is not scalable for more massive and dense traffic conditions.

However, the proposed system focuses on the global problem space by constructing a crosslinked data structure with different buckers. This makes the process of algorithm execution faster and ensures that incoming queued jobs are faster processed. Each cost matrix keeps track of indegree and outdegree, which are always balanced if there is more indegree than outdegree, causing better control of latency performance.

Fig. 11 highlights that the proposed system's cost performance is approximately 40% better than the existing approach. The deterministic approach performs calculation of the efficient algorithms in advance and then fixed its topology to perform query processing towards buckets. This allocation's static nature often contradicts the allocated weight of the links with the incoming traffic, which causes an increased cost of allocation.

On the other hand, a randomized approach exhibits better performance in the prior set of simulation rounds; however, this approach requires additional effort to explore a greater number of network indicators for efficient content placement with increasing rounds. So, it is only slightly better than the deterministic approach. The proposed approach offers a better cost reduction capability as it can carry out all task processing and allocation using a parallel approach and hence same data structure is reused for parallel allocation of job over the buckets. This drastically reduces the cost associated with allocation in order to find a better content placement.

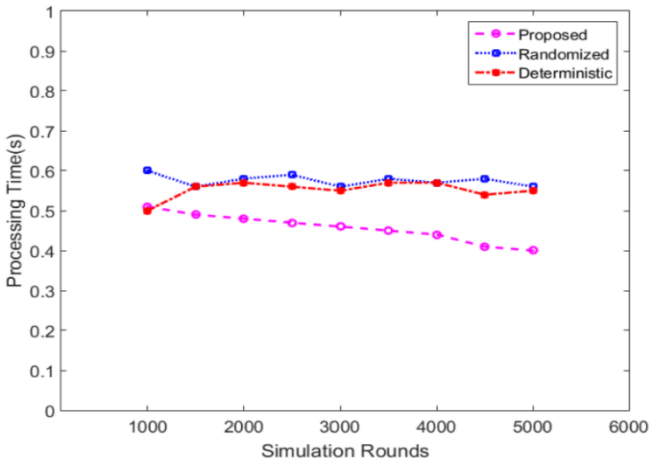

Fig. 9. Comparative Analysis of Processing Time.

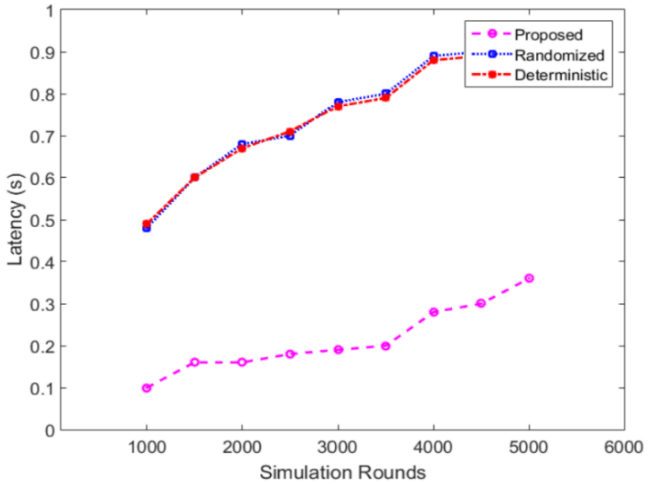

Fig. 10. Comparative Analysis of Latency.

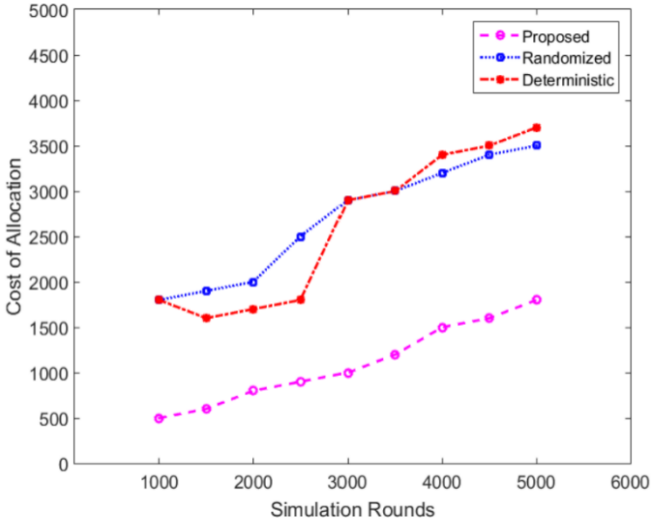

Fig. 11. Comparative Analysis of Cost.

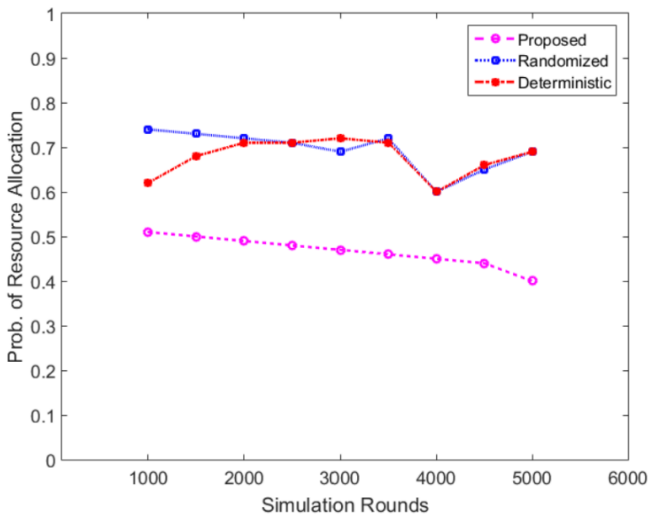

Fig. 12. Comparative Analysis of Allocation.

Fig. 12 highlights that the probability of resource allocation of the proposed system is approximately 35% better than the existing approach for a similar cause. Hence, based on this outcome, it can be said that the proposed system is capable of offering cost-effective content placement with reduced computational complexity and at par with meeting the demands of practical networks over a cloud environment.

VII. Conclusion

This paper has presented a novel framework where a unique architecture towards cost modelling is carried out for content placement. The overall contribution of this paper is as follows: i) the paper introduces a novel way of using cost matrix using crosslinking data structure which can minimize the dependencies of replica in order to optimize the cost of content delivery, ii) the study achieves faster processing time over a stream of continuous data (or request) from the user making it suitable for practical world application over Cloud-based CDN, iii) the applicability of proposed system will have a higher score of quality of experience as well as the quality of service due to the following reason: it offers reduced dependency of resources thereby making it resource optimized approach, it uses data sparsity and offers reduced allocation cost capable of processing multiple jobs at one instance. This makes the system support both parallel processing over a distributed storage environment over Cloud. The key novelty is that the scheme presented in this paper, jointly addresses multiple issues such as content server placement and optimal content placement on the content server to support maximum content request and content delivery with less delay and higher throughput. The proposed system can be viewed as a support system in CCDN to enhance the user experience. Our future work will be towards further optimizing the performance.

References

[1] B.Zolfaghari,G. Srivastava, S. Roy, H. R Nemati, "Content Delivery Networks: State of the Art, Trends, and Future Roadmap", ACM Computing SurveysVol. 53, No. 2Content Delivery Networks: State of the Art, Trends, and Future Roadmap, 2019.

[2] Salahuddin, M. A., Sahoo, J., Glitho, R., Elbiaze, H., &Ajib, W. (2017). A Survey on Content PlacementAlgorithms for Cloud-based Content Delivery Networks. IEEE Access: Practical Innovations, Open Solutions, 6, 91–114. doi:10.1109/ACCESS.2017.2754419.

[3] Silva, Fabrício A., Azzedine Boukerche, Thais RM Braga Silva, Linnyer B. Ruiz, Eduardo Cerqueira, and Antonio AF Loureiro. "Vehicular networks: A new challenge for content-delivery-based applications." ACM Computing Surveys (CSUR) 49, no. 1 (2016): 1-29.

[4] Yubao Zhang, Hao, Shuai, , Haining Wang, and Angelos Stavrou. "End-users get maneuvered: Empirical analysis of redirection hijacking in content delivery networks." In 27th {USENIX} Security Symposium ({USENIX} Security 18), pp. 1129-1145. 2018.

[5] S. Qabil, U. Waheed, S. M. Awan, Y. Mansoor and M. A. Khan, "A Survey on Emerging Integration of Cloud Computing and Internet of Things," 2019 International Conference on Information Science and Communication Technology (ICISCT), Karachi, Pakistan, 2019, pp. 1-7, doi: 10.1109/CISCT.2019.8777438.

[6] Banu, S. Sajitha, and S. R. Balasundaram. "Cost effective approaches for content placement in cloud CDN using dynamic content delivery model." International Journal of Cloud Applications and Computing (IJCAC) 8, no. 3 (2018): 78-117.

[7] Gkatzikis, Lazaros, Vasilis Sourlas, Carlo Fischione, Iordanis Koutsopoulos, and György Dán. "Clustered content replication for hierarchical content delivery networks." In 2015 IEEE International Conference on Communications (ICC), pp. 5872-5877. IEEE, 2015.

[8] R. W. L. Coutinho, A. Boukerche and A. A. F. Loureiro, "Design Guidelines for Information-Centric Connected and Autonomous Vehicles," in IEEE Communications Magazine, vol. 56, no. 10, pp. 85-91, OCTOBER 2018, doi: 10.1109/MCOM.2018.1800134.

[9] Gupta, R. K., Hada, R., & Sudhir, S. (2017). 2-Tiered Cloud based Content Delivery Network Architecture: An Efficient Load Balancing Approach for Video Streaming. IEEE International Conference on Signal Processing and Communication. doi:10.1109/CSPC.2017.8305885

[10] P. Osypanka and P. Nawrocki, "Resource Usage Cost Optimization in Cloud Computing Using Machine Learning," in IEEE Transactions on Cloud Computing, doi: 10.1109/TCC.2020.3015769.

[11] Aral, Atakan&Ovatman, Tolga. (2018). A Decentralized Replica Placement Algorithm for Edge Computing. IEEE Transactions on Network and Service Management. 1-1. 10.1109/TNSM.2017.2788945.

[12] Qazi, Faiza, Osman Khalid, Rao Naveed Bin Rais, and Imran Ali Khan. "Optimal content caching in content-centric networks." Wireless Communications and Mobile Computing 2019 (2019).

[13] Khabbiza, El Hassane, Rachid El Alami, and Hassan Qjidaa. "Peer-Assisted Content Delivery to Reduce the Bandwidth of TSTV Service in IPTV System." International Journal of Digital Multimedia Broadcasting 2019 (2019).

[14] Liu, Yujie, Dianjie Lu, Guijuan Zhang, Jie Tian, and Weizhi Xu. "Q-learning based content placement method for dynamic cloud content delivery networks." IEEE Access 7 (2019): 66384-66394.

[15] Duan, Jie, Yuan Xing, Ruilin Tian, Guofeng Zhao, Shuai Zeng, Yuanni Liu, and Chuan Xu. "SCDN: A novel software-driven CDN for better content pricing and caching." IEEE Communications Letters 22, no. 4 (2018): 704-707.

[16] Qu, Hua, Gongye Ren, Jihong Zhao, Zhenjie Tan, and Shuyuan Zhao. "Joint Optimization of Content Placement and User Association in Cache-Enabled Heterogeneous Cellular Networks Based on Flow-Level Models." Wireless Communications and Mobile Computing 2018 (2018).

[17] Benkacem, Ilias, Tarik Taleb, Miloud Bagaa, and Hannu Flinck. "Optimal VNFs placement in CDN slicing over multi-cloud environment." IEEE Journal on Selected Areas in Communications 36, no. 3 (2018): 616-627.

[18] Alghamdi, Fatimah, Saoucene Mahfoudh, and Ahmed Barnawi. "A novel fog computing based architecture to improve the performance in content delivery networks." Wireless Communications and Mobile Computing 2019 (2019).

[19] Asheralieva, Alia, and Dusit Niyato. "Game theory and Lyapunov optimization for cloud-based content delivery networks with device-to-device and UAV-enabled caching." IEEE Transactions on Vehicular Technology 68, no. 10 (2019): 10094-10110.

[20] Bosunia, Mahfuzur Rahman, and Seong-Ho Jeong. "Efficient Content Delivery for Mobile Communications in Converged Networks." Wireless Communications and Mobile Computing 2019 (2019).

[21] Liu, Ling, Yiqing Zhou, Jinhong Yuan, Weihua Zhuang, and Ying Wang. "Economically optimal MS association for multimedia content delivery in cache-enabled heterogeneous cloud radio access networks." IEEE Journal on Selected Areas in Communications 37, no. 7 (2019): 1584-1593.

[22] Wang, Ning, Gangxiang Shen, Sanjay Kumar Bose, and Weidong Shao. "Zone-based cooperative content caching and delivery for radio access network with mobile edge computing." IEEE Access 7 (2018): 4031-4044.

[23] Lau, Chun Pong, Abdulrahman Alabbasi, and Basem Shihada. "An efficient content delivery system for 5G CRAN employing realistic human mobility." IEEE Transactions on Mobile Computing 18, no. 4 (2018): 742-756.

[24] Haghighi, Ali A., Shahram Shah Heydari, and Shahram Shahbazpanahi. "Dynamic QoS-aware resource assignment in cloud-based content-delivery networks." IEEE Access 6 (2017): 2298-2309.

[25] Haghighi, Ali A., Shahram Shahbazpanahi, and Shahram Shah Heydari. "Stochastic QoE-aware optimization in cloud-based content delivery networks." IEEE Access 6 (2018): 32662-32672.

[26] Sinky, Hassan, Bassem Khalfi, Bechir Hamdaoui, and Ammar Rayes. "Responsive content-centric delivery in large urban communication networks: A LinkNYC use-case." IEEE Transactions on Wireless Communications 17, no. 3 (2017): 1688-1699.

[27] Siracusano, Giuseppe, Roberto Bifulco, Martino Trevisan, Tobias Jacobs, Simon Kuenzer, Stefano Salsano, Nicola Blefari-Melazzi, and Felipe Huici. "Re-designing dynamic content delivery in the light of a virtualized infrastructure." IEEE Journal on Selected Areas in Communications 35, no. 11 (2017): 2574-2585.

[28] Fan, Qilin, Hao Yin, Zexun Jiang, Haojun Huang, Yan Luo, and Xu Zhang. "Adaptive Content Management for UGC Video Delivery in Mobile Internet Era." Mobile Information Systems 2016 (2016).

[29] Papagianni, Chrysa, Aris Leivadeas, and Symeon Papavassiliou. "A cloud-oriented content delivery network paradigm: Modeling and assessment." IEEE Transactions on Dependable and Secure Computing 10, no. 5 (2013): 287-300.

[30] Salahuddin, Mohammad A., Amina Mseddi, Halima Elbiaze, and Roch H. Glitho. "Popularity and Correlation-aware Content Placement for Hierarchical Surrogates in Cloud-based CDNs." In GLOBECOM 2017-2017 IEEE Global Communications Conference, pp. 1-6. IEEE, 2017.

[31] Ghalehtaki, Razieh Abbasi, Somayeh Kianpisheh, and Roch Glitho. "A Bee Colony-based Algorithm for Micro-cache Placement Close to End Users in Fog-based Content Delivery Networks." In 2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 1-4. IEEE, 2019.

[32] Ha, Minkeun, and Daeyoung Kim. "On-demand cache placement protocol for content delivery sensor networks." In 2017 international conference on computing, networking and communications (ICNC), pp. 207-216. IEEE, 2017.

[33] Retal, Sara, Miloud Bagaa, Tarik Taleb, and Hannu Flinck. "Content delivery network slicing: QoE and cost awareness." In 2017 IEEE International Conference on Communications (ICC), pp. 1-6. IEEE, 2017.