ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

A FUZZY BASED APPROACH FOR EFFECTIVE DATA REPLICATION IN PEER-TO-PEER NETWORKS

Anju Rani[1], Rahul Kumar Yadav[2]

Computer Science and Engineering Deptt, PDM College of Engineering, Bahadurgarh, Haryana[1,2]

Abstract: A P2P network is one of the most usable local area network that provides the interconnectivity between the nodes. But as the communication increases over the network, the load on some centralized nodes or the server is increased. In such case, there is the requirement of setup some sub system over the network to distribute the network load. Such sub systems are called replication servers. These systems can have full or the partial copy of the actual centralized server. The main problem in such systems is to find optimal number of replication servers required in the system. The presented research is about to identify the required number of such sub systems over the network. In this work, a statistical analysis is been presented based on the network capabilities, network distribution and the node requirements.

Keywords: P2P, Replication, Distribution, Optimization, Statistical Analysis

I. INTRODUCTION

A Storage Area networking has generated tremendous interest worldwide among both Internet surfers and computer networking professionals. In such kind of network, heavy data is been stored in the peered networks. This kind of network gives effective distribution of large amount of data over the network. Such kind of technology is been improved effectively over the period of time. SAN is one of the major current and the future technology to perform the equalized distribution of data over the network.

In such kind of network each node is having its on capabilities as well the responsibilities so that a client server network works as the peer to peer network. This peered network architecture provides the dedicated services to the user. Some of these nodes work as the dedicated nodes that can perform one kind of work distribution over the network. But as of other centralized system, this kind of network system also suffers the problem of overloading. In the overload conditions, a server suffers the problem of multiple queries that are required to handle in optimized time frame.

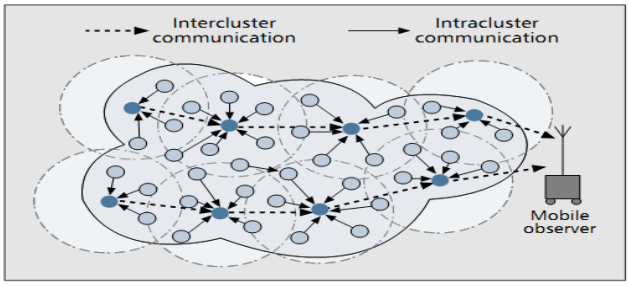

Figure 1: Storage Area Network

Here figure 1 is showing a typical SAN network. Here the light blue nodes are the requesting nodes and the dark blue nodes represent the centralized servers or the replication servers.

The SAN overlay network consists of all the participating peers as network nodes. There are links between any two nodes that know each other: i.e. if a participating peer knows the location of another peer in the SAN network, then there is a directed edge from the former node to the latter in the overlay network. Based on how the nodes in the overlay network are linked to each other, we can classify the SAN networks as unstructured or structured.

A) Structured Storage Area systems

Such kind of network system is globally consistent that ensures the communication over the network. The network is able to identify the node respective to the user required file query. It performs a pattern based search over the network by using the Hash Table or some other algorithm concept to search the file effectively. In most of the peer to peer network, such kind of network construction or the structured architecture is been used.

B) Unstructured Storage Area systems

An unstructured storage area network is used to identify the overlay links. Such kind of systems uses the concept of replicating data to the other nodes so that the search will be done effectively. In these networks, as the search is been performed, the request message is broadcast

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

to all the peers and identify the list of peers that can provide the desired data. The major drawback of such system is the concept of flooding the request. That increases the overall communication over the network, so that the network goes slow as the queries increases over the network. If there is peer, looking for rare data shared only with few peers, it can be performed effectively in such network.

There also exist hybrid SAN systems, which distribute their clients into two groups: client nodes and overlay nodes. Typically, each client is able to act according to the momentary need of the network and can become part of the respective overlay network used to coordinate the SAN structure. This division between normal and 'better' nodes is done in order to address the scaling problems on early pure SAN networks. Examples for such networks are for example Gnutella (after v0.4) or G2.

In SAN networks, all the available nodes are fully capable with some defined features. These features include the available space, bandwidth and the computing power. As the number of requested queries over the system increases, the demand of different resources increases so that to improve the efficiency of system, the capacity of the system is required to increase. In this work, an improvement over the system is been defined to setup some sub systems over the network so that network effectiveness will be increased.

II. LITERATURE SURVEY

The earlier researches performed lot of work on Storage Area network as well as on the replication servers. There are number of replication methods proposed by many researchers. These available techniques provide the higher degree of network maintainability and the reliability to provide the distribution of the data over the network effectively. In this section some of the work already done by different researchers is discussed. One of the major replication based work is perfomed by Evjola Spaho. The author presented the fuzzy based approach to improve the replication over the P2P network. The author has presented an effective distribution of the computational burden over the network [1]. The author also performs the simulation of the work in an open environment. Another work is performed by Shyam Antony in P2P system by performing the sharing of desirable features under the effective of scalability and the maintenance of the system. The author has defined a consistent clustering approach perform the distribution. The author presented a protocol to share the data over the network [2].

In Year 2009, Mesaac MAKPANGOU proposes a large-scale peer-to-peer database hosting system capable to efficiently replicate and manage databases accessed worldwide. This paper focuses on the system architecture and how this architecture is deployed over a P2P network. The author argues that this database replica hosting system will boost the performance of database-intensive applications accessed by clients that are distributed world-wide [3]. Qin Lv performed a work on data replication strategy, and network topology. The author proposes a query algorithm based on multiple random walks that resolve queries almost as quickly as Gnutella's flooding method while reducing the network traffic by two orders of magnitude in many cases [4]. In Year 2009, Joshua Reich presented a work on opportunistic network. In this work the author shows that an optimal allocation can be efficient computed or approximated. As users become increasingly impatient, the optimal allocation varies steadily between uniform and highly-skewed towards popular content. Moreover, in opportunistic environments, the global cache state may be difficult or impossible to obtain, requiring that replication decisions be made using only local knowledge. Author develop a reactive distributed algorithm, Query Counting Replication (QCR) that for any delay-utility function drives the global cache towards the optimal allocation - without use of any explicit estimators or control channel information [5].

In Year 2008, Bin Cheng proposes and evaluates a framework for lazy replication. Lazy replication postpones replication, trying to make efficient use of bandwidth. In Presented framework, two predictors are plugged in to create the working replication algorithm. Lazy replication with several predictors is compared with a naive eager replication algorithm. The author finds that lazy replication is more efficient than eager replication, even when using two simple predictors [6]. In Year 2010, Sanaullah Nazir presents a replication strategy to improve data availability in P2P Networks. The focus of the paper is to replicate data to nodes which are highly available and complement one another in terms of uptimes. In Presented evaluation Author show that a life pattern along with the availability of nodes improves overall data availability [7]. Geunyoung Park presents Chordet, which is an efficient and transparent replication scheme for Chord-based P2P networks. Chordet replicates the data stored in the participating nodes so that they are evenly distributed. Using Chordet, the nodes can reduce lookup failure rates and lookup path length. Moreover, not all nodes need to implement proposed replication algorithm [8]. Leonard Barolli proposes a fuzzy-based system, which improves the QoS of MANETs via data replication. In this paper, Author use fuzzy logic and build a system which has three input linguistic parameters and one output linguistic parameter. The author evaluates by simulations the proposed system. The simulation results show that the proposed system has a good decision [9]. Anna Saro Vijendran proposes a new popularity based

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

QOS(Quality Of Service)-aware smart replica placement algorithm for content distribution in peer to-peer overlay networks which overcomes the access latency, fault tolerance, network traffic and redundancy problems with low cost. The paper also describes briefly the literature survey of the existing algorithms and their merits and demerits [10]. Manu Vardhan presents a threshold based file replication model that replicates the file on other servers based on the number of file accesses [11].

Jian Zhou proposes an On-line Pointer1 Replication (OPR) algorithm in structured P2P networks which yields significantly low worst case query latency. Also, the degree of replication achieved by OPR is dynamically adaptable to the instantaneous query arrival rate and churn characteristics of the system in order to reduce total control traffic [12]. Saurabh Tewari complements those results to show that this distribution has network-wide advantages as well. Given these benefits of proportional replication, the next issue is achieving proportional replication in a decentralized manner [13].

III. PROPOSED WORK

The proposed work is about to define a network model for a distributed P2P system so that the replication over the network can be done in an easy and effective way. The basic proposed model based on the concept of cost estimation. The model consists of following characterstics:

- Multiple access points are connected wirelessly to each other.
- The location of all access points is predefined and constant.
- Only part of the access points have a physical link to the Internet, and thus act as gateways.
- Multiple mobile clients may connect to the Internet through the gateways. At each time point, a client is connected to one gateway.
- The connection to the gateways is either direct or through a series of forwarding access points.
- Clients may switch gateways and/or routes (access-points) dynamically throughout the simulation. The assignment of a client to a gateway is done by a nomadic service assignment algorithm.

As the network model is constructed, it is under different cost estimation over the network. These cost based framework is having the following components given as under

A) Setup Cost

The setup cost is a onetime cost paid when a client switches from one gateway to another, and represents the cost to setup a new connection with the new gateway. In this project, we assume that the setup cost is constant and identical to all clients and gateways at all times, and is equal to losing all of the files in the timeslot.

B) Access Cost

The access cost is a contiguous cost, paid for each timeslot in which the client is connected to a certain gateway. It represents the quality of the connection between the client and the gateway, and is measured using the percentage of data access in each slot. The data access is largely influenced by the number of hops between the client and the gateway: each additional hop (access point) usually increases the packet loss, and therefore worsens the connection and increases the hold cost.

The Access cost is usually small, but its affect on the total average cost increases with time, since access cost associated with maintaining a bad connection for a long time accumulate to a large amount. The average total cost per node, which is defined as the sum of the total hold and setup costs for the users in the group[1], divided by the number of users (N) in the group.

C) Algorithm

This topology definition algorithm is defined as follow:

- N defines number of Nodes.
- Randomly place the nodes inside 1x1 box.
- (0, 0), (1, 1) are defines as the source and target nodes.
- Calculate the distance between each pair of nodes (d<=sqrt(2)).
- Place edge between each pair(I,J) with probability of
- A, B are parameters which can help control the number of edges in the Graph. A increase the number of edges linearly and B exponentially.

After building the topology we added delay and capacity values to the existing edges. We used random values (within a reasonable range) for both the delay and capacity as we thought it will model a "real" network the best. Since our simulations run for one source and one target the topology we use should reflect a network state at a certain given time, which mean the capacity and delay values are actually reflecting a given state of the network on which we try to add the flow from source to target. Since this state can be any state, we chose to randomly pick these values.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

IV. RESULTS

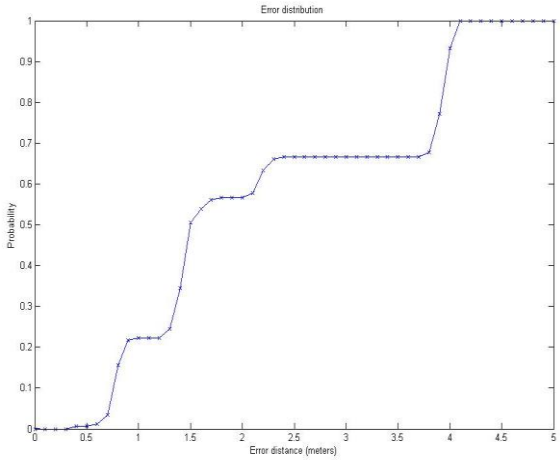

Figure 2: Error analysis (Network Architecture 1)

Figure 2 is showing the effect of distance of nodes from the server. The result is presented respective to architecture 1. In this figure, as the distance from the server is increased, the error rate in the system is also increased. It shows the requirement of the replication server in the system. Distance is presented as crucial factor for deciding position to position and distance from the server.

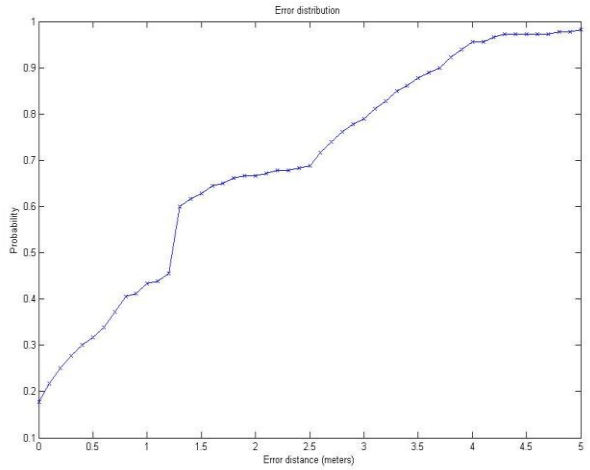

Figure 3: Error Analysis (Network Architecture 2)

Figure 3 is showing the effect of distance of nodes from the server. The results are taken based on the Network architecture 2 in the system. In this figure, as the distance from the server is increased, the error rate in the system is also increased. It shows the requirement of the replication server in the system. Distance is presented as crucial factor for deciding position to position and distance from the server.

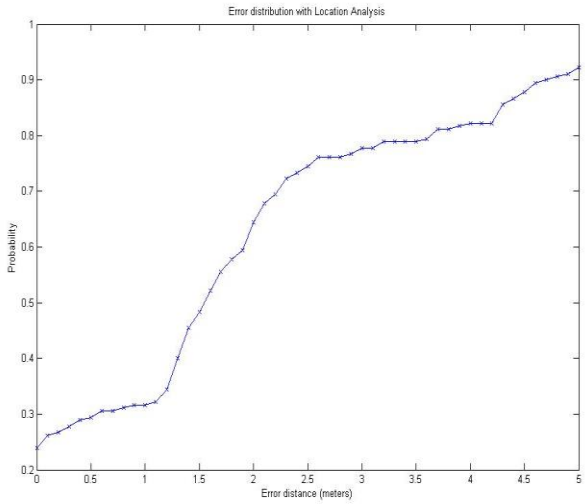

Figure 4: Error Analysis (Network Architecture 3)

Figure 4 is showing the effect of distance of nodes from the server. The results are taken based on the third Network architecture in the system. In this figure, as the distance from the server is increased, the error rate in the system is also increased. It shows the requirement of the replication server in the system. Distance is presented as crucial factor for deciding position to position and distance from the server.

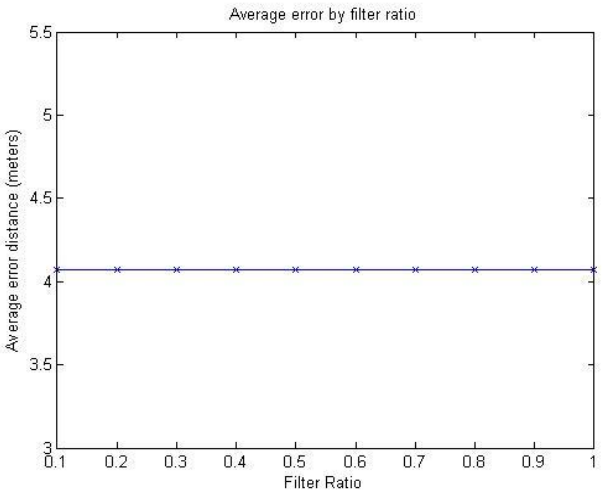

Figure 5: Average Error Analysis (Network Architecture 3)

Figure 5 is showing the average error analysis of different network architectures. It is showing effect of distance of nodes from the server. In this figure, as the distance from the server is increased, the error rate in the system is also increased. It shows the requirement of the

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

replication server in the system. Distance is presented as crucial factor for deciding position to position and distance from the server.

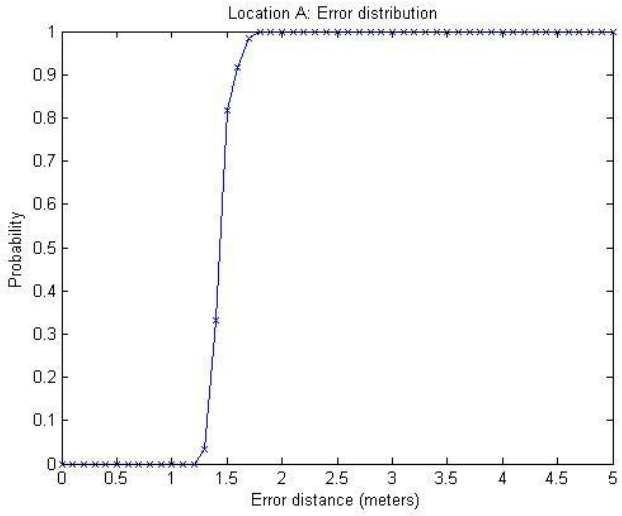

Figure 6: Error Analysis with Replication Server (Network Architecture 1)

Figure 6 is showing the effect of implementing the replication server in distance of nodes from the server. The results are taken based on the first Network Architecture in the system. In this figure, as the distance from the server is node controlled by the replication server. So that if the distance from the main server is increased, even then the error rate is not much increased.

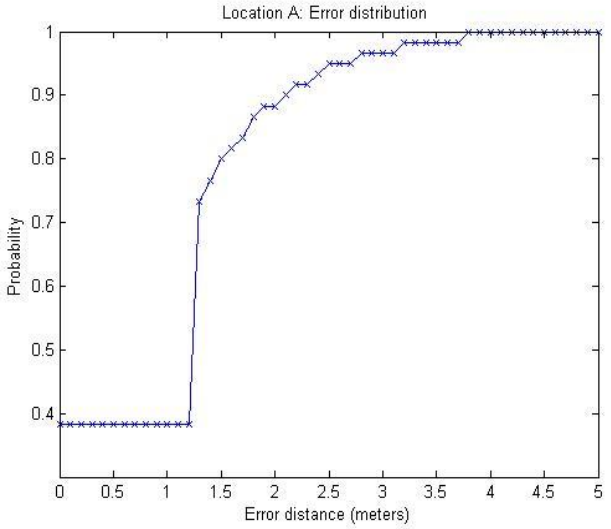

Figure 7: Error Analysis with Replication Server (Network Architecture 2)

Figure 7 is showing the effect of implementing the replication server in distance of nodes from the server. The results are taken based on the second Network Architecture in the system. In this figure, as the distance from the server is node controlled by the replication server. So that if the distance from the main server is increased, even then the error rate is not much increased.

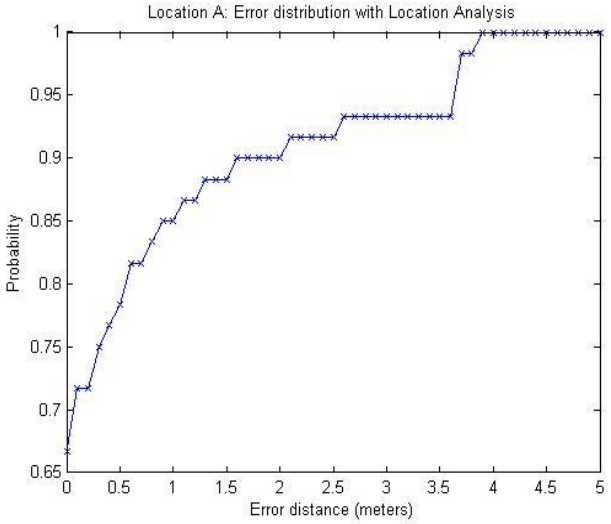

Figure 8: Error Analysis with Replication Server (Network Architecture 3)

Figure 8 is showing the effect of implementing the replication server in distance of nodes from the server. The results are taken based on the third Network Architecture in the system. In this figure, as the distance from the server is node controlled by the replication server. So that if the distance from the main server is increased, even then the error rate is not much increased.

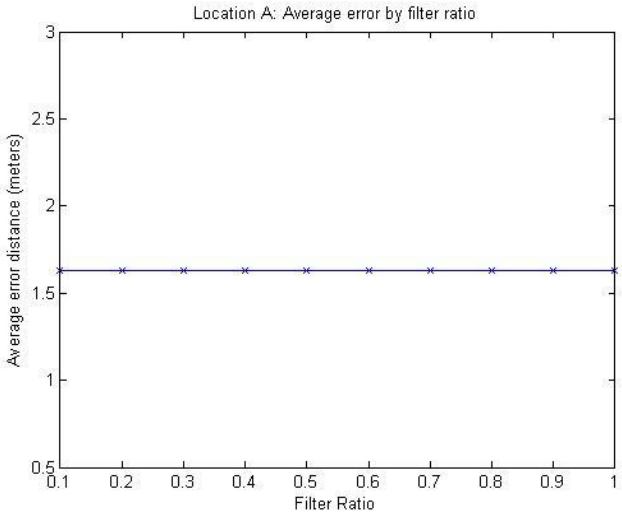

Figure 9: Average Error Analysis with Replication Server

Figure 9 is showing the effect of implementing the replication server in distance of nodes from the server. The results are taken based on the all Network Architectures in

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

the system. In this figure, as the distance from the server is node controlled by the replication server and the error rate probability is decreased with the inclusion of replication server.

V. CONCLUSION

The proposed work is about to define a network replication system so that the distribution of data over the network will be performed. The work is about to define a parametric algorithmic approach that will analyze the cost of setup of the network as well as the communication cost based on which the estimation of the replication node placement will be decided.

.

REFERENCES

[1] Evjola Spaho," A Fuzzy-based System for Data Replication in P2P Networks", 2011 Third International Conference on Intelligent Networking and Collaborative Systems 978-0-7695-4579-0/11 © 2011 IEEE (pp 373-377)
[2] Shyam Antony," P2P Systems with Transactional Semantics", EDBT'08, March 25–30, 2008, Nantes, France. ACM 978-1-59593-926-5/08/0003 (pp 4-15)
[3] Mesaac MAKPANGOU," P2P based Hosting System for Scalable Replicated Databases", DAMAP 2009, March 22, 2009, Saint Petersburg, Russia. ACM 978-1-60558-650-2 (pp 47-54)
[4] Qin Lv," Search and Replication in Unstructured Peer-to-Peer Networks", ICS'02, June 22-26, 2002, New York, New York, USA ACM 1-58113-483-5/02/0006 (pp 84-95)
[5] Joshua Reich," The Age of Impatience: Optimal Replication Schemes for Opportunistic Networks", CoNEXT'09, December 1–4, 2009, Rome, Italy. ACM 978-1-60558-636-6/09/12 (pp 85-96)
[6] Bin Cheng," A Framework for Lazy Replication in P2P VoD", NOSSDAV '08 Braunschweig,Germany ACM 978-1-60588-157-6/05/2008 (pp 93-98)
[7] Sanaullah Nazir," Using Monte Carlo simulation for improving Data Availability in P2P network", IDEAS10 2010, August 16-18, Montreal, QC [Canada] Editor: Bipin C. DESAI ACM 978-1-60558-900-8/10/08 (pp 179-185)
[8] Geunyoung Park," Chordet: An Efficient and Transparent Replication for Improving Availability of Peer-to-Peer Networked Systems", SAC'10, March 22-26, 2010, Sierre, Switzerland. ACM 978-1-60558-638-0/10/03 (pp 221-225)
[9] Leonard Barolli," A Fuzzy-based Data Replication System for QoS Improvement in MANETs", MoMM2012, 3-5 December, 2012, Bali, Indonesia ACM 978-1-4503-1307-0/12/12
[10] Dr. Anna Saro Vijendran," Survey of Caching and Replica Placement Algorithm for Content Distribution in Peer to Peer Overlay Networks", CCSEIT-12, October 26-28, 2012, Coimbatore [Tamil nadu, India] ACM 978-1-4503-1310-0/12/10 (pp 248-252)
[11] Manu Vardhan," A Demand Based Fault Tolerant File Replication Model for Clouds", CUBE 2012, September 3–5, 2012, Pune, Maharashtra, India. ACM 978-1-4503-1185-4/12/09 (pp 561-566)
[12] Jian Zhou," An Effective Pointer Replication Algorithm in P2P Networks".
[13] Saurabh Tewari," Proportional Replication in Peer-to-Peer Networks".