International Journal of Computer and Communication Technology

Volume 6 | Issue 4 Article 9

October 2015

AN EFFICIENT ALGORITHM FOR CLUSTERING NODES, CLASSIFYING AND REPLICATION OF CONTENT ON DEMAND BASIS FOR CONTENT DISTRIBUTION IN P2P OVERLAY NETWORKS

ANNA SARO VIJENDRAN
Department of MCA SNR Sons College (Autonomous), (Affiliated to Bharathiar University) Coimbatore, saroviji@rediffmail.com

S. THAVAMANI
Department of Computer Applications, SNR Sons College (Autonomous), (Affiliated to Bharathiar University), Coimbatore, thavamaniphd11@gmail.com

AN EFFICIENT ALGORITHM FOR CLUSTERING NODES, CLASSIFYING AND REPLICATION OF CONTENT ON DEMAND BASIS FOR CONTENT DISTRIBUTION IN P2P OVERLAY NETWORKS

ANNA SARO VIJENDRAN[1] & S. THAVAMANI[2]

[1]Department of MCA SNR Sons College (Autonomous), (Affiliated to Bharathiar University) Coimbatore
[2]Department of Computer Applications, SNR Sons College (Autonomous), (Affiliated to Bharathiar University), Coimbatore
Email:saroviji@rediffmail.com, thavamaniphd11@gmail.com

Abstract-This paper proposes an efficient algorithm for clustering nodes, classifying and replication of content on demand basis for content distribution in p2p overlay networks. Peer-to-peer overlay networks are a good solution for distributed computing than client-server model since nodes in P2P networks act both as client and server. With such functionality of managing replicas is an astonishing task and we designed an architecture for content distribution. In the proposed architecture, nodes are grouped into strong, medium and weak clusters based on their weight vector. More copies of Class I content are replicated in strong clusters that are having high weight values. Class II and Class III clusters consumes medium and low queries. Routing is performed hierarchically by broadcasting the query to the strong clusters first, then to medium clusters and finally to weak clusters. From extensive simulations using NS2 simulator, the proposed architecture achieves less bandwidth consumption, reduced latency, reduced maintenance cost, strong connectivity and query coverage.

1. INTRODUCTION

Peer-to-peer which is a distributed computer architecture are designed for sharing the computer resources such as storage, content, CPU cycles directly irrespective using a centralized server. P2P networks are classified based on their failure adaptation abilities, connectivity along with maintenance of suitable connectivity. Remarkable research attention has been pertained to content distribution which is a significant peer-to-peer application on the internet. Overlay networks are supple and easy to deploy which allows users to achieve distributed operations without changing or modifying the available underlying physical network.

Various peer-to-peer networks try to address the performance problem through using different mechanisms in order to enhance the Quality of Service shortly termed as QoS . Since there are several metrics meant for QoS [1, 2], for the replica placement problem the metrics such as throughput, delay also called as latency, bandwidth utilization and query efficiency are chosen which best suits.

A distributed hash table which is a type of decentralized distributed system is capable enough to offer a lookup service. It is like hash table from which any nodes present in the peer-to-peer network can efficiently retrieve the data / content required. The usage of distributed hash table is wide and it best suits to build distributed file systems, DNS, instant messaging, and can be extensively used in the peer-to-peer file sharing along with replica placement strategies.

A distributed replication group consists of several servers dedicating some storage for the replicas. A server has to serve requests from its clients and also from other servers in the group. When a server receives a request from a client, it immediately responds to the client if the object is in its local storage. Otherwise, the object is fetched from other servers within the group at a higher access cost or from the origin server, at an even higher cost, in the case no server within the group stores a replica of the object. [4]

1.1. Problem Statement

The motivation behind this work is to design and develop an efficient algorithm for clustering nodes, classifying and replication of content on demand basis for content distribution in p2p overlay networks. For attaining the above it is needed to develop a query search retrieval algorithm which is adaptive to the peer-to-peer overlay networks. In this research three clusters are formed based on the node's weight. The weight of each node is measured using a strategy. The clusters are classified into three in order to improve the QoS performance. Hence, an efficient algorithm for clustering nodes, classifying and replication of content on demand basis for content distribution in p2p overlay networks is proposed in this research.

2. LITERATURE REVIEW

In [3] the authors proposed protocol which combine DHT efficiency with gossip robustness and take into account the interests and localities of peers. Their scheme Flower-CDN provides a hybrid and locality-aware routing infrastructure for user queries. PetalUp-CDN is a highly scalable version of Flower-CDN that dynamically adapts to variable rates of participation and prevent overload situations. In addition, they ensured the robustness of their P2P CDN via low-cost maintenance protocols that can detect and recover from churn and dynamicity. The extensive performance evaluation shows that their protocols yielded high performance gains under both static and highly dynamic environments. Furthermore, it incurred acceptable and tunable overhead. Finally they provided important guidelines to deploy Flower-CDN for the public use.

In [6], the authors presented a paper which studied the Quality-of-Service (QoS)-aware replica placement problem in a general graph model. Since the problem was proved NP-hard, heuristic algorithms are the current solutions to the problem. However, those algorithms cannot always find the effective replica placement strategy. Hence they proposed two algorithms which can obtain better results

within the given time period. The first algorithm is called Cover Distance algorithm, which is based on the Greedy Cover algorithm. The second algorithm is an optimized genetic algorithm, in which we use random heuristic algorithms to generate initial population to avoid enormous useless searching. Then, the 0-Greedy-Delete algorithm is used to optimize the genetic algorithm solutions. According to the performance evaluation, their Cover Distance algorithm can obtain relatively better solution in time critical scenarios. Whereas, the optimized genetic algorithm is better when the replica cost is of higher priority than algorithm execution time. The QoS-aware data replication heuristic algorithms are applied into the data distribution service of an astronomy data grid pipeline prototype, and the operation process is studied.

In [7], Caviglione and Cervellera have presented a paper entitled introduced "Design, optimization and performance evaluation of a content distribution overlay for streaming". In their perspective, the paper introduced an overlay Content Distribution Network (CDN) able to sustain the real-time delivery of data streams. To better use resources, and to face the churn affecting users, the control and optimization of the CDN are performed through a model predictive control scheme. Simulations of two use cases are provided to show the effectiveness of the proposed solution. In particular, the stream of multimedia and interactive grid data was considered.

In [8], the authors proposed a proactive replication strategy to improve search efficiency for rare objects which uses an object-probing technique for peers to decide whether or not to establish replications for their objects when they join the network. The strategy can effectively increase the popularity of rare objects in order to enhance search efficiency. A rare object search algorithm is used to reduce the overhead caused by the replication strategy. When a peer forwards a search request, the forward probability is calculated according to its neighbors' degrees and the number of neighbors' objects. Therefore, the search request is forwarded to the peers more likely containing target objects. Simulations showed that their proactive replication strategy greatly improves search efficiency for rare objects with moderate communication overhead. The rare object search algorithm not only improves search efficiency for rare objects, but also achieves load balance in search.

In [9] the authors have presented a structured P2P system called Donuts for range queries, which takes two important yet somewhat conflicting issues into account of proximity and load balance. Proximity allows physically close nodes to be arranged near each other in the overlay so as to reduce the cost of neighbor communications that occur quite often in a range-queriable system. Load balance is crucial because object distribution in a semantically meaningful key space is often skewed. Efficient load balance, however, requires flexible node position in the overlay, and thus conflicts with proximity. Donuts resolve the problem by separating physically close nodes into several overlay sections. By dynamically switching between these sections, they help one another balance their loads without altering overlay proximity too much. Still, breaking apart physically close nodes inevitably compromises overlay proximity. Therefore, the authors put effort in the overlay construction to ensure that load balance can be performed effectively and efficiently, with minimal damage to overlay proximity.

Jian Zhou et al. [10] have shown that the replica placement problem in P2P networks has represented as a Clustered KCenter problem (which essentially differed from the classic kcenter problem) and is proven to be NP-complete. To solve the problem, they bring forward an approximation algorithm in the form of a distance graph for the network topology; when their defined feasibility condition has hold at a certain point; the replica placement solution has built out of (m-1) power of current distance graph.

Yan Chen et al. [11] have proposed the dissemination tree, a dynamic content distribution system built on top of a peer-topeer location service. They have presented a replica placement protocol that has built the tree while meeting QoS and server capacity constraints. The number of replicas as well as the delay and bandwidth consumption for update propagation was significantly reduced.

3. AN EFFICIENT ALGORITHM FOR CLUSTERING NODES, CLASSIFYING AND REPLICATION OF CONTENT ON DEMAND BASIS FOR CONTENT DISTRIBUTION IN P2P OVERLAY NETWORKS

3.1. Overview

In our proposed Quality of Service (QoS) conscious topology, nodes are clustered into three categories namely strong, medium and weak depending on their weight. The weight is estimated based on the parameters namely available capacity, CPU speed, Size of the memory and access delay. Routing is performed hierarchically by broadcasting the query only to the strong and medium clusters. Hence the proposed mechanism achieves less bandwidth consumption, reduced delay, lesser maintenance cost along with sturdy connectivity and query coverage. The system model is briefly dealt in the next section. This research work focuses on design and development of smart replica management which is much aware on QoS metrics such as throughput, delay, query efficiency and bandwidth utilization. This model is proposed for peer-to-peer overlay networks for content distribution. The queries are first sent to the nearby node and if it is not found it will be retrieved from the originating server.

3.2. System model

It is considered with a collection of N server nodes that form a peer to peer (P2P) overlay network. Along the part of the overlay, each node in the network functions as a server responding to queries which come from clients outside of the overlay network. As an example could be that each node is a web server with the overlay linking the servers and clients being web browsers on remote machines requesting content from the servers. It is assumed that each node always stores one copy of its own content item which it serves to clients and that it has additional storage space to store the replicated content items from other nodes which it can also serve. The object is associated with an authoritative origin server (OS) in the network where the content provider makes the updates to the object. The object copy located at the origin server is called the origin copy and an object copy at any remaining server is called a replica.

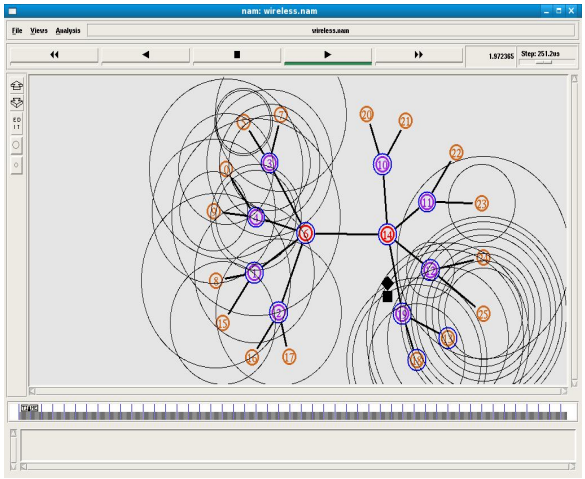

Fig1. Clustering Nodes

3.3. Clustering nodes

For each node Ni, i=1 to n. Let

POB_i – Bandwidth obtained, PCS_i- CPU Speed

PAD_i – Access Delay PMZ_i – Size of the memory.

The weight is calculated using the below equation

$$PW_i = (POB_i + PCS_i + PMZ_i) / PAD_i$$

Clustering the nodes is done using the following methodology

β *is the individual node weight*

$\beta_1 = \beta + 10$

$\beta_2 = \beta - 10$

If weight $> \beta_1$ *then Strong cluster*

ElseIf weight $\leq \beta_1$ *and* $\geq \beta_2$ *then*

Medium cluster

ElseIf weight $< \beta_2$ *then Weak cluster*

Algorithm steps are as follows

1.The weight of the peer Pi can be calculated as follows

$$PW_i = (POB_i + PCS_i + PMZ_i) / PAD_i$$

2. Form the vector $PW = \{ P_i, PW_i \}$, where P_i denotes the node ids and PW_i denotes their corresponding weight values, sorted on the decreasing order.

3. i) Let $\{P_{SC}\}$ denote the set of strong cluster nodes $(0 < = P_{SC} <= n)$, which satisfies the following condition

$$PW_i > \beta_1$$

where β_1 is the high level minimum threshold value for the weight. The remaining nodes are $\{R_N\} = \{N\} - \{P_{SC}\}$.

ii) Let $\{S_{WC}\}$ denote the set of weak cluster nodes $(0 <= P_{WC} <= n)$, which satisfies the following condition

$$PW_i < \beta_2$$

where β_2 is the low level maximum threshold value for the weight. The remaining nodes are $\{R_{N1}\}=\{R_N\} - \{P_{WC}\}$.

iii) Then the set $\{P_{MC}\} = \{Ni\} - \{\{P_{SC}\} + \{P_{WC}\}\}$, denote the set of medium cluster nodes $(0 < = P_{MC} < n)$, which satisfies the following condition

$$PW_i <= \beta_1 \quad and \; PW_i >= \beta_2$$

A server knows the request rates of its local users for all objects. For the placement matrix m X n matrix is represented by PM

$$PM_{ij} = \begin{cases} -1, if \; o_j in \; Localcatch \\ 0, if \; o_j inStrong \\ 1, if \; o_j inMedium \\ > 1, if \; o_j inWeak \end{cases}$$

3.4 Content Classification and Replica Placement Algorithm

It is considered that QrySrv be the Query Server that registers the query of each client. The query server saves the cluster information of each node along with the node id as S, M, and W for strong, medium and weak clusters respectively. At time θ, let p number of clients sends the query request. The requested content of the queries are classified as class I, class II and class III based on the frequency of access.

A query Qj, j<m,

If($n(Qj) > \beta_{max}$) then O_{req} is considered to be C_I

Elseif ($n(Qj) <= \beta_{max}$ and $n(Qj) >= \beta_{min}$) then Q_{req} is considered to be C_{II}

Elseif ($n(Qj) < \beta_{min}$) is considered to be $C_{III.}$

where n(Qj) is the number of access of the content pattern for the given query and β_{max} is the low level maximum access threshold value, β_{min} is the high level minimum access threshold value.

Create-Replica()

{

1.For each O_{req} form C_i

2.If (O_{req} € C_I) then Assign O_{req} to P_{SC}

Else If (O_{req} € C_{II}) then Assign O_{req} to P_{MC}

Else If (O_{req} € C_{III}) then Assign O_{req} to P_{WC}

3.If Replica of O_{req} then Serve request

Eelse If PBWi && PSPi && PMZi available

Create replica(O_{req})

}

The QrySrv assigns the class I contents to strong cluster nodes, class II contents to medium cluster nodes and class III contents to weak cluster nodes. The assignment of such thing is completed and the pattern information will be replicated to the origin server (OS). The origin server (OS) will carry out replication placement, concurrently to the pattern information got from the QrySrv. The cost value COST of each node is stored. The Origin Server broadcasts the replication information to the respective clients.

$$\min \sum_{i=1}^{m} \left(\sum_{j:PM_{ij}=-1} r_{ij} t_l + \sum_{j:PM_{ij}=0 and (PW_i > \beta_{max})} r_{ij} t_S + \sum_{j:PM_{ij}=1 and (PW_i \leq \beta_{max} and PW_i \geq \beta_{min})} r_{ij} t_M + \sum_{j:PM_{ij}>1 and (PWi < \beta_{min})} r_{ij} t_W \right)$$

Here the first term represent the requested object accessed from local hit, and second, third and fourth term represents the requested object is accessed from strong group, medium group and weak group of clusters respectively.

4. SIMULATION RESULTS

From the Fig.3 it can be observed that packet delivery ratio of the proposed protocol EACNCRC is constantly improved than the QIRMA. Also from Fig.4 it is shown that EACNCRC has consumed less energy than QIRMA. The total number of strong, medium and weak clusters formed and number of replications made is shown in Fig.5 and Fig.6 respectively. The delay of EACNCRC and QIRMA is compared in Fig. 7 and it can be observed that the proposed EACNRC has got less delay. Also the throughput gets enhanced and overhead is reduced in the proposed protocol which is shown in Fig.8 and Fig.9 respectively.

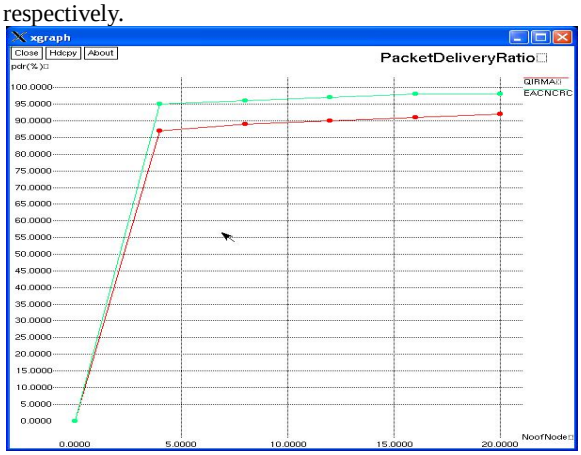

Fig.3 No. of Nodes Vs Packet Delivery Ratio

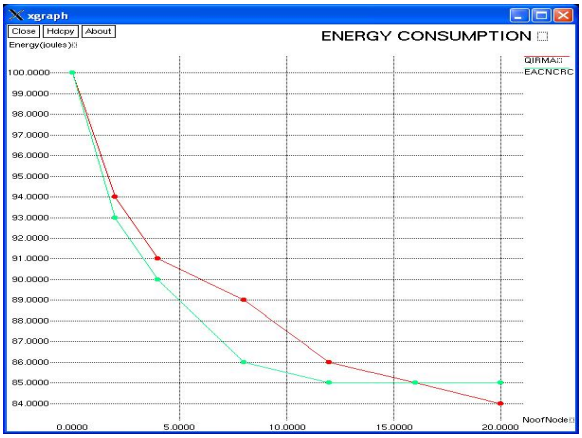

Fig.4 No. of Nodes Vs Energy Consumption

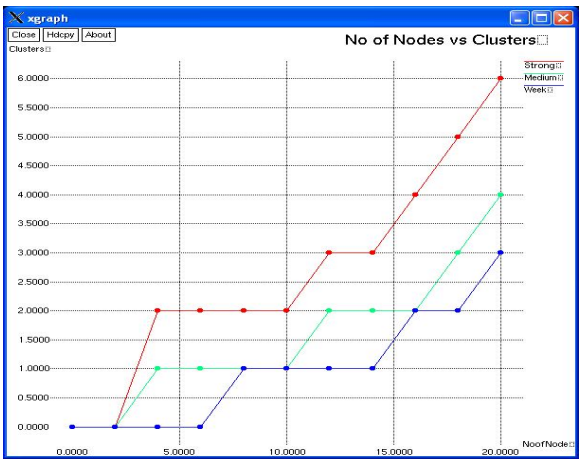

Fig.5 No. of Nodes Vs Clusters

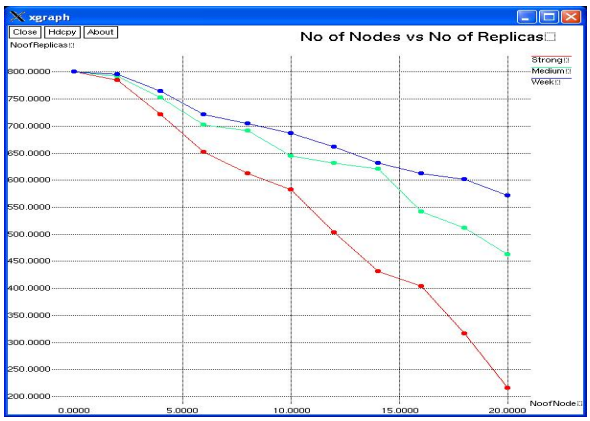

Fig.6 No. of Nodes Vs No. of Replicas

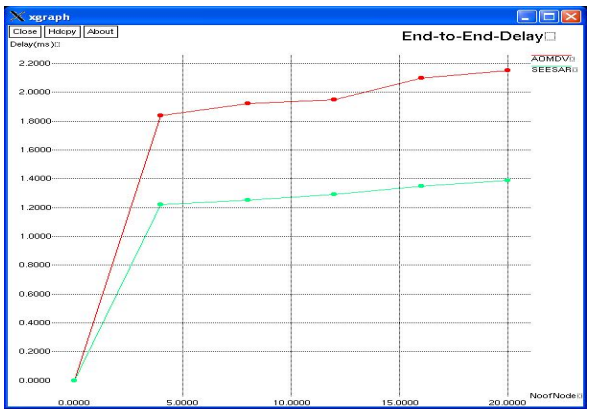

Fig.7 No. of Nodes Vs Delay

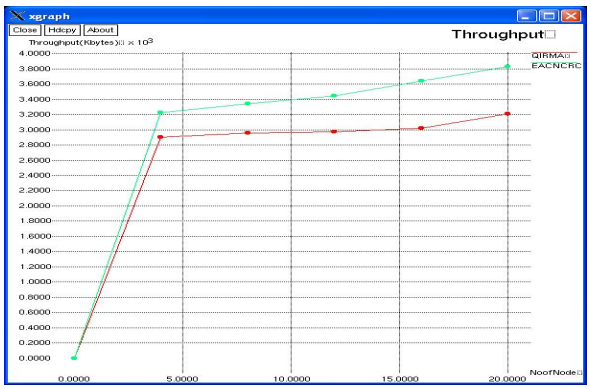

Fig.8 No. of Nodes Vs Throughput

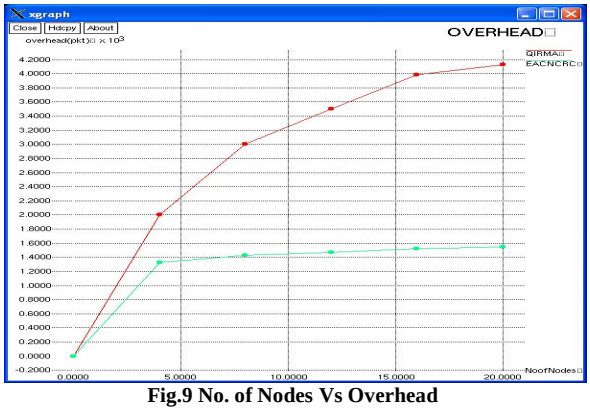

Fig.9 No. of Nodes Vs Overhead

3. CONCLUSION AND FUTURE WORKS

This research work aims at design and development of proposes an efficient algorithm for clustering nodes, classifying and replication of content on demand basis for content distribution in p2p overlay networks. We designed architecture for content distribution. In the proposed architecture, nodes are grouped into strong, medium and weak clusters based on their weight vector. In the proposed replica placement algorithm, the contents are classified as class I, class II and class III, based on their access patterns. More copies of Class I content are replicated in strong clusters that are having high weight values. Routing is performed hierarchically by broadcasting the query only to the strong clusters. From wide simulations using NS2 simulator, the proposed architecture achieves less bandwidth consumption, reduced latency, reduced maintenance cost, strong connectivity and query coverage. The proposed protocol is comparatively adaptive. It is observed from the results that the proposed work better in terms of throughput, query efficiency and bandwidth utilization. In future the proposed work can be enhanced with delay tolerant mechanisms. At the whole the proposed protocol is better than that of the QIRMA.

REFERENCES

[1] Dr. Anna saro vijendren and S.Thavamani “Survey of Caching and Replica Placement Algorithm for Content Distribution in Peer to Peer Overlay Networks”. The Second International conference on Computational Science, Engineering and Information Technology (CCSEIT-2012) October 2012, Coimbatore. India. Conference proceedings published by ACM that available in ACM digital library.

[2] Dr. Anna saro vijendren and S.Thavamani “Analysis Study on Caching and Replica Placement Algorithm for Content Distribution in Distributed Computing Networks”. International Journal of Peer-to-Peer Networks. Nov-2012, Vol.3, No:6.PP.13-21.

[3] N. Laoutaris, O. Telelis, V. Zissimopoulos, and I. Stavrakakis, “Distributed Selfish Replication,” IEEE Trans. Parallel and Distributed Systems, vol. 17, no. 12, pp. 1401-1413, Dec. 2006.

[4] Sharrukh Zaman, and Daniel Grosu, "A Distributed Algorithm for the Replica Placement Problem", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 22, NO. 9, SEPTEMBER 2011, pp.1455 - 1468.

[5] Paraskevi Raftopoulou and Euripides G.M. Petrakis, "iCluster: a Self-Organizing Overlay Network for P2P Information Retrieval", in proc. of ECIR, 30 March- 3 April 2008.

[6] Seung Chul Han, Ye Xia, "Network load-aware content distribution in overlay networks", Computer Communications, ELSEVIER, Volume 32, Issue 1, 23 January 2009, Pages 51–61.

[7] Manal El Dick, Esther Pacitti, Reza Akbarinia, Bettina Kemme, "Building a peer-to-peer content distribution network with high performance, scalability and robustness", Information Systems, ELSEVIER, Volume 36, Issue 2, April 2011, Pages 222–247.

[8] Zhihui Du, Jingkun Hu, Yinong Chen, Zhili Cheng, Xiaoying Wang, "Optimized QoS-aware replica placement heuristics and applications in astronomy data grid ", Journal of Systems and Software, ELSEVIER, Volume 84, Issue 7, July 2011, Pages 1224–1232.

[9] Luca Caviglione, Cristiano Cervellera, "Design, optimization and performance evaluation of a content distribution overlay for streaming", Computer Communications, ELSEVIER, Volume 34, Issue 12, 2 August 2011, Pages 1497–1509.

[10] Jian Zhou, Xin Zhang, Laxmi Bhuyan and Bin Liu,"Clustered KCenter: Effective Replica Placement in Peer-to-Peer Systems", IEEE conference on Global Telecommunications, pp.2008-2013 Nov. 2007.

[11] Yan Chen, Randy H. Katz and John D. Kubiatowicz"Dynamic Replica Placement for Scalable Content Delivery", Lecture Notes In Computer Science; Vol. 2429, pp.306-318, 2002.

AUTHORS

Dr. Anna Saro Vijendran is the Director – MCA in SNR Sons College, Coimbatore, India. She has a teaching experience of 20 years in the field of Computer science. Her area of Specialization is Digital Image Processing and Artificial Neural Networks .She has presented more than 30 Papers in various Conferences and her research works have been published in International Journals. She is currently a Supervisor for research works of various Universities and also Reviewer for reputed Journals.

S, Thavamani is an Assistant Professor in Department of Computer Applications, SNR Sons College, Coimbatore, India. She has a teaching experience of 12 years in the field of Computer science. Her area of Specialization is Distributed Computing and Networks. She has presented more than 15 Papers in various International and National Conferences. She is currently a supervisor for M.Phil research works of various Universities. She is currently pursuing Ph.D Degree in SNR Sons College, under Bharathiar University, Coimbatore.