International Journal of Database Theory and Application
Vol.9, No.10 (2016), pp.419-428
http://dx.doi.org/10.14257/ijdta.2016.9.10.36

Design and Research on the Special System Architecture for Peer-to-Peer Distributed Storage System Based on Node Grouping

Cai Liang[1]* and Huang Hao[2]

[1] Hubei University of Automotive Technology, Shiyan, China

[2] Hunan University, Changsha, China
** CAI Liang, 87240898@qq.com*

Abstract

The organization pattern of P2P (Peer-to-Peer), has become an important form of internet application of new generation because of its such features, as good expansibility, fault tolerance and high-performance. The persistent data storage is the key to inhibiting the development of global storage system, and is also difficulty issues of the research. This paper describes a design idea and a system architecture of distributed storage system established on P2P in detail, and proposes an effective mechanism according to resource topic and predicted network distance for node grouping, and verifies the persistent data storage performance of P2P global storage system.

Keywords*: P2P storage system, node grouping, resource topic, network distance*

1. Introduction

At present, P2P storage system has become an important researching field. Constructing storage system on P2P network has many advantages. It can effectively make use of the network bandwidth, storage space and data resources between the nodes. In addition, the system has high extensibility, data availability and reliability. The aim of P2P system is to organize the nodes in the system to store data. In order to improve the effectiveness and to improve the effective management of nodes towards the storage resources, this paper proposes a method to manage nodes by grouping according to P2P network storage resource topic and predicted network distance, and finally to form a layered overlay hybrid P2P storage system.

2. Related Works

At present, the node management based on P2P storage system can be mainly divided into two ways: centralized and decentralized. Typical centralized system Napster [1] utilizes centralized server to take charge of service of catalog management. Due to the limitation of the server, there often exist the problems of damage of systematic topological structure, low availability of node storage data and the decrease of reliability of node storage data due to single node failure. While typical decentralized system Gnutella [2] and Freenet [3], due to the lack of centralized server, massages will be spreaded on the internet by means of flooding when querying data, and exists the problems of taking up a lot of system bandwidth, blizzards of messages and poor system extensibility.

All printed material, including text, illustrations, and charts, must be kept within the parameters of the 8 15/16-inch (53.75 picas) column length and 5 15/16-inch (36 picas) column width. Please do not write or print outside of the column parameters. Margins are 3.3cm on the left side, 3.65cm on the right, 2.03cm on the top, and 3.05cm on the bottom. Paper orientation in all pages should be in portrait style.

ISSN: 2005-4270 IJDTA

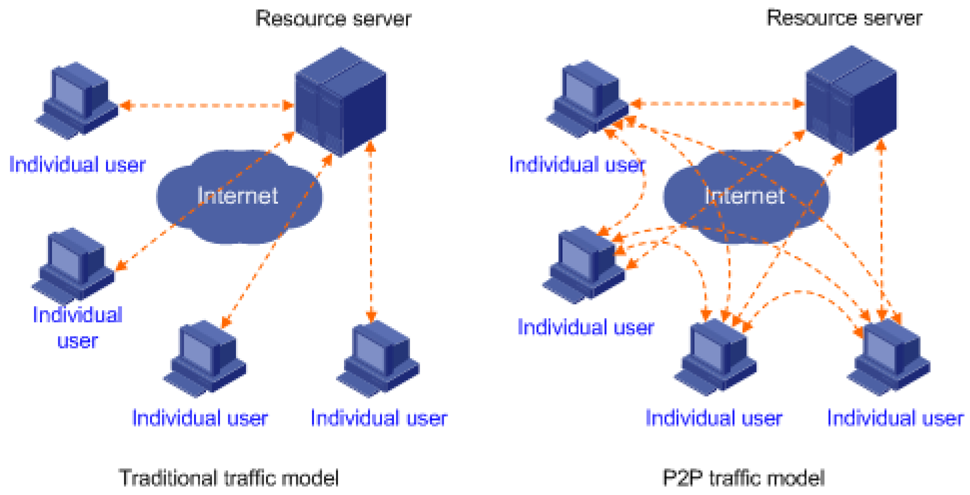


Figure 1. P2P Storage System

3. The Node Grouping Algorithm and Data Storage Mechanism

3.1. Algorithm Based on Storage Resource Topic and Predicted Network Distance

In real network, the storage resource contained in each node have specific topic. Supposing the topic set of all the storage resources in P2P network is T {Ti |1 i n, n N} , where n is the maximum number of possible topics; each topic Ti contains P = {Pj |1 j m, m N} subtopics, where m is the maximum number of possible subtopics in Ti . | T | is the number of topics of network storage resources, and | Pj | is the number of subtopics contained in Ti topics. | T | is the benchmark of network grouping number, and | Pj | is benchmark of the number of super nodes in each group.

In order to improve the storage space allocation rate of P2P storage system and routing efficiency when storing data, and to make network nodes distributed more uniformly in each group, the method of describing circle with predicted radio in applied to node grouping. The specific means is as follows: randomly take a node k in large-scale network, and judge the network data transmission delay among k and its adjacent nodes. Supposing there are n adjacent nodes, the network available bandwidth among k and these n adjacent nodes are BIT1 , BIT2 . . . BITn , supposing the size of transmitting data blocks areDATAk (kb) . The network data transmission delay among k and n adjacent nodes are $TIME_i = DATA_k / BIT_i, (i = 1, 2...n)$, Then, take the averrage value of network data transmission delay $D_1 = (\sum_{i=1}^{n} TIME_i)/n = (\sum_{i=1}^{n} DATA_k / BIT_i)/n, (i = 1, 2...n)$. By this analogy, take the average values of network data transmission delay of m random nodes $D_1, D_2 ... D_m$ .We suppose the predicted value of network distance is the average value of the mean value of network data transmission delay obtained for each sample node, namely, the predicted network distance is $dis = (\sum_{i=1}^{m} D_i)/m$. After obtaining the predicted network distance, randomly take one node from the initial network, and use this node as the center, put all the nodes in the radius of predicted network distance in a group, then remove the nodes from initial network after obtaining the group. Grouping for | T | times by this means, the nodes in initial network can be divided into | T | groups. For each group Ti , select | Pj | nodes in the group as super nodes(considering the complexity of large-scale network, we mainly select nodes with long online time). The super nodes connect with adjacent normal nodes to form chord structure[4], and super nodes also form chord structure with other super nodes. In order to guarantee the connectivity in among the groups, take a super node in each group and connect them to form net structure, as shown in the Figure 1. Then, a three-layer overlay network is formed. In order to maintain the integrity of the network topological structure, for each super node in the group, select an adjacent node as its backup node; for the each super node which connects each group, select two adjacent nodes as its backup nodes. For each super node and its backup node, save the location information of two previous and subsequent nodes to guarantee that the loop will not disconnect when node failure happens.

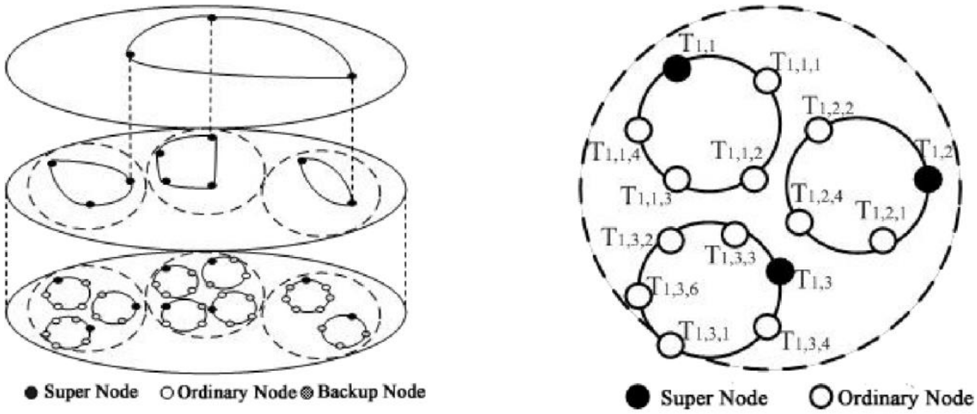


Figure 2. Three Layer Overlay Network and Node Naming Diagram

3.2. Node Naming and Data Storage Mechanism

Because the mechanism in which overlay network finally forms for P2P network node grouping is applied, the distance among the nodes in each group is similar. In order to make full use of bandwidth and to lower routing overhead when storing data, multiplex naming mechanism of system nodes according to storage resource topic is applied.Namely, for $|T|$ groups in the system, each group is firstly named T_i $(1 \quad i \quad |T|)$. Then, each super node in the group is named $T_{i,j}$ $(1 \quad i \quad |T|, 1 \quad j \quad |P_j|)$. according to its subtopic number. Finally, according to chord rule[4], the identifier of the normal nodes connected with each super node is named using Hash function, as is shown in Figure 2. Because the name of each group and the corresponding identifier of the node in each group are different, the name of each node in the system will be unique. In order to lower the system overhead when querying storage location of the data, the location information of the super nodes in the third layer will be disclosed. In order to avoid overload of the super nodes in the third layer, they will not store data, only take charge of maintaining the information table of its connecting nodes. When querying data storage, just by knowing the topic of the storage data, according to the similarity of the topics and the information maintained by the super nodes on the third layer, the group for the storage and the subtopic super node storing the node information on the second layer can be determined. Finally, by querying the information table of the normal node maintained by subtopic super node, map the filename of the storage data according to Hash function, and the nearest node of the identifier can always be found so that data is stored on this node. When there is node needed adding to the system, firstly calculate the network distance among this node and the super nodes in each group. Add it to the group with the minimum average distance. Then, connect to the nearest super node ring in the group and register its information on this super node. When node failure or exiting system happens, delete the information of this node directly, and then connect its previous and subsequent nodes. When super node fails, use backup node to replace its location immediately.

4. Data Backup and Storage Space Recovery Mechanism

4.1. Data Backup Mechanism

The objective of data backup is to improve the availability and reliability of the data in P2P storage system, which prevents data loss because of node failure in the system[5]. The statistic analysis in reference[6] indicates that in specific group subnetwork, the repetition rate in data topic querying process is very high, which is because there exist some nodes with high access frequency on which some hotspot topics are stored. The backup of these hotspot data topics with high access frequency will make a great difference to the improvement of the availability and reliability of the the whole system. Therefore, this paper applies different data backup strategies according to the difference in the topic access frequency. For the topic data with relatively low access frequency, the method of intragroup backup is applied. The specific means is as follows: in the nodes that need backing up, randomly and uniformly select n(this parameter is the number of replications, which needs to be predicted by users) intragroup nodes. Create the replications of the data on theses nodes, and write the information into information table of the super nodes connected with these nodes. For topic data with high access frequency, the method of intragroup backup combined with intergroup backup is applied. Namely, while intragroup backing up, selectively backup the data to other nodes in other groups and then write the backup information into information table of the super nodes connected with these nodes. Because the overlay layered super node management topological structure is applied in storage network, when there is node storage data failure, the system can find the needed backup data information by querying super node information table. Of the two methods, intragroup backup method have higher disaster resistance than intragroup backup, and it is more suitable for the storage of important data information in large-scale storage system, and thus effectively guarantee the complete repairability of the data with high access frequency, and will improve the performance of the system without increasing the overburden on the system.

4.2. Recovery of Storage Space

When node failure or existing system happens, the system will require corresponding nodes to delete the information saved on them. But part of the nodes may not be online when deleting the data, so that the corresponding information on these nodes is not deleted completely, which will produce storage garbage and result in the waste of storage space. For this purpose, on each node, for each storage data, attach a record field for last access time property. For each data whose last access time limit is out of certain range, nodes will have the right to delete it. Therefore, as time goes on, the data which are not visited will be gradually deleted and thus system resource is saved.

5. Performance Analysis

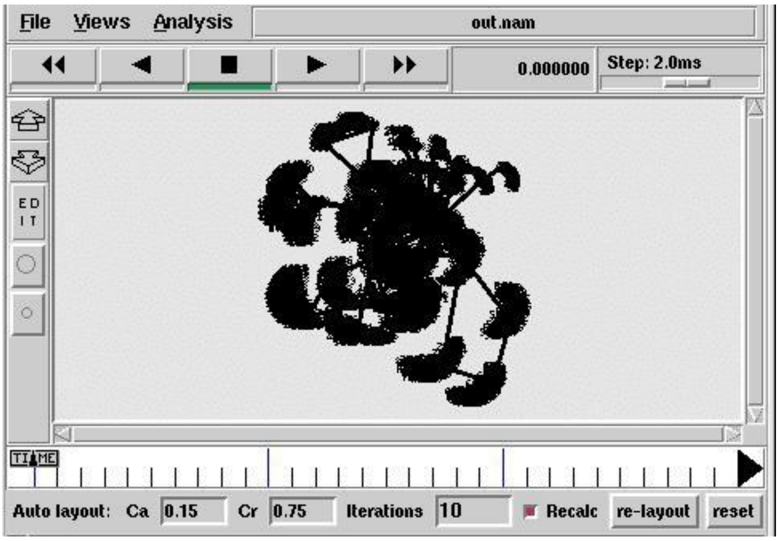
5.1. Evaluation Indexs

All P2P wide-area storage systems are based on structured P2P overlay network, which is a kind of organization method to maintain the connection of nodes in the application layer. It can connect the nodes in the system according to certain logical topology, and make it possible for any two nodes to intercommunicate by route message. In the situation in which node dynamically enters or exits the system, structured P2P overlay network needs to guarantee the interconnectivity of the nodes. MIT designed Chord algorithm is a well-known structured network algorithm, which applies one-dimensional annular structured nodes. This algorithm is simple and clear, and has very high application value. However, in Chord system, the nodes which are adjacent logically may not be adjacent in physical location (it is likely that the nodes which are adjacent logically may be far apart in physical location), so that in large-scale network system, when there are nodes entering or exiting the system frequently, there may exist many problems, such as the decrease of routing efficiency of the system, the decrease of the consistency and reliability of node stored data and excess number of messages handled by nodes in unit time. In the composite-structured system designed in this paper, the interconnection among the supernodes in the second layer in overlay network applies the organization method similar to Chord, which is similar to the characteristics of Chord routing model that for the nodes in the system, the routing information of other nodes needs not be saved, but we only need to save the message of routing items. But the algorithm designed in the paper divides the nodes which are close in network distance into the same group to be managed by supernodes, which makes the nodes which are adjacent logically close in physical position (the network distance), in order to improve the storage property of the system. Therefore, the property evaluation for this paper will be mainly compared with Chord algorithm.

In storage system, the routing querying hop for data storage positioning is of great significance for measuring the routing efficiency and time overhead of the storage system [7]. And the average number of message handling when node entering for exiting the system is of great significance for measuring the overhead for maintaining the topology and storing data [7]. When nodes in the system fail, the data stored on it will be inevitably damaged. At this time, the lost data need to be supplemented by repairing, which is the essential measure to maintain the durability of data storage [8]. The average timed needed for repairing the damaged data will directly influence the reliability and availability of the system data storage. If the average time needed for repairing is short, the reliability and availability of the system data storage will be better. Therefore, the average time needed for repairing when node fails is also an important indicator for measuring the durability of system data storage. The property of storage system designed in this paper will be mainly measured by three indicators of the routing querying hop for data storage positioning, the average number of message handling when node entering for exiting the system and the average time needed for repairing when node fails.

5.2. Simulation Results and Analysis

Under the processor Intel Core 2 Duo 1.73G and the environment of Linux Redhat 9.0, we used network simulator NS-2 to simulate the actual condition of network, realized chord and text algorithm respectively by Otcl and C++ program design and used GT-ITM topology generator to generate the typology graph of network. In order to simulate the actual network situation as possible as we can, we combined the literature [9] conclusion and the current network situation to set bandwidth which the nodes had to four levels. Among these, about 5% of the nodes which simulated the advanced users had the bandwidth of the 100Level. 15% of the nodes had the bandwidth of the 10Level. 60% of the nodes which simulated most of the users of the DSL Internet had a bandwidth of 1Level. 20% of the nodes which simulated dial-up users had the bandwidth of the 0.1Level. The node of each in the time line determined randomly values, the online unit of time is one hour. Considering respectively 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000nodes in the network, we divided it into groups by taking impact factors and data block ri =1, DATA K =10000 (kb). Each group took 50 nodes to select super codes and backup codes according to impact factors m1 = 0.5, m2= 0.5 and the formula $Snodei$ $m1$ $timei$ $m2$ Bi (Bi is the value of Level). In order to obtain more accurate simulation results, the system of different number of nodes carried out experiment repeatedly for 20 times and got the average value. The simulation network topology of three layer generated in accordance with the above parameters was shown in Figure 3.



Auto layout: Ca 0.15 Cr 0.75 Iterations 10 Recalc

Figure 3. The Simulation Network Topology of Three Layer

Considering the performance requirement of the storage system, we will determine the route hops that are needed for the location of storing data and the average number of massages when nodes enter or exit the system as important factors to measure the system performance.

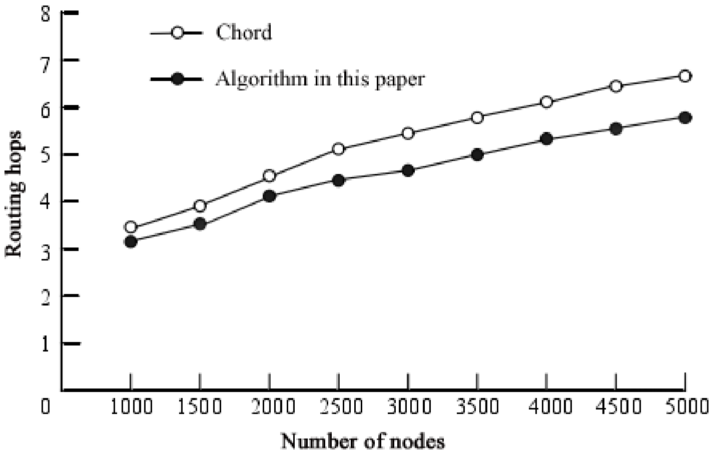


Figure 4. Data Storage Querying Comparison Result

In order to test its performance, the initial group number is set 32. The specific data storage operation is set the space range data storage querying. Randomly select an intergroup super node as the initiation node of data storage querying, and the performance monitor collect hop-measuring parameters. Aim at different querying range and node number to conduct the experiment for many times, and record the results respectively and then take the average value. Figure 4 provides the data storage querying comparison result of the method designed in this paper and that in Chord system model on the condition of different node numbers. The result indicates that when node number is relatively low, the effects of the two methods are nearly the same. With the increase of node number, the method provided in this paper will have better effect on decreasing hops.

Table 1. The Route Query Hops

Number of nodes	Chord	Algorithm in this paper
1000	3.502	3.213
1500	3.916	3.501
2000	4.523	4.148
2500	5.175	4.256
3000	5.505	4.521
3500	5.826	4.932
4000	6.111	5.123
4500	6.415	5.521
5000	6.806	5.808

Table 1 shows that when the number of nodes is less (less than 1500), the algorithm in this section, compared to Chord algorithm, improves little for the route query hop of data storage and location with minor difference of average route query hops. This article argues that when the number of nodes is small, the impact on route query performance of Chord system by the performance of system nodes themselves and the physical location of the nodes is minor, thus the routing overhead is not complex. However, when the number of nodes is larger (the number is more than 3000), compared to Chord algorithm, the algorithm in this section has certain improvements on the route query hop of data storage and location. This article argues that when the node number is larger, in the designed system of this section, the node number within each group is far less than the one in Chord system and the network distance between nodes is shorter; the routing process searching between each node under the management of supernodes is completed internally within the groups so as to avoid bad performance of node itself existed in Chord system and detouring problem (mainly caused by the physical location between nodes). Thus, the efficiency to find the target node can be improved and the complexity of system routing overhead can be reduced.

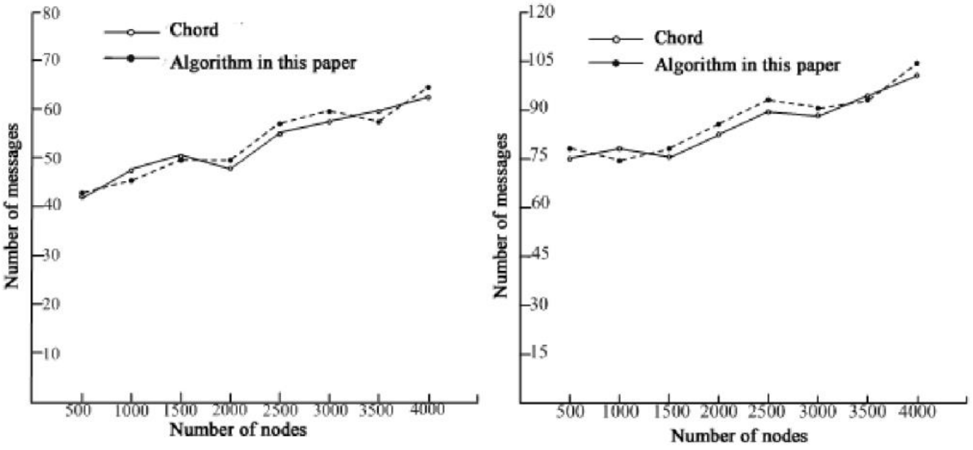


Figure 5. The Number of Sent Messages when Nodes Enter or Exit

The simulation results in Figure 5 respectively record the relationship between the number of messages sent by system and entering or exiting system of the nodes in algorithm provided in this paper and in Chord system model. As can be seen in the figure, with the increase of network node number, the average number of sent messages when nodes enter or exit system in the algorithm provided in this paper and in Chord system model tend to be the same, which indicates that the performances of the two tend to be consistent. This result effectively verify that the increase of routing maintenance overhead based on resource topic designed in this paper is within the acceptable range of the system.

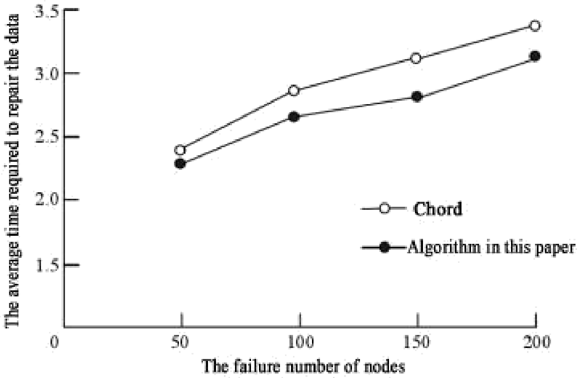


Figure 6. The Average Time Required to Repair the Damaged Data

In this performance simulation test, the size of system node was 1000. Node failure was that we selected at random a fixed number of nodes from the entire system and made it fail, considering 50, 100, 150 and 200 node failure respectively. Then the Chord algorithm and the algorithm of this paper was to investigate the average time required to repair the damaged data. As shown in Figure 6, with the increase of the number of failed nodes, the average time of the algorithm in this section was improved compared with Chord system. When the failure number of nodes was 50, the algorithm proposed in this paper had no difference with the Chord algorithm. This paper analyzed that when the size of system nodes was small and failure nodes are few, chord topology structure of the system could still maintain well, and each backup copy of chord system was less damaged, and Chord algorithm damage repair data was still valid. When the number of nodes was more than 100, the average time of this algorithm to repair the data was improved compared with Chord algorithm. This paper thought that when the number of failure nodes increased, topology structure Chord algorithm was destroyed and then maintenance was more difficult, and data backup also damaged greatly, and the algorithm in this paper designed maintenance mechanism of topological structure more completely than Chord algorithm. In the backup mechanism, the stored data objects mostly backed up in nodes of close distance. The immediate repair strategy is used to the hot data within group and between group backup and hot data. Therefore, when the number of failure nodes was large, the average time of this algorithm to repair the data was improved compared with Chord algorithm. The performance of data persistence storage was better.

6. Conclusion

This paper describes a design idea and system architecture of distributed storage system established on P2P in detail, and proposes an effective mechanism according to resource topic and predicted network distance for node grouping. After grouping, a three-layer overlay network on super node management is established, and a method to maintain this topological structure is proposed. On this basis, a multiplex naming mechanism for node naming and data storage mechanism according to storage resource topic are proposed, and mechanism of elaborate data replication backup and data storage space recovery is designed. By simulation experiment, the effectiveness of this storage system is verified.

Acknowledgments

This work was supported by the Humanities-Society Scientific Research Program of Hubei Province Education Department under grant No.16Q170, and the Foundation of Hubei Province Educational Committee under grant No. 2014B184.T

References

[1] S. Saroiu, K. P. Gummadi and S. D.Gribble, "Measuring and analyzing the characteristics of Napster and Gnutella hosts", Multimedia Systems., vol. 9, no. 2, **(2003)**, pp. 170-184.

[2] M. Ripeanu, "Peer-to-Peer Architecture Case Study: Gnutella Network", International Conference on Peer-to-peer Computing, Linkoping, Sweden, **(2001)**, pp. 99-100.

[3] I. Clarke, O. Sandberg, B. Wiley and T. W. Hong, "Designing Privacy Enhancing Technologies", EditedF.Hannes, Springer Berlin Heidelberg, Berlin, vol. 2009, **(2001)**, pp. 46-66.

[4] I. Stoica, R.Morris, D. Karger, M. F. Kaashoek and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup service for internet applications". SIGCOMM Comput, New York, USA, **(2001)**, pp.149-160.

[5] YANG Lei, "The Research on WAN Distributed Storage Technologies based on P2P Architecture", Hunan University, Changsha, **(2013)**.

[6] S. Ratnasamy, I. Stoica and S. Shenker, "Routing Algorithms for DHTs: Some Open Questions". Peer-to-Peer Systems: First InternationalWorkshop, IPTPS 2002 Cambridge, MA, USA, **(2002)**, pp. 45-52.

[7] Z. P. Li, J. H. Huang and H. Tang, "A P2P computing based self-organizing network routing model", Journal of Software, vol.16, no. 5, **(2005)**, pp. 916−930.

[8] J. Tian and Y. F. Dai, "Study on durable peer-to-peer storage techniques", Journal of Software, vol.18, no. 6, **(2007)**, pp. 1379−1399.

[9] S. Saroiu, P. Gummadi and S. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems", Multimedia Computing and Networking, Philadelphia, USA, **(2002)**, pp. 156-170.