Context-Aware Replica Placement in Peer-to-Peer Overlay Networks

Mohammad H. Al Shayeji, M. D. Samrajesh, Saud Abdulaziz Al-Behairy, and Khalid Assaf Al-Enazi

***Abstract*—The high demand for rich on-line media content such as Video-On-Demand (VOD) has lead to a rapid increase in internet congestion. Today peer-to-peer overlay networks play a key role in content delivery. Replication in P2P networks can significantly reduce congestion and minimize access time. However, the challenge is to decide on which content to replicate, where to place replicas, which replica to replace and which replica is to be designated as the primary copy. The current distribution and delivery of content can be improved by making replica content placements, replacement and management context aware.**

In this paper we propose a context-aware replica placement algorithm that reduces access time by placing multiple copies of content at various strategic locations using a number of context-aware parameters; we also present a replica replacement algorithm considering the size of the content as well as other important factors. Moreover, we propose a primary copy assignment algorithm that minimizes the content update overhead by choosing the right replica as the primary copy. Our discussions and comparative study with other algorithms shows that the proposed algorithm is effective.

***Index Terms*—Access latency, overlay networks, peer-to-peer (P2P), replica placement.**

I. Introduction

The massive growth of digital content has created a digital world where intelligent user-friendly devices and wide variety of communication infrastructures are less significant if the required digital content is not easily accessible to the user [13],[11]. Replication improves content availability, reduces access latency and improves system reliability. However, replication in large scale Content Delivery Networks (CDN) is costly [12]. The alternative solution is to use Peer-to-Peer (P2P) overlay networks.

An overlay network is a network which is built on top of another network. Nodes in the overlay are considered to be connected by virtual or logical links, each of which corresponds to a path. A path may actually use many physical links in the underlying network [15], [6]. P2P networks are overlay networks because their nodes run on top of the Internet

Today, P2P file sharing is the dominant traffic type in the Internet, exceeding all other Internet based data transfer including the Web [16]. P2P Overlay networks offers placement of replicated content, searching and sharing. However, one of the main challenges of replication in overlay networks is to find the right number of replicas that would achieve optimal performance [10]. Replication in Overlays Network can be defined as a graph $G = (V, E)$, where V is the set of nodes and $E \subset V \times V$ is the set of links between the nodes in the overlay network. Each is node associated with a bandwidth $Nb(i)$.

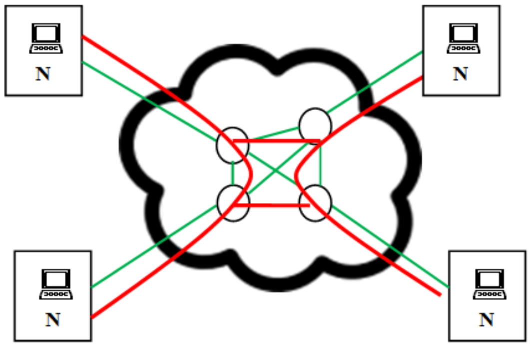

Fig. 1. A typical layout of overlay networks.

Our main contribution in this paper is a context-aware replica placement algorithm that reduces access time by placing multiple copies of content at various strategic locations using various context-aware parameters; the algorithm calculates the Cost of Transfer (CoT) based on the size of the content, request rate for the content to be replicated and the CoT of replicas held in the cluster in deciding replica placement and replacement. Moreover, we propose a primary copy assignment algorithm that minimizes the content update overhead by choosing the right node for assigning the primary replica.

The paper is structured as follows: Section II discusses related work, Background information on overlay networks is presented in Section III. Context aware replica placement, replacement and primary copy assignment algorithms is presented in Section IV, Our discussions by comparative study is presented in Section V and finally in Section VI we have our conclusion and future work.

II. Related Work

Many replica placement algorithms have been proposed for CDN [14] and some for P2P overlay networks [20]. In general, the goal is to optimize performance and minimize the infrastructure cost. In [1] content is distributed and placed uniformly across the nodes of a hierarchical naming sub-tree, however this is not practically effective since internet architecture is not similar to tree structure.

Manuscript received September 12, 2012; revised November 28, 2012.

The authors are with the Computer Engineering Department, Kuwait University, Kuwait (e-mail: alshayej@eng.kuniv.edu.kw, sam@differentmedia-kw.com).

DOI: 10.7763/IJCTE.2013.V5.714

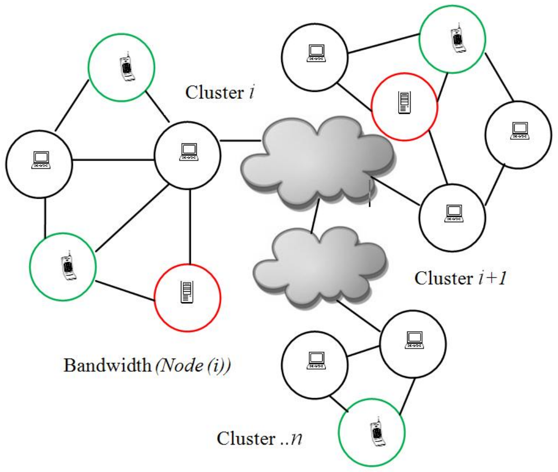

Fig. 2. Different type of devices connected in a P2P cluster.

Distributed paging technique proposed in [2] deals with the dynamic allocation of copies of content in a distributed network such as to minimize the communication cost over a series of read and write requests, but the life time of the cache and overhead associated with caching and distributed caching are of more concerns.

A replica update approach that scales in term of number of users or in term of number of editions was proposed in [3] which ensure causality, consistency and intention preservation criteria. However, this was not a general applicable approach, since it was presented only in context of Wikipedia and has additional overhead for a causal broadcast to achieve convergence.

Minimum or optimal replication problem has been discussed for file sharing applications in [4], [5] but most of the work only provides replications based on a centralized approach. Replica placement QoS requirements were considered in [7] for content delivery among content servers. This also only provided a centralized greedy-based heuristic algorithm.

Creation and deletion of file replica by dynamically adapting to time-varying file popularity index in a decentralized manner based on the query traffic was proposed in [9], however it is not an optimal solution since no other criteria's other than popularity index are considered in replication creation and deletion.

Recent research on latency associated with file replication in P2P system has been studied in [17]. The paper emphasized the importance of search time and the time required by the peers to transmit the file, however the model is applicable for multipart downloads and had not considered the file transfer delay related to replication in the peers.

Our approach of context-aware replica placement is different from the earlier replica placements. We try to place replica in the cluster of node which is relatively efficient and that guarantees the delivery of content to every other node in the overlay network with minimum latency. Secondly, in case of replica replacement the size of the content which is an important factor is also considered while evicting/replacing replicas. Finally, once replica is created a tracking mechanism will track the replica and assign the primary replica copy based on access pattern such that the update overheads are minimal.

III. Background

A. Overlay Networks

An overlay network is a network which is built on top of another network. Nodes in the overlay are considered to be connected by virtual or logical links, each of which corresponds to a path. A path may actually use many physical links in the underlying network. Overlay networks offer guaranteed data retrieval, automatic load balancing, and self-organization [15],[8].

B. Peer-to-Peer (P2P)

Peer-to-peer (P2P) overlay systems can make portion of their resources including disk storage available to other network participants. The attraction of these systems, when compared to client/server frameworks, is in their robustness, reliability and cost efficiency. Unlike traditional distributed computing, P2P networks aggregate large number of computers and possibly mobile or handheld devices, which tend to join and leave the network frequently [21].

Nodes in a P2P network are called peers; their roles vary based on the interaction with other peers. When requesting information's they are clients. When providing information's to other peers they are servers. When they forward information to other peers they are routers. This type of interaction creates application level virtual networks with their own overlay topology [21]. To search for data or resources, messages are sent from one peer to another with each peer responding to the request for information it has stored locally.

IV. Context Aware Replica Placement

A. System Model

Replication in Overlays network can be defined as graph $G = (V, E)$, where V is the set of nodes and $E \subset V \times V$ the set of links between the nodes in the overlay network. Each node is associated with a bandwidth $Nb(node(i))$.

B. Assumptions

- Each node including the requesting node can store replica copy.
- Node storage space is limited and additional replica if needed can only be placed by replica replacement.
- Primary copy can reside only in one node.
- Distributed list of replica information is stored across the overlay network in a node in each cluster.
- Cluster of peers is a set of nodes grouped physically. (e.g. nodes under a ISP, nodes in a university campus)

C. Context-Aware Replica Placement Algorithm

The CAR (Context-Aware Replication) algorithm is invoked when a node requests access to a content which is not available in the local cluster. The algorithm decides whether or not to create local replica and where to place the new replica in cluster.

```
Algorithm: Create New Replica
Input   : Context sensitive values
Output: Place Replica  or not
            // Nbi – Max available bandwidth (kbps)
  For (i) < number of nodes in cluster Ni= 1...n
            Nbi= bandwidth(node(i)).
             Sort Nbi(Descending), Choose first Nbi
  If (Node(i) is full) then
           Choose next node from the sorted list;
  Else // none is empty
      If (CoT(request(i)) >   threshold) then
      Call ReplaceReplica()
           Else  Access content directly. End if
  End if ; Propagate updates across the network.
```

Fig. 3. Create new replica algorithm

Initially the nodes are sorted on descending order based on their available bandwidth. The node with the highest available bandwidth is picked as a possible replica host since the node with the highest available bandwidth would most likely serve future request with minimum latency.

In case the top node is not able to hold content due to space constraint then the algorithm chooses the next node from the list. If all nodes in the cluster are not able to host the new replica then the algorithm has to decide whether to replicate the content or access it remotely. To do so the algorithm tracks the number of request originating from the cluster nodes for a specific content. When the numbers of requests from the cluster nodes reach a specific threshold then, the Replica Replacement algorithm is invoked else the content is directly accessed from the origin source. The threshold is a dynamically calculated value that depends on the popularity of local content (i.e. the number of request served by local replicas) and replica size.

As stated above the replica replacement algorithm is invoked only when no node in the cluster is able to satisfy the placement criteria and when the demand for the content has been substantial i.e. above the specified threshold. The algorithm compares the Cost of Transfer (CoT) of the remote content by multiplying the number of requests by content size. The algorithm also tracks the CoT of all local replicas. To calculate the threshold of a remote content the algorithm sorts the CoT of all replicas in each peer in the cluster. The threshold is them calculated by the minimum total CoT of the local replicas that needs to be removed to accommodate the remote content.

The algorithm selects one of more victim replicas in the same node/peer to avoid fragmentation. Moreover, the algorithm ensures that the sum of the CoT's of all the victim replicas is less than the CoT of the new replica otherwise a replacement is not performed.

Consider the scenario in Fig. 5 for requested content R_{19}, the CAR algorithm will sort all local replicas in the cluster based on the CoT and picks a victim(s) to replace. The replica with the minimum CoT is R_2 in *Node 1*, since its size is less than R_{19} it is not possible to accommodate the new content. The next minimum i.e. R_6 from *Node 3* is also considered as a possible victim. The combined size of both R_2 and R_6 is greater or equal to the size of R_{19} and the combined CoT is less than R_{19} CoT, but, since R_2 and R_6 belong to different node the replacement will not take place. CAR will try to find the next victim ignoring R_6 in *Node 3*.

```
Algorithm: Replica Replacement
Input   : Replica ID, context sensitive values
Output: Replacement of Replica
   Sort(ascending) based on CoT
    Pick top(i)
     If (size(top(i))>=size(request) and CoTi
         <=CoT(request))
         Replace top(i) with new content
       Else
         Find slots such that (∑size(i)) >=  size(request) and
         such that each CoTi<= CoT(request) and ∑ CoTi<=
        CoT(request) and all slots belongs to same node.
        If (available)
           Replace(top(i)..(i+n)) replica with new content
  Else
        Replacement currently not possible. End if
     End if;   Propagate updates across the network.
```

Fig. 4. Replace replica algorithm

R_2 and R_1 from *Node 1* are selected since their combined CoT is less than the requesting contents CoT, they belongs to the same node, and their combined size is greater or equal to the size of R_{19}. So a new replica is created in *Node 1* replacing R_1 & R_2. However, In case of requested content R_{20} since its CoT is less than the combined CoT of R1 & R_2, replacement will not take place under the current scenario.

D. Assign Primary Copy Algorithm

We assume that the replica with large requests in a region has a higher possibility of updates from that region, hence to reduce the overhead of content updates the algorithm chooses the cluster that has the high request count for a specific replica and assign the replica of the node as the primary copy.

A key feature of P2P networks is decentralization we have distributed list that hold information's of replica. This has many advantages such as robustness, availability of information and fault-tolerance, In case the primary copy node fails, the next node that has the highest Request count from the P2P network is promoted to be the primary copy node.

Requests at time 't'

Request	Size	Request	CoT
(R_{19})	135	5	675
(R_{20})	100	6	600

Node-1

Replica	Size	Request	CoT
R_2	75	4	300
R_1	85	4	340
R_7	80	10	800
R_{13}	90	20	1800
R_{17}	75	25	1875
R_{18}	25	100	2500

Node-2

Replica	Size	Request	CoT
R_{15}	100	8	800
R_{14}	50	20	1000
R_3	150	30	4500
R_4	200	30	6000
R_8	200	40	8000
R_9	100	82	8200

Node-3

Replica	Size	Request	CoT
R_6	80	4	320
R_{10}	75	10	750
R_{11}	100	15	1500
R_{12}	225	10	2250
R_5	250	10	2500
R_{16}	150	20	3000

Fig. 5. Replica replacement scenarios

Algorithm: Primary Copy Assignment
Input : Replica ID, context sensitive values
Output: Primary Copy Assignment
For (i) < number of Replica
Find a cluster that has the high Request Count for a specific replica, Select the node that holds the replica.
If (selected node is not a primary copy) then
Set the nodes Replica as primary copy
End if
Propagate updates across the network

Fig. 6. Assign primary copy replica algorithm

V. Discussions

We compare our algorithm with Top-K LRU algorithm in [19]. Here when there is a request for a file *j* a new replica of *j* is obtained and stored in the current first-place node for *j* and a replica of another file is evicted from the node based on LRU.

The main advantage of CAR algorithm is that the algorithm considers the CoT based on the size of the file and the number of requests. A typical scenario is illustrated in the Table I.

Assume that there are requests to replicate content in the cluster of size 40 MB and the content had met the criteria of minimum CoT for creating a replica in the cluster, further we assume that there is no free space for moving in the new replica, in this scenario replacement algorithm will be invoked. In the case of CAR algorithm the replica R_3 in *Node* N_2 will be replaced since its CoT is minimum and its size is >= 40, whereas in the case of K-Top algorithm since R_5 is LRU from the cluster, R_5 is evicted from Node N_3 to accommodate the new replica in spite of its large size and CoT. In Internet where recency differs in few seconds to minutes CAR outperforms LRU in the above scenario.

Secondly we compare our algorithm with Top-K Most Frequently Requested (MFR) Algorithm [19] where each node *i* maintains a table for all files for which it has received a request. A file *j* in the table, the node maintains an estimate of $\lambda j(i)$, the local request rate for the file. In the simplest form, $\lambda j(i)$ is the number of requests node *i* has seen for file *j* divided by the amount of time node *i* has been up.

Assume that there is a request to replicate a content of size 100 MB and also the requested content had met the criteria of minimum CoT for moving it into the cluster and further, we assume that there is no free space for moving in the new content and all nodes are up all the time. In case of CAR replacement algorithm R_5 from N_3 will be evicted, whereas in case of MFR replica R_2 from N_1 in spite of its size of 400 MB will be is evicted to replace the new content.

We compare our replica assignment algorithm against Criteria Based Primary-copy Assignment (CBPA) in [18] where a list is created on descending order based on the availability of the system the primary copy is assigned to the system which is in the top of the list. Here only the availability of the system is considered and the request from the client is not an important criterion in deciding the primary copy.

TABLE I: Replica Tracking CoT Vs Last Access

Node ID	Replica ID	Size (MB)	Request Count	CoT	Last Access
N_1	R_1	100	90	9000	10:55:10
N_1	R_2	200	70	14000	11:55:10
N_2	R_3	50	20	1000	12:55:10
N_2	R_4	400	30	12000	13:55:10
N_3	R_5	200	10	2000	2:55:10

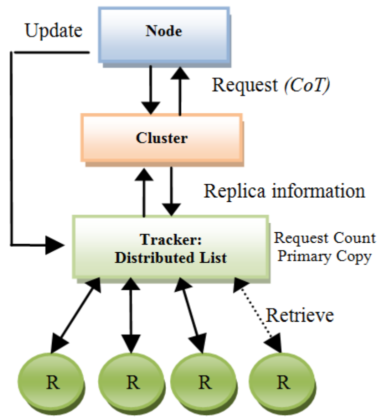

Fig. 7. The car architecture

Maintaining multiple replicated copies consistent requires substantial bandwidth and is not directly related to the availability of the system. Hence in CAR Primary copy assignment is depended on the access pattern of the replica and the size of replica. Since, the replica is initially placed based on the best bandwidth in the cluster, performance of CAR primary copy assignment is better than CBPA.

In Table III when CBPA algorithm is used to choose the primary copy it will select Replica R_1 in Node N_1 of cluster C_1 based on the availability whereas in case of CAR R_1 from Node N_1 in cluster C_2 will be chosen based on the request count, in case N_1 of C_2 fails the next popular node in the cluster for the specific replica will be selected.

TABLE II: Replica Tracking List CoT Vs Request Count

Node ID	Replica ID	Size (MB)	Request Count	CoT
N_1	R_2	400	5	2000
N_1	R_1	100	90	9000
N_2	R_3	150	40	6000
N_2	R_4	400	30	12000
N_3	R_5	100	10	1000

TABLE III: Primary Copy – CBPA (Availability)

Cluster ID	Node ID	Replica ID	Size (MB)	Request Count	Availability
C_1	N_1	$R_1[p]$	400	10	100%
C_1	N_2	R_2	100	70	90%
C_1	N_3	$R_3[p]$	150	80	100%
C_2	N_1	R_1	400	30	80%
C_3	N_1	$R_2[p]$	100	25	100%

TABLE IV: Primary Copy – CAR

Cluster ID	Node ID	Replica ID	Size (MB)	Request Count	Availability
C_1	N_1	R_1	400	10	100%
C_1	N_2	$R_2[p]$	100	70	90%
C_1	N_3	$R_3[p]$	150	80	100%
C_2	N_1	$R_1[p]$	400	30	80%
C_3	N_1	R_2	100	25	100%

[p]- Primary Replica

The above comparative study demonstrates the effectiveness of our proposed algorithm. CAR significantly reduces the file transfer overhead by placing the right replica at the right place.

VI. Conclusion and Future Work

With digital content growing exponentially and internet congestion on rise innovative ways are required to access the digital content available in the network. P2P overlay networks offers ways by which content can be widely distributed using replication. However, the challenge is where to place the replica, which replica to replace and which replica to be assigned as the primary copy. The distribution and delivery of digital content can be improved by making replica content placements, replacement and management context aware.

In this paper we have proposed a context-aware replica placement algorithm that reduces access time by placing multiple copies of content at various strategic locations using various context-aware parameters. We also proposed the replica replacement algorithm considering the Cost of Transfer (CoT) that is based on the size of the content and request rate. Moreover, we proposed a primary copy assignment algorithm that minimizes the content update overheads by choosing the right node as primary replica. Our discussions demonstrate the effectiveness of the proposed algorithms.

In future work we plan to compare the performance of the CAR replication algorithm with various other replication algorithms using a simulation study.

References

[1] N. J. A. Harvey, J. Dunagan, M. B. Jones, S. Saroiu, M. Theimer, and A. Wolman, "Skip net: A scalable overlay network with practical locality properties," *Proceedings of the 4th conference on USENIX Symposium on Internet Technologies and Systems*, vol. 4, 2009

[2] B. Awerbuch, Y. Bartal, and A. Fiat, "Distributed paging for general networks," in *The Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, Atlanta, Georgia, 1996, pp. 574-583.

[3] S. Weiss, P. Urso, and P. Molli, "Logoot: A scalable optimistic replication algorithm for collaborative editing on P2P networks," *ICDCS*, 2009

[4] J. Kangasharju, K. W. Ross, and D. A. Turner, "Optimal content replication in P2P communities," *Manuscript*, 2002.

[5] E. Cohen and S. Shenker, "Replication strategies in unstructured peer-to-peer networks," in *ACM SIGCOMM '02, Pittsburgh, Pennsylvania*, August 2002.

[6] S. Banerjee, C. Kommareddy, K. Kar, B. Bhattacharjee, and S. Khuller, "Construction of an efficient overlay multicast infrastructure for realtime applications," in *IEEE INFOCOM 2003*, 2003.

[7] X. Tang and J. Xu, "On replica placement for QOS-aware content distribution," *in IEEE INFOCOM 2004*, 2004.

[8] D. Doval and D. O'Mahony, "Overlay networks: A scalable alternative for P2P," *IEEE Internet Computing*, vol. 7, no. 4, pp. 79-82, July/Aug. 2003

[9] H. Shen, "EAD: An efficient and adaptive decentralized file replication algorithm in P2P file sharing systems," *Eighth International Conference on Peer-to-Peer Computing*, pp. 99-108, 2008

[10] E. K. Lua , J. Crowcroft , M. Pias , R. Sharma, and S. Lim, "A survey and comparison of peer-to-peer overlay network schemes," *IEEE Communications Surveys*, 2005

[11] L. Ma, H. Shen, and Q. Zhang, "The key technologies for a large-scale real-time interactive video distribution system," *ICACC*, 2010

[12] A. M. K. Pathan and R. Buyya, "Economy-based content replication for peering content delivery networks," *CDN-CCGrid*, 2007

[13] G. Hofmann and G. Thomas, "Digital lifestyle 2020," *IEEE MultiMedia*, vol. 15, no. 2, pp. 4-7, Apr.-June 2008

[14] L. Qiu, V. N. Padmanabhan, and G. M. Voelker, "On the placement of Web server replicas," *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3, pp. 1587-1596, 2001.

[15] J. W. Kim, S. W. Han, D.-H. Yi, N. Kim, and C.-C J. Kuo, "Media-oriented service composition with service overlay networks: challenges," *Approaches and Future Trends Journal of Communications*, vol. 5, no. 5, pp. 374-389, May 2010.

[16] M. Kucharzak and K. M. Walkowiak, "File sharing-based heuristics for flow assignment in p2p systems," *Logistics and Industrial Informatics, 2009. LINDI 2009. 2nd International*, pp. 1-6, 10-12 Sept. 2009.

[17] Ramachandran and Sikdar, "A queuing model for evaluating the transfer latency of peer-to-peer systems," *IEEE Transactions on Parallel and Distributed Systems*, March 2010

[18] S. Abdalla, I. Ahmad, E. H. Tat, G. Aik, and Y. L. Kee, "Towards achieving a highly available distributed file system," *The 9th ICACT*, 2007.

[19] J. Kangasharju, K. W. Ross, and D. A. Turner, "Adaptive content management in structured P2P communities," *ACM International Conference Proceeding Series*, vol. 152, 2006

[20] W. K. Y. Lin and C. Chiu, "Decentralized replication algorithms for improving file availability in P2P networks," *IEEE IWQoS*, 2007

[21] M. Srivatsa, B. Gedik, and L. Liu, "Large scaling unstructured peer-to-peer networks with heterogeneity-aware topology and routing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 17, no. 11, pp. 1277-1293, Nov. 2006.

Mohammad H. Al Shayeji received his B.Sc. (Eng), from University of Miami, and M.S. (Computer Science) from University of Central Florida. He got his Ph.D. in the field of Computer Science and Engineering from University of Southern California. Currently he is working as Assistant Professor in Computer Engineering Department, College of Engineering and Petroleum, Kuwait University. He has several publications in different international Journals and Conferences. His research interest includes VOD, Video Servers, Multimedia DMS, and Distributed Systems.

M. D Samrajesh received his B.Sc, from Bharathiar University, and MCA from Bharathidasan University. He got his M.Phil in the field of Computer Science from MS University. He is a member of IACSIT, IAENG, CSI. He has many publications in various national and international Conferences and Journals. His research interest includes Software Engineering, Distributed Systems, and Video-on-Demand

Saud Abdulaziz Al-Behairy and Khalid Assaf Al-Enazi are graduate students at Computer Engineering Department, Kuwait University.