Cycloid: A Constant-Degree and Lookup-Efficient P2P Overlay Network

Haiying Shen and Cheng-Zhong Xu
Department of Electrical & Computer Engg.
Wayne State University, Detroit, MI 48202
{shy,czxu}@ece.eng.wayne.edu

Guihai Chen
State Key Lab of Novel Software Technology
Nanjing University, Nanjing, China
gchen@nju.edu.cn

Abstract

There are many structured P2P systems that use DHT technologies to map data items onto the nodes in various ways for scalable routing and location. Most of the systems require $O(\log n)$ hops per lookup request with $O(\log n)$ neighbors per node, where n is the network size. In this paper, we present a constant-degree P2P architecture, namely Cycloid*, which emulates a Cube-Connected-Cycles (CCC) graph in the routing of lookup requests. It achieves a time complexity of $O(d)$ per lookup request by using $O(1)$ neighbors per node, where $n = d \cdot 2^d$. We compare Cycloid with other two constant-degree systems, Viceroy and Koorde in various architectural aspects via simulation. Simulation results show that Cycloid has more advantages for large scale and dynamic systems that have frequent node arrivals and departures. In particular, Cycloid delivers a higher location efficiency in the average case and exhibits a more balanced distribution of keys and query loads between the nodes.*

1. Introduction

Over the past years, the immerse popularity of P2P resource sharing services has produced a significant stimulus to content-delivery overlay network research. An important class of the overlay networks is distributed hash tables (DHTs) that map keys to the nodes of a network based on a consistent hashing function. Representatives of the DHTs include CAN [13], Chord [15], Pastry [14], Tapestry [16], Kademlia [10]. They organize the nodes in various ways for efficient location of data items. Most of the DHTs require $O(\log n)$ hops per lookup request with $O(\log n)$ neighbors per node, where n is the network size.

The network degree determines the number of neighbors with which a node must maintain continuous contact. In order to reduce the cost for maintenance, in this paper we present a new constant-degree DHT, namely Cycloid. It achieves a lookup path length of $O(d)$ with $O(1)$ neighbors, where d is the network dimension and $n = d \cdot 2^d$. It combines Chord and Pastry and emulates a cube-connected-cycles (CCC) graph in the routing of lookup requests between the nodes.

There exist other two constant-degree DHTs: Viceroy [9] and Koorde [6]. Both of them feature a time complexity of $O(\log n)$ hops per lookup request with $O(1)$ neighbors per node. But they are different in maintenance of the connectivity between a changing set of nodes and in routing for efficient key location. Koorde embeds a de Bruijn graph on the identifier circle for forwarding lookup requests. It bears much resemblance to Chord in routing and connectivity maintenance. Viceroy emulates a butterfly network by assuming a real number id space in $[0, 1)$. It requires to select a butterfly level parameter of each node according to an estimate of the network size. Due to the dynamic nature of peer-to-peer systems, the level of a node may change with time. By contrast, Cycloid specifies each node by a pair of cyclic and cubic indices. It emulates a CCC graph by using a routing algorithm similar to the one in Pastry. Although the lookup complexity of all the three constant-degree DHTs are of the same order $O(\log n)$, our simulation results show that Cycloid has a much shorter path length per lookup request in the average case than Viceroy and Koorde. Cycloid distributes keys and lookup load more evenly between the participating nodes than Viceroy. Also, Cycloid is more robust as it continues to function correctly and efficiently with frequent node joins and leaves.

The rest of this paper is structured as follows. Section 2 presents a concise review of representative DHTs. In particular, Viceroy and Koorde are discussed in detail. Section 3 details the architecture of Cycloid, with an emphasis on its nodal routing table, routing algorithm, and self-organization considerations. Section 4 shows the performance of Cycloid, in comparison with Viceroy and Koorde. Finally, Section 5 concludes this paper with remarks on possible future work.

2. Related Work

There are two classes of peer-to-peer content-delivery overlay networks: unstructured and structured. Unstructured networks such as Gnutella [1] and Freenet [4] do not assign responsibility for data to specific nodes. Nodes join and leave the network according to some loose rules. Currently, the query method is either flooding [1] where the query is propagated to all neighbors, or random-walkers [7] where

Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS'04)
IEEE COMPUTER SOCIETY

Table 1: A Comparison of representative DHTs; d=log n in CAN and n=$d \cdot 2^d$ in Cycloid.

Systems	Base Network	Lookup Cost	Routing Table Size
Chord	cycle	$O(\log n)$	$O(\log n)$
CAN	mesh	$O(dn^{1/d})$	$O(d)$
eCAN	mesh	$O(dn^{1/d})$	$O(d)$
Pastry Tapestry	hypercube	$O(\log n)$	$O(\|L\|)$+ $O(\|M\|)$+ $O(\log n)$
Viceroy	butterfly	$O(\log n)$	7
Koorde	de Bruijn	$O(\log n)$	≥ 2
Cycloid	CCC	$O(d)$	7

the query is forwarded to randomly chosen neighbors until the object is found.

Flooding-based search mechanism brings about heavy traffic in a large-scale system because of exponential increase in messages generated per query. Though random-walkers reduce flooding by some extent, they still create heavy overhead to the network due to the many requesting peers involved. Furthermore, flooding and random walkers cannot guarantee data location. They do not ensure that querying terminates once the data is located, and they cannot prevent one node from receiving the same query multiple times, thus wasting bandwidth.

Structured networks have strictly controlled topologies. The data placement and lookup algorithms are precisely defined based on a distributed hash table (DHT) data structure. The node responsible for a key can always be found even if the system is in a continuous state of change. Because of their potential efficiency, robustness, scalability and deterministic data location, structured networks have been studied intensively in recent years. Representative DHTs include CAN [13], Chord [15], Pastry [14], Tapestry [16], Kademlia [10].

In the following, we review and compare some of the structured DHTs by focusing on their topological aspects. Space limitation prevents us from a detailed discussion of each system. Instead, we give more detailed descriptions of constant-degree Viceroy and Koorde DHTs for comparison. Hypercube-based Pastry is discussed in detail, as well because it serves as a base of our Cycloid DHT. We summarize their architectural characteristics in Table 1. In [3], we presented an abstract and generic topological model that captured the essence of the structural P2P architectures.

Hypercube-Based. Plaxton *et al.* [11] developed perhaps the first routing algorithm that could be scalably used for P2P systems. Tapestry and Pastry use a variant of the algorithm. The approach of routing based on address prefixes, which can be viewed as a generalization of hypercube routing, is common to all theses schemes. The routing algorithm works by correcting a single digit at a time in the left-to-right order: If node number 12345 received a lookup query with key 12456, which matches the first two digits, then the routing algorithm forwards the query to a node which matches the first three digits (e.g., node 12467). To do this, a node needs to have, as neighbors, nodes that match each prefix of its own identifier but differ in the next digit. For each prefix (or dimension), there are many such neighbors (e.g., node 12467 and node 12478 in the above case) since there is no restriction on the suffix, i.e., the rest bits right to the current bit. This is the crucial difference from the traditional hypercube connection pattern and provides the abundance in choosing cubical neighbors and thus a high fault resilience to node absence or node failure. Besides such cubical neighbors spreading out in the key space, each node in Pastry also contains a leaf set L of neighbors which are the set of $|L|$ numerically closest nodes (half smaller, half larger) to the present node ID and a neighborhood set M which are the set of $|M|$ geographically closest nodes to the present node.

Ring-based. Chord uses a one-dimensional circular key space. The node responsible for the key is the node whose identifier most closely follows the key numerically; that node is called the key's successor. Chord maintains two sets of neighbors. Each node has a successor list of k nodes that immediately follow it in the key space and a finger list of $O(\log n)$ nodes spaced exponentially around the key space. The i^{th} entry of the finger list points to the node that is 2^i away from the present node in the key space, or to that node's successor if that node is not alive. So the finger list is always fully maintained without any null pointer. Routing correctness is achieved with such 2 lists. A lookup(key) is, except at the last step, forwarded to the node closest to, but not past, the key. The path length is $O(\log n)$ since every lookup halves the remaining distance to the home.

Constant-degree DHTs. Viceroy [9] maintains a connection graph with a constant-degree logarithmic diameter, approximating a butterfly network. Each Viceroy node in butterfly level l has 7 links to its neighbors, including pointers to its predecessor and successor pointers in a general ring, pointers to the next and previous nodes in the same level ring, and butterfly pointers to its left, right nodes of level $l+1$, and up node of level $l-1$, depending on the node location. In Viceroy, every participating node has two associated values: its identity $\in [0,1)$ and a butterfly level index l. The node id is independently and uniformly generated from a range $[0,1)$ and the level is randomly selected from a range of $[1, \log n_0]$, where n_0 is an estimate of the network size. The node id of a node is fixed, but its level may need to be adjusted during its life time in the system.

Viceroy routing involves three steps: ascending to a level 1 node via up links, descending along the down link until a node is reached with no down links, and traversing to the destination via the level ring or ring pointers. Viceroy takes $O(\log n)$ hops per lookup request.

Koorde [6] combines Chord with de Bruijn graphs. Like Viceroy, it looks up a key by contacting $O(\log n)$ nodes with $O(1)$ neighbors per node. As in Chord, a Koorde node and a key have identifiers that are uniformly distributed in a 2^d identifier space. A key k is stored at its successor, the first node whose id is equal to or follows k in the identifier space.

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

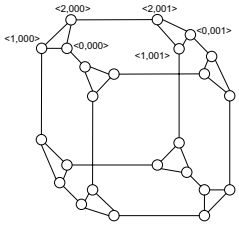


Figure 1: A 3-dimensional Cube-Connected-Cycles.

Node $2^d - 1$ is followed by node 0.

Due to the dynamic nature of the P2P systems, they often contain only a few of the possible 2^d nodes. To embed a de Bruijn graph on a sparsely populated identifier ring, each participating node maintains knowledge about its successor on the ring and its first de Bruijn node. To look up a key k, the Koorde routing algorithm must find the successor of k by walking down the de Bruijn graph. Since the de Bruijn graph is usually incomplete, Koorde simulates the path taken through the complete de Bruijn graph, passing through the immediate real predecessor of each imaginary node on the de Bruijn path.

3. Cycloid: A Constant-Degree DHT

Cycloid combines Pastry with CCC graphs. In a Cycloid system with $n = d \cdot 2^d$ nodes, each lookup takes $O(d)$ hops with $O(1)$ neighbors per node. Like Pastry, it employs consistent hashing to map keys to nodes. A node and a key have identifiers that are uniformly distributed in a $d \cdot 2^d$ identifier space.

3.1. CCC and Key Assignment

A d-dimensional CCC graph is a d-dimensional cube with replacement of each vertex by a cycle of d nodes. It contains $d \cdot 2^d$ nodes of degree 3 each. Each node is represented by a pair of indices $(k, a_{d-1}a_{d-2}\ldots a_0)$, where k is a cyclic index and $a_{d-1}a_{d-2}.....a_0$ is a cubical index. The cyclic index is an integer, ranging from 0 to $d-1$ and the cubical index is a binary number between 0 and 2^d-1. Figure 1 shows the 3-dimensional CCC. A P2P system often contains a changing set of nodes. This dynamic nature poses a challenge for DHTs to manage a balanced distribution of keys among the participating nodes and to connect the nodes in an easy-to-maintain network so that a lookup request can be routed toward its target quickly. In a Cycloid system, each node keeps a routing table and two leaf sets with a total of 7 entries to maintain its connectivity to the rest of the system. Table 2 shows a routing state table for node (4,10111010) in an 8-dimensional Cycloid, where x indicates an arbitrary binary value, inside leaf set maintains the node's predecessor and successor in the local cycle, and outside leaf set maintains the links to the preceding and the succeeding remote cycles. Its corresponding links in both cubical and cyclic aspects are shown Figure 2.

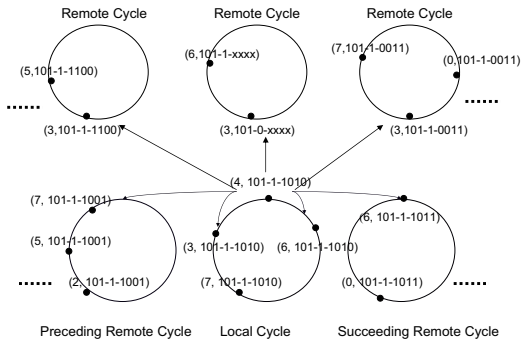


Figure 2: Cycloid node routing links state.

Table 2: Routing table of a Cycloid node (4,101-1-1010).

NodeID(4,101-1-1010)	
Routing table	
cubical neighbor: (3,101-0-xxxx)	
cyclic neighbor: (3,101-1-1100)	
cyclic neighbor: (3,101-1-0011)	
Leaf Sets (half smaller, half larger)	
Inside Leaf Set	
(3,101-1-1010)	(6,101-1-1010)
Outside Leaf Set	
(7,101-1-1001)	(6,101-1-1011)

In general, a node $(k, a_{d-1}a_{d-2}\ldots a_k\ldots a_0)$ $(k \neq 0)$ has one cubical neighbor $(k-1, a_{d-1}a_{d-2}\ldots \overline{a}_k xx...x)$ where x denotes an arbitrary bit value, and two cyclic neighbors $(k-1, b_{d-1}b_{d-2}\ldots b_0)$ and $(k-1, c_{d-1}c_{d-2}\ldots c_0)$. The cyclic neighbors are the first larger and smaller nodes with cyclic index $k-1$ mod d and their most significant different bit with the current node is no larger than $k-1$. That is,

(k-1, $b_{d-1}\ldots b_1b_0$)
= min$\{\forall$(k-1, $y_{d-1}\ldots y_1y_0)|y_{d-1}\ldots y_0 \geq a_{d-1}\ldots a_1a_0\}$
(k-1, $c_{d-1}\ldots c_1c_0$)
= max$\{\forall$(k-1, $y_{d-1}\ldots y_1y_0)|y_{d-1}\ldots y_0 \leq a_{d-1}\ldots a_1a_0\}$

The node with a cyclic index $k = 0$ has no cubical neighbor and cyclic neighbors. The node with cubical index 0 has no small cyclic neighbor, and the node with cubical index $2^d - 1$ has no large cyclic neighbor.

The nodes with the same cubical index are ordered by their cyclic index mod d on a local cycle. The left inside leaf set node points to the node's predecessor and the right inside leaf set node points to the node's successor in the local cycle. The largest cyclic index node in a local cycle is called the primary node of the local cycle. All local cycles are ordered by their cubical index mod 2^d on a large cycle. The left outside leaf set node points to the primary node in the node's preceding remote cycle and the right outside leaf set node points to the primary node in the node's succeeding remote cycle in the large cycle.

The cubical links allow us to change cubical index from left to right, in the same left-to-right order as in Pastry. The cyclic links allow us to change the cyclic index. It is easy to see that the network will be the traditional cube-connected

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

cycles if all nodes are alive. Our connection pattern is resilient in the sense that even if many nodes are absent, the remaining nodes are still capable of being connected. The routing algorithm is heavily assisted by the leaf sets. The leaf sets help improve the routing efficiency, check the termination condition of a lookup, and wrap around the key space to avoid the target overshooting. How the routing table and leaf sets are initialized and maintained is the subject of Section 3.3.

The Cycloid DHT assigns keys onto its id space by the use of a consistent hashing function. The key assignment is similar to Pastry, except that the Cycloid associates a pair of cyclic and cubic indices with each node. For a given key, the cyclic index of its mapped node is set to its hash value modulated by d and the cubical index is set to the hash value divided by d. If the target node of a key's id $(k, a_{d-1} \dots a_1 a_0)$ is not a participant, the key is assigned to the node whose id is first numerically closest to $a_{d-1} a_{d-2} \dots a_0$ and then numerically closest to k.

3.2. Cycloid Routing Algorithm

Cycloid routing algorithm emulates the routing algorithm of CCC [12] from source node $(k, a_{d-1} \dots a_1 a_0)$ to destination $(l, b_{d-1} \dots b_1 b_0)$, incorporating the resilient connection pattern of Cycloid. The routing algorithm involves three phases, assuming MSDB be the most significant different bit of the current node and the destination.

1. *Ascending*: When a node receives a request, if its k<MSDB, it forwards the request to a node in the outside leaf set sequentially until cyclic index k≥MSDB.
2. *Descending*: In the case of k≥MSDB, when k=MSDB, the request is forwarded to the cubical neighbor, otherwise the request is forwarded to the cyclic neighbor or inside leaf set node, whichever is closer to the target, in order to change the cubical index to the target cubical index.
3. *Traverse cycle*: If the target Id is within the leaf sets, the request is forwarded to the closest node in the leaf sets until the closest node is the current node itself.

Figure 3 presents an example of routing a request from node (0,0100) to node (2,1111) in a 4-D Cycloid DHT. The MSDB of node (0,0100) with the destination is 3. As (0,0100) cyclic index k=0 and k<MSDB, it is in the ascending phase. Thus, the node (3,0010) in the outside leaf set is chosen. Node (3,0010)'s cyclic index 3 is equal to its MSDB, then in the descending phase, the request is forwarded to its cubical neighbor (2,1010). After node (2,1010) finds that its cyclic index is equal to its MSDB 2, it forwards the request to its cubical neighbor (1,1110). Because the destination (2,1111) is within its leaf sets, (1,1110) forwards the request to the closest node to the destination (3,1111). Similarly, after (3,1111) finds that the destination is within its leaf sets, it forwards the request to (2,1111) and the destination is reached.

Each of the three phases is bounded by *O(d)* hops, hence the total path length is *O(d)*. The key idea behind this algorithm is to keep the distance decrease repeatedly. The correctness of the routing algorithm can be shown by showing its convergence and reachability. By convergence, we mean that each routing step reduces the distance to the destination. By reachability, we mean that each succeeding node can forward the message to the next node. Because each step sends the lookup request to a node that either shares a longer prefix with the destination than the current node, or shares as long a prefix with, but is numerically closer to the destination than the current node, the routing algorithm is convergent. Also, the routing algorithm can be easily augmented to increase fault tolerance. When the cubical or the cyclic link is empty or faulty, the message can be forwarded to a node in the leaf sets. Our discussion so far is based on a 7-entry Cycloid DHT. It can be extended to include two predecessors and two successors in its inside leaf set and outside leaf set, respectively. We will show via simulations in the next section that the 11-entry Cycloid DHT has better performance.

3.3. Self-Organization

Peer-to-Peer systems are notoriously dynamic in the sense that nodes are frequently joining in and departing from the network. Cycloid deals with node joining and leaving in a distributed manner, without requiring hash information to be propagated through the entire network. This section describes how Cycloid handles node joining and leaving.

3.3.1. Node Join When a new node joins, it needs to initialize its routing table and leaf sets, and inform other related nodes of its presence. Like Chord and Viceroy, Cycloid assumes that any new node initially knows about a live node. Assume the first contact node is $A = (k, a_{d-1} a_{d-2} \dots a_0)$ and the new node is $X = (l, b_{d-1} b_{d-2} \dots b_0)$. According to the routing algorithm in Section 3.2, the node A will route the joining message to the existing node Z whose id is numerically closest to the id of X. Z's Leaf Sets are the basis for X's Leaf Sets. In particular, the following two cases are considered:

1. If X and Z are in the same cycle, Z's outside leaf set becomes the X's outside leaf set. X's inside leaf set is initiated according to Z's inside leaf set. If Z is X's successor, Z's predecessor and Z are the left node and right node in X's inside leaf set respectively. Otherwise, Z and Z's successor are the left node and right node.
2. If X is the only node in its local cycle, then Z is not in the same cycle as X. In this case, two nodes in X's inside leaf set are X itself. X's outside leaf set is initiated according to Z's outside leaf set. If Z's cycle is the succeeding remote cycle of the X, Z's left outside leaf set node and the primary node in Z's cycle are the left node and right node in X's outside leaf set. Otherwise, the primary node in Z's cycle and Z's right out-

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

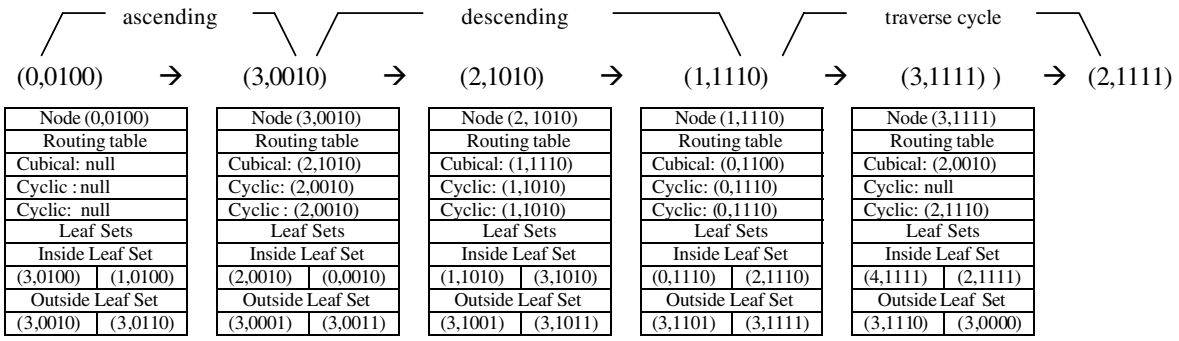


Figure 3: An example of routing phases and routing table states in Cycloid.

side leaf set node are the left node and right node in X's outside leaf set.

We use a local-remote method to initialize the 3 neighbors in the X's routing table. It searches for a neighbor in the local cycle in a decreasing order of the node cyclic index. If the neighbor is not found, then its neighboring remote cycle is searched. The remote cycle search sequence depends on the k^{th} bit in the cubical index. If a_k is 1, the search direction is counter-clockwise, otherwise the direction is clockwise. This is done in order to enhance the possibility and the speed of finding the neighbors.

After a node joins the system, it needs to notify the nodes in its inside leaf set. It also needs to notify the nodes in its outside leaf set if it is the primary node of its local cycle. Once the nodes in the inside leaf set receive the joining message, they will update themselves. When the nodes in the outside leaf set receive the joining message, in addition to update themselves, they need to transfer the message to the nodes in their inside leaf set. Thus, the message is passed along in the joining node's neighboring remote cycle until all the nodes in that cycle finish updating.

3.3.2. Node Departure Before a node leaves, it needs to notify its inside leaf set nodes. In Cycloid, a node only has outgoing connections and has no incoming connections. Therefore, a leaving node cannot notify those who take it as their cubical neighbor or cyclic neighbor. The need to notify the nodes in its outside leaf set depends on whether the leaving node is a primary node. Upon receiving a leaving notification, the nodes in the inside and outside leaf sets update themselves. In addition, the nodes in the outside leaf set need to notify other nodes in their local cycle one by one, which will take at most d steps. As a result, only those who take the leaving node as their inside leaf set or outside leaf set are updated. Those nodes who take the leaving node as their cubical neighbor or cyclic neighbor cannot be updated. Updating cubical and cyclic neighbors are the responsibility of system stabilization, as in Chord.

3.3.3. Fault Tolerance Undoubtedly, low degree P2P networks perform poorly in failure-prone environments, where nodes fail or depart without warning. Usually, the system maintains another list of nodes to handle such problems, such as the successor list in Chord [15] and the bucket in Viceroy [9]. In this paper, we assume that nodes must notify others before leaving, as the authors of Koorde argued that the fault tolerance issue should be handled separately from routing design.

4. Cycloid Performance Evaluation

In [6], Kaashoek and Karger listed five primary performance measures of DHTs: degree in terms of the number of neighbors to be connected, hop count per lookup request, degree of load balance, degree of fault tolerance, and maintenance overhead. In this section, we evaluate Cycloid in terms of these performance measures and compare it with other two constant-degree DHTs: Viceroy and Koorde. Recall that each Cycloid node maintains connectivity to 7 neighbors in its routing table. Cycloid can be extended to include more predecessors and successors in its inside and outside leaf sets for a trade-off for lookup hop count. The results due to 11-neighbor Cycloid are included for a demonstration of the trade-off. Similarly, Koorde DHT provides a flexibility to making a trade-off between routing table size and routing hop count. For a fair comparison, in our simulations, we assumed the Koorde DHT maintained connectivity to 7 neighbors, including 1 de Bruijn node, 3 successors and 3 immediate predecessors of the de Bruijn node. Since all of the constant-degree DHTs borrowed ideas from Chord and other DHTs with $O(\log n)$ neighbors, we also include the results of Chord as references. The actual number of participants varied in different experiments.

4.1. Key location efficiency

It is known that all of the constant-degree DHTs have a complexity of $O(\log n)$ or $O(d)$ hops per lookup request with $O(1)$ neighbors. Although Cycloid contains more nodes than the others for the same network dimension, its average routing performance relative to Viceroy and Koorde is unknown. In this experiment, we simulated networks with $n = d \cdot 2^d$ nodes and varied the dimension d from 3 to 8. Each node made a total of $n/4$ lookup requests to random destinations. Figure 4 plots the mean of the measured path lengths of the lookup requests due to various DHT routing algorithms. The path length of each request is measured

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

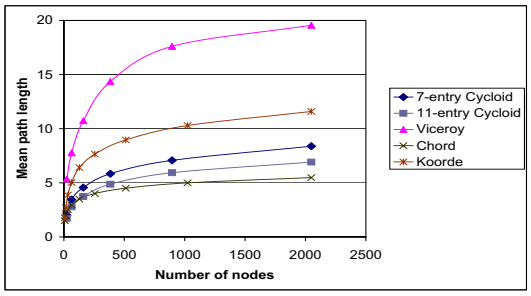


Figure 4: Path lengths of lookup requests in various DHTs of different network sizes.

by the number of hops traversed during its search. From the figure, we can see that the path lengths of Viceroy are more than 2 times than those of Cycloid, although key locations in both Cycloid and Viceroy involve the same ascending, descending and traverse ring/cycle phases. There are two reasons. First, the ascending phase in Cycloid usually takes only one step because the outside leaf set entry node is the primary node in its cycle. But the ascending phase in Viceroy takes $(\log n)/2$ steps on average because each step decreases the level one at a time. Figures 5(a) and (b) present breakdowns of the lookup cost in different phases in Cycloid and Viceroy, respectively. From the figures, we can see that the ascending phase in Viceroy constitutes about 30% of the total path length, but only up to 15% in Cycloid. Second, the descending phase in Cycloid takes d steps because each step redirects the request to a node with longer prefix or is numerically closer to the target. It is followed by another d hops of search in local cycles or cubic neighbor cycles. In Viceroy, the distance to the target can be halved each step in the second descending phase. But Figure 5(b) shows that the descending phases constitutes around only 20% of the total searching path. More than half of the cost is spend in the third traverse ring phase. In the traverse ring phase, the lookup request approaches the destination step by step along ring links or level ring links and needs another $(\log n)/2$ steps on average.

In Koorde, each node redirects an incoming lookup request to its first de Bruijn node or a successor. Each selection of a first de Bruijn node would reduce the distance by half. Since the first de Bruijn node may not be the immediate predecessor of the imaginary node of the destination, selection of a successor is to find the immediate predecessor. Figure 5(c) shows a breakdown of the cost between the two selections. The selection of successors constitutes about 30% of the total path length, which implies some nodes might interpose land in between the current node's first de Bruijn node and the imaginary node. In this case, the current node's successors have to be passed in order to reach the immediate predecessor of the imaginary node. Because of the dense network in which every node is alive, there are only a few nodes at interpose between de Bruijn node and the imaginary node, consequently, the path length of taking successors takes a reasonable percentage of the whole path length. However, Koorde's lookup efficiency is reduced in sparse network. We will discuss this in Section 4.5.

The principle of Cycloid routing algorithm is almost the same as that of Koorde. In both algorithms, starting from a specific chosen node, the node id bits are changed one by one until the target node id is reached. Both of their path lengths are close to d, the dimension of the network in simulation. Since a d-dimensional Cycloid contains more $(d\text{-}1)\cdot 2^d$ nodes than Koorde of the same dimension, Cycloid leads to shorter lookup path length than Koorde in networks of the same size, as shown in Figure 4.

From Figure 4, we can also see that the path length of Viceroy increases faster than the dimension $\log n$. Its path length increases from 4 in a 4-dimensional network to 12.91 in an 8-dimensional network. This means the more nodes a Viceroy network has, the less the key location efficiency.

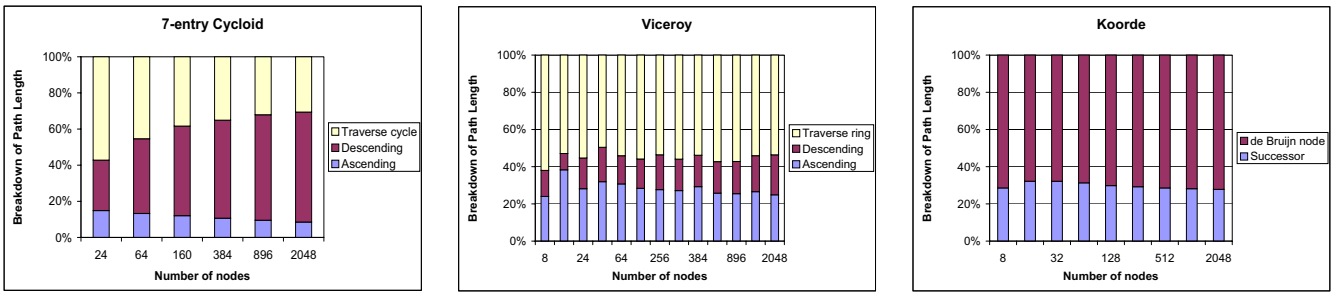
4.2. Load balance

A challenge in the design of balanced DHTs is to distribute keys evenly between a changing set of nodes and to ensure each node experiences even load as an intermediate for lookup requests from other nodes. Cycloid deals with the key distribution problem in a similar way to Koorde, except that Cycloid uses a pair of cyclic and cubical indices to represent a node. Viceroy maintains a one-dimensional id space. Although both Cycloid and Viceroy nodes have two indices to represent their place in the overlay network, the cyclic index is part of the Cycloid node id but the level is not part of the Viceroy node id. Also, Viceroy stores keys in the keys' successors.

In this experiment, we simulated different DHT networks of 2000 nodes each. We varied the total number of keys to be distributed from 10^4 to 10^5 in increments of 10^4. Figure 6(a) plots the mean, the 1st and 99th percentiles of the number of assigned keys per node when the network id space is of 2048 nodes. The number of keys per node exhibits variations that increase linearly with the number of keys in all DHTs. The key distribution in Cycloid has almost the same degree of load balance as in Koorde and Chord because Cycloid's two-dimensional id space is reduced to one-dimension by the use of a pair of modula and divide operations. By comparison, the number of keys per node in Viceroy has much larger variations. Its poor balanced distribution is mainly due to the large span of real number id space in $[0, 1)$. In Viceroy, the key is stored in its successor; that is, a node manages all key-value between its counter-clockwise neighbor and itself. Because of Viceroy's large id span, its node identifiers may not uniformly cover the entire space, some nodes may manage much more keys than the others.

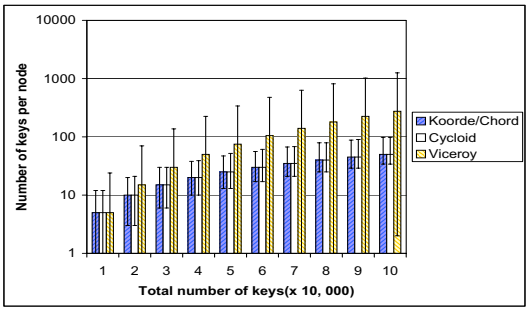
Figure 6(b) plots the mean, the 1st and 99th percentiles of the number of keys per node in Cycloid and Koorde DHTs when there are only 1000 participants in the network. From the figure, it can be seen that Cycloid leads to a more balanced key distribution than Koorde for a sparse network. In Koorde, the node identifiers do not uniformly cover the en-

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

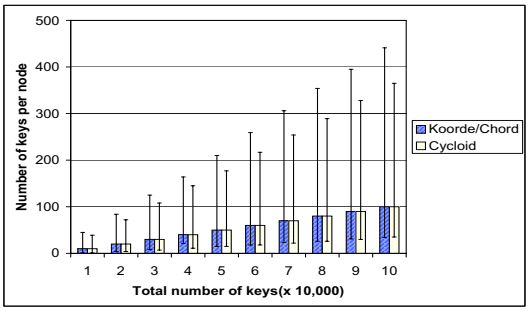


(a) Cycloid (b) Viceroy (c) Koorde

Figure 5: Path length breakdown in various DHTs of different sizes.



(a) Network of 2000 nodes



(b) Network of 1000 nodes

Figure 6: Key distribution in networks of various sizes

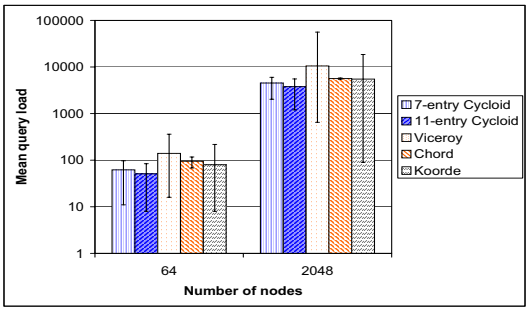


Figure 7: Query load variances in DHTs of different sizes.

tire identifier space, leading to unbalanced key allocation as a result of storing the key in its successor. By comparison, using 2 dimension key allocation method, Cycloid achieves better load balance by storing the key in its numerically closest node; that is, the keys between a node's counter-clockwise neighbor and itself will be allocated to that neighbor or the node itself rather than to itself totally. Chord solved this problem by replicating each node into $O(\log n)$ "virtual nodes", but such replication would destroy the optimality of constant degree in Koorde. In [6], Kaashoek and Karger put forward a question of finding a system that is both degree optimal and load balanced. Cycloid should be an answer.

In summary, when the entire identifier space is mostly occupied, Cycloid's load balance is as good as Chord. When the actual nodes only occupy small part of the total entire identifier space, Cycloid's load balance is better than Chord.

Key distribution aside, another objective of load balancing is to balance the query load between the participating nodes. The query load is measured as the number of queries received by a node for lookup requests from different nodes.

Figure 7 plots the the mean, the 1st and 99th percentiles of query loads of various DHT networks of 64 nodes and 2048 nodes. The figures shows that Cycloid exhibits the smallest variation of the query load, in comparison with other constant-degree DHTs. This is partly due to the symmetric routing algorithm of Cycloid.

In Viceroy, the ascending phase consists of a climb using up connections until level 1 is reached and the descending phase routes down the levels of tree using the down links until no down links. As a result, the nodes in the higher levels will be the hot spots, on the other hand, the nodes of the lower levels have smaller workload, which leads to the great workload variation, especially in the large-scale network. In Koorde, the first de Bruijn of a node with id m is the node immediately precedes 2m. So, all the first de Bruijn nodes' identifiers are even in a "complete"(dense) network and with high probability the ids are even in an incomplete(sparse) network. Consequently, the nodes with even ids have heavy workload while nodes with odd ids have light workload according to the lookup algorithm of Koorde. In Cycloid, because of the leaf sets, the nodes with small cyclic index, typically 0 or 1, will be light loaded. However, these nodes constitute only small part of the Cycloid network, in comparison with the hot spots in Viceroy and Koorde.

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

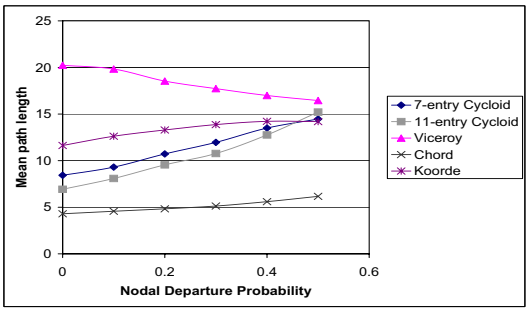


Figure 8: Path lengths of various DHTs with different nodal departure probabilities.

Table 3: Timeout numbers in different DHTs as more nodes depart/fail.

p	7-Cycloid	11-Cycloid	Chord	Koorde
0.1	0.53(0, 4)	0.56(0, 4)	0.62(0, 6)	0.02(0, 1)
0.2	1.24(0, 8)	1.38(0, 8)	1.37(0, 8)	0.04(0, 1)
0.3	2.46(0, 11)	2.14(0, 10)	2.38(0, 11)	0.06(0, 2)
0.4	4.09(0, 17)	3.67(0, 15)	3.91(0, 16)	0.08(0, 3)
0.5	5.88(0, 24)	5.24(0, 25)	6.53(0, 26)	0.09(0, 4)

4.3. Simultaneous node failures/departures

In this section, we evaluate the impact of massive node failures and departures on the performance of various DHTs, and on their capability to performing correct lookups without stabilization. We use the term of departure to refer to both failure and departure. We assume that node departures are graceful—a node informs its relatives before its departure. Ungraceful departure is not discussed in this paper.

In this experiment, we simulated a network of 2048 nodes. Once the network becomes stable, each node is made to fail with probability p ranging from 0.1 to 0.5. After a failure occurs, we performed 10,000 lookups with random sources and destinations. We recorded the number of timeouts occurred in each lookup, the lookup path length, and whether the lookup found the key's correct storing node. A timeout occurs when a node tries to contact a departed node. The number of timeouts experienced by a lookup is equal to the number of departed nodes encountered. Figure 8 shows the mean path length of the lookups with the change of departure probability p in different DHTs. The mean, the 1st and 99th percentiles of the number of timeouts of each DHTs are presented in Table 3.

In Cycloid, the path length increases due to the increasing of the number of timeouts as the p increases. Recall that when a departed node is met, the leaf sets have to be turned to for the next node. Therefore, the path length increases. All lookups were successfully resolved means that the Cycloid is robust and reliable.

We can see from the Figure 8 that unlike Cycloid and Koorde, the lookup path length in Viceroy decreases with the increase of p. In Viceroy, a node has both outgoing and incoming connections. A node notifies its outgoing and incoming connections before its departure. Therefore, all related nodes are updated before the node departs. Based on this characteristic, a massive departures has no adverse effect on Viceroy's ability to perform correct lookups and hence Viceroy has no timeouts. The decrease of the path length is caused by the decrease of the network size. We can see from Figure 8 that when the departure probability is 0.5, the path length is 16.45, which is very close to the average path length (16.92) in a 1024-node "complete" network, as shown in Figure 4.

In order to eliminate the impact of simultaneous node departures in Viceroy, a leaving node would induce $O(\log n)$ hops and require $O(1)$ nodes to change their states. This causes a large amount of overhead. In Cycloid, the path length increased a little with a small fraction of departed nodes. Even though the path length of Cycloid increases slightly, it is still much less than that of Viceroy.

In Figure 8, Koorde's path length increased not so much as in Cycloid when the node departure probability p exceeds 0.3. Unlike Cycloid and Viceroy, Koorde has lookup failures when p becomes larger than 0.3. Our experiment results show that there are 791, 1226, and 4259 lookup failures when p=0.3, 0.4, and 0.5, respectively.

In Koorde, when a node leaves, it notifies its successors and predecessor. Then, its predecessor will points to its successor and its successor will point to its predecessor. By this way, the ring consistency is maintained. The nodes who take the leaving node as their first de Bruijn node or their first de Bruijn node's predecessor will not be notified and their update are the responsibility of stabilization.

Each Koorde node has 3 predecessors of its first de Bruijn node as its backups. When the first de Bruijn node and its backups are all failed, the Koorde node fails to find the next node and the lookup is failed. When the failed node percentage is as low as 0.2, all the queries can be solved successfully at a marginal cost of query length with increase path length as shown in Figure 8. When p exceeds 0.3, with increasing of timeouts as shown in Table 3, the number of failure increases, the path length increases not so much as before because less backups are taken.

From Table 3, we can see that although Koorde has much less timeouts than Cycloid, it still has a large number of failures. In Koorde, the critical node in routing is the de Bruijn node whose backups cannot always be updated. In contrast, Cycloid relies on updated leaf sets of each node for backup. Therefore, Koorde is not as robust as Cycloid in response to massive node failures/departures. The experiment shows that Cycloid is efficient in handling massive node failures/departures without stabilization.

4.4. Lookups during node joining and leaving

In practice, the network needs to deal with nodes joining the system and with nodes that leave voluntarily. In this paper, we assume that multiple join and leave operations do not overlap. We refer the reader to [8] for techniques to

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

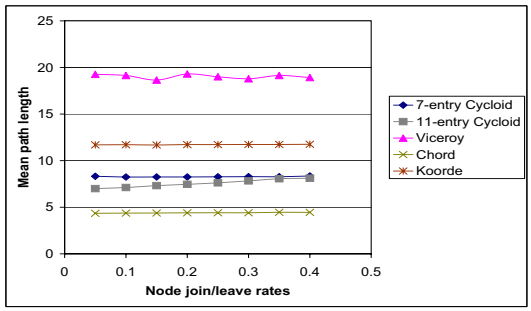


Figure 9: Path length of lookup requests in different DHTs as the node join/leave rates change.

achieve concurrency and to handle failures in the system. In this experiment, we compare Cycloid with Viceroy and Koorde when nodes join and leave continuously.

The setting of this experiment is exactly the same as the one in [15]. Key lookups are generated according to a Poisson process at a rate of one per second. Joins and voluntary leaves are modeled by a Poisson process with a mean rate of R, which ranges from 0.05 to 0.40. A rate of R = 0.05 corresponds to one node joining and leaving every 20 seconds on average. In Cycloid, each node invokes the stabilization protocol once every 30 seconds and each node's stabilization routine is at intervals that are uniformly distributed in the 30 seconds interval. Thus, R ranges from a rate of 1.5 joins and 1.5 leaves per one stabilization periods to a rate of 12 joins and 12 leaves per one stabilization period. The network starts with 2048 nodes.

Figure 9 shows the mean path length of lookup operations in different DHTs as the node join/leave rate R changes. The statistics of the number of timeouts are shown in Table 4. There are no failures in all test cases. From the path length evaluation in Section 4.1, we know that the mean path length of Cycloid in steady states is 8.38. From Figure 9, we can see that the measured path lengths in the presence of node joining and/or leaving are very close to this value and do not change with the rate R. This is because with the help of stabilization, there are less needs for a node to turn to its leaf sets in the case of meeting an absent or departure node. Consequently, a lookup request would experience less timeouts and its path length remains unchanged. Compared with the timeout results in Table 3, we can see that stabilization avoids most of the timeouts.

In Koorde, the path lengthes changed little compared to 11.59 in stable network though the timeouts increases with the rate of node joins and leaves. The failure time is reduced to 0 compared to the large failure time in the Section 4.3. It is because stabilization updates the first de Bruijn node of each node and the de Bruijn node's predecessors in time. When the first de Bruijn node and all of its predecessors are failed, passed lookups would fail with high probability.

The results show that Viceroy's performance is not affected by the node leaving and joining. It is because, before a node leaves and after a node joins, all the related nodes are updated. Although Viceroy has no timeouts, its path length is much longer compared to Cycloid's path.

Table 4: Timeout numbers as the node join/leave rate changes.

R	7-Cycloid	11-Cycloid	Chord	Koorde
0.05	.005(0, 0)	.059(0, 2)	.033(0, 1)	.003(0, 0)
0.10	.009(0, 0)	.103(0, 2)	.078(0, 2)	.013(0, 1)
0.15	.014(0, 1)	.171(0, 2)	.130(0, 2)	.008(0, 0)
0.20	.031(0, 1)	.205(0, 3)	.125(0, 2)	.013(0, 1)
0.25	.047(0, 2)	.246(0, 2)	.151(0, 2)	.016(0, 1)
0.30	.052(0, 2)	.289(0, 3)	.191(0, 3)	.016(0, 1)
0.35	.058(0, 2)	.367(0, 4)	.220(0, 3)	.023(0, 1)
0.40	.070(0, 2)	.374(0, 4)	.233(0, 3)	.023(0, 1)

4.5. Impact of network sparsity in the ID space

Due to the dynamic nature of peer-to-peer systems, a DHT needs to maintain its location efficiency, regardless of the actual number of participants it has. But, in most of the DHTs, some node routing table entries are void when not all nodes are present in the id space. For example, if a local cycle in Cycloid has only one node, then this node has no inside leaf set nodes. It is also possible that a node cannot find a cubical neighbor, or cyclic neighbor. We define the degree of sparsity as the percentage of non-existent nodes relative to the network size. To examine the impact of sparsity on the performance of other systems, we did an experiment to measure the mean search path length and the number of failures when a certain percentage of nodes are not present. We tested a total of 10,000 lookups in different DHT networks with an id space of 2048 nodes. Figure 10 shows the results as the degree of network sparsity changes. There are no lookup failures in each test case. From the figure, we can see that Cycloid keeps its location efficiency and the mean path length decreases slightly with the decrease of network size. In Viceroy, it's impossible for nodes to fully occupy its id space because the node id $\in [0, 1)$. Therefore, it is very likely that the some links of a node are void and hence the sparsity imposes no effect on the location efficiency. In Koorde, the path length increases with the actual number of participants drops. This is because a sparse Koorde DHT exhibits a large span between two neighboring nodes. Since Koorde routes a lookup request through the immediate real predecessor of each imaginary node on the de Bruijn path, the distance between the imagination node and its immediate predecessor in the sparse DHT leads to a longer distance between the predecessor's first de Bruijn node and the imagination node in the next step. Therefore, more successors need to be taken to reach the immediate predecessor of the imagination, thus more path length.

In summary, the sparsity does not have adverse effect on the location efficiency in Cycloid. However, Koorde's performance degrades with the decrease of the number of actual participants.

0-7695-2132-0/04/$17.00 (C) 2004 IEEE

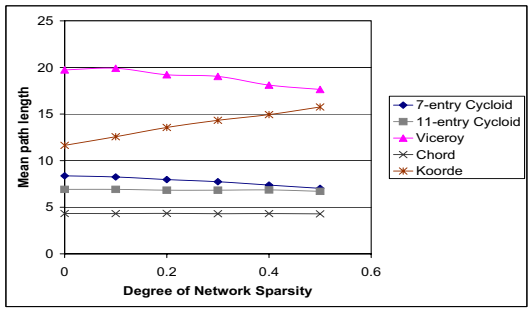


Figure 10: Path length of lookup requests in different DHTs with the change of the degree of network sparsity.

5. Conclusions

In this paper, we have presented a constant-degree DHT, namely Cycloid, and compared it with other two constant-degree DHTs: Viceroy and Koorde. Cycloid is based on Pastry and CCC, while Viceroy and Koorde emulate butterfly and de Bruijn graphs, respectively. Cycloid resembles Viceroy and Koorde in appearance because CCC is a subgraph of butterfly network and de Bruijn is a coset graph of butterfly, as recently proved in graph theories [2, 5]. But they are different in connectivity maintenance of dynamic participants and in routing for key location.

We have evaluated the performance of the DHTs in terms of the lookup hop count, degree of load balance, degree of fault tolerance, cost for maintenance. Experiment results show that (1) Cycloid yields the best average-case location efficiency; (2) Cycloid distributes keys and query load more evenly between the participants than Viceroy. In comparison with Koorde, Cycloid results in higher degrees of load balance for sparse networks and the same degree of load balance for dense networks; (3) Cycloid is more robust because it continues to function correctly when a node's information is only partially correct. By contract, Koorde cannot perform well with partial correct information and incomplete network; (4) Cycloid scales well with the number of nodes, recovers from large numbers of simultaneous node departures and joins, and answers lookups correctly even during recovery. By contrast, Viceroy handles massive node failures/departures at a high cost for connectivity maintenance, especially in the case when a node needs to change its level.

A common problem with constant-degree DHTs is their weakness in handling node leaving without warning in advance. Keeping more information like successor list in Chord and Bucket in Viceroy helps resolve the problem, but destroys the optimality of constant degree. Because of this disadvantage, whenever a node joins or leaves, Cycloid needs to notify its inside leaf set. Especially, if the joining or leaving node is the primary node of a cycle in Cycloid, the updates might produce much more overhead. In addition, the initialization and updates of three neighbors in the routing table may also cause much overhead. These issues need to be further addressed.

Acknowledgments This research was supported in part by U.S. NSF grants ACI-0203592 and CCR-9988266. Guihai Chen's work was partly supported by China NSF Grant (No. 60073029) and China 973 project (No. 2002CB312002).

References

[1] Gnutella. http://www.gnutella.com.

[2] F. Annexstein, M. Baumslag, and A. L. Rosenberg. Group action graphs and parallel architecture. *SIAM J. Computing*, 19:544–569, 1990.

[3] G. Chen, C. Xu, H. Shen, and D. Chen. P2p overlay networks of constant degree. In *Lecture Notes in Computer Science: Proc. of the Int'l Workshop on Grid and Cooperative Computing*, pages 285–192, December 2003.

[4] I. Clarke, O. Sandberg, B. Wiley, and T.W. Hong. Freenet: A distributed anonymous information storage and retrieval system. In *Proc. of the ICSI Workshop on Design Issues in Anonymity and Un-observability*, 2000.

[5] R. Feldmann and W. Unger. The cube-connected cycles network is a subgraph of the bufferfly network. *Parallel Processing Letters*, 2(1):13–19, 1991.

[6] M. F. Kaashoek and R. Karger. Koorde: A simple degree-optimal distributed hash table. In *2nd International workshop on P2P Systems(IPIPS'03)*, 2003.

[7] Q. Lv and S. Shenker. Search and Replication in Unstructured Peer-to-Peer Networks. In *Proceedings of ACM SIGGRAPH', 02*, August 2002.

[8] N. Lynch, D. Malkhi, and D. Ratajczak. Atomic Data Access in Distributed Hash Tables. In *Proc. of the International Peer-to-Peer Symposium*, 2002.

[9] D. Malkhi, M. Naor, and D. Ratajczak. Viceroy: A scalable and dynamic emulation of the butterfly. In *Proc. of Principles of Distributed Computing(PODC 2002)*, 2002.

[10] P. Maymounkov and D. Mazires. Kademlia: A peer-to-peer information system based on the xor metric. *The 1st Interational Workshop on Peer-to-Peer Systems (IPTPS'02)*, 2002.

[11] C. Plaxton, R. Rajaraman, and A. Richa. Accessing nearby copies of replicated objects in a distributed environment. In *Proc. of ACM SPAA*, pages 311–320, 1997.

[12] F. Preparata and J. Vuillemin. The cube-connected cycles: A versatile network for parallel computation. *CACM*, 24(5):300–309, 1981.

[13] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content-addressable network. In *Proceedings of ACM SIGCOMM*, pages 329–350, 2001.

[14] A. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. *18th IFIP/ACM Int'l Conf. on Distributed Systems Platforms(Middleware 2001)*, 2001.

[15] I. Stoica, R. Morris, D. Liben-Nowell, Kaashoek M. F. Karger, D. R. Karger, F. Dabek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup protocol for Internet applications. In *IEEE/ACMTrans. on Networking, 2002. for internet applications. In Proc. SIGCOMM (2001)*, August 2001.

[16] B.Y. Zhao, J. Kubiatowicz, and A.D. Oseph. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. *Technical Report UCB/CSD-01-1141, University of California at Berkeley*, 2001.

0-7695-2132-0/04/$17.00 (C) 2004 IEEE