Parallel Processing Letters

Hierarchical Peer-to-peer Systems

L. Garcés-Erice, E.W. Biersack, K.W. Ross, P.A. Felber and G.Urvoy-Keller
Institut EURECOM
06904 Sophia Antipolis B.P. 193, France

Received (received date)
Revised (revised date)
Communicated by (Name of Editor)

ABSTRACT

Structured peer-to-peer (P2P) lookup services organize peers into a flat overlay network and offer distributed hash table (DHT) functionality. Data is associated with keys and each peer is responsible for a subset of the keys. In hierarchical DHTs, peers are organized into groups, and each group has its autonomous intra-group overlay network and lookup service. Groups are organized in a top-level overlay network. To find a peer that is responsible for a key, the top-level overlay first determines the group responsible for the key; the responsible group then uses its intra-group overlay to determine the specific peer that is responsible for the key. We provide a general framework and a scalable hierarchical overlay management. We study a two-tier hierarchy using Chord for the top level. Our analysis shows that by using the most reliable peers in the top level, the hierarchical design significantly reduces the expected number of hops. We also provide a hierarchical construction of groups, which is built according to the Internet topology, and guarantees very short intra-group communication delay. The results demonstrate the feasibility of locality-based peer groups, which allow P2P systems to take full advantage of the hierarchical design.

Keywords: P2P DHT Hierarchy Topology-aware Superpeers

1. Introduction

Peer-to-peer (P2P) systems are gaining increased popularity, as they make it possible to harness the resources of large populations of networked computers in a cost-effective manner. A central problem of P2P systems is to assign and locate resources among peers. This task is achieved by a P2P *lookup service.*

Several important proposals have been recently put forth for implementing distributed P2P lookup services, including Chord [1], CAN [2], Pastry [3] and Tapestry [4]. In these lookup services, each key for a data item is assigned to the live peer whose node identifier is "closest" to the key (according to some metric). The lookup service determines the peer that is responsible for a given key. The lookup service is implemented by organizing the peers in a structured overlay network, and routing a message through the overlay to the responsible peer. The efficiency of a lookup service is generally measured as a function of the number of peer hops needed to route a message to the responsible peer, as well as the size of the routing table maintained by each peer. For example, Chord requires $O(\log N)$ peer hops and $O(\log N)$ routing table entries when there are N peers in the overlay. Implementations of the distributed lookup service are often referred to as *Distributed Hash Tables (DHTs).*

Chord, CAN, Pastry, and Tapestry are all flat DHT designs without hierarchical routing. Each peer is indistinguishable from another in the sense that all peers use the same rules for determining the routes for lookup messages. This approach is strikingly different from routing in the Internet, which uses hierarchical routing. Hierarchical routing in the Internet offers several benefits over non-hierarchical routing, including scalability and administrative autonomy.

Inspired by hierarchical routing in the Internet, we examine two-tier DHTs in which (i) peers are organized in disjoint groups, and (ii) lookup messages are first routed to the destination group using an inter-group overlay, and then routed to the destination peer using an intra-group overlay.

We present a general framework for hierarchical DHTs. Each group maintains its own overlay network and intra-group lookup service. A top-level overlay is defined among the groups. Within each group, a subset of peers are labeled as "superpeers." Superpeers, which are analogous to gateway routers in hierarchical IP networks, are used by the top-level overlay to route messages among groups. We consider designs for which peers in the same group are locally close, and we provide a solution for organizing such groups in a hierarchy. We describe a cooperative caching scheme that can significantly reduce average data transfer delays. Finally, we also provide a scalable algorithm for assigning peers to groups, identifying superpeers, and maintaining the overlays.

After presenting the general framework, we introduce results from our recent research on TOPLUS [5], an experimental low-stretch DHT that allow for the construction of local hierarchical groups. Then we explore in detail a particular instantiation in which Chord is used for the top-level overlay. Using a novel analytical model, we analyze the expected number of peer hops that are required for a lookup in the hierarchical Chord instantiation. Our model explicitly captures inaccuracies in the routing tables due to peer failures.

The paper is organized as follows: We first discuss related work in Section 2. We then present the general framework for hierarchical DHT's in Section 3 and we illustrate the construction of hierarchical groups with locality properties. Then we discuss the particular case of a two-tier Chord instantiation in Section 4, and we quantify the improvement of lookup latency due to the hierarchical organization of the peers.

2. Related Work

P2P networks can be classified as either unstructured or structured. Chord [1], CAN [2], Pastry [3], Tapestry [4], and P-Grid [6], which use highly structured overlays and use hashing for targeted data placement, are examples of structured P2P networks. These P2P networks are all flat designs (P-Grid uses a virtual distributed search tree only for routing purposes). Gnutella [7] and KaZaA [8], whose overlays grow organically and use random data placement, are examples of unstructured P2P networks.

Ratnasamy et al. [9] explore using landmark nodes to bin peers into groups. The basic idea is for each peer to measure its round-trip time (RTT) to M landmarks, order the resulting RTTs, and then assign itself to one of $M!$ groups. Our hierarchical DHT schemes bear little resemblance to the scheme in [9]. Although in [9] the peers are organized in groups according to locality, the lookup algorithm applies only to CAN, does not use superpeers, and is not a multi-level hierarchical algorithm.

Our approach has been influenced by KaZaA, an enormously successful unstructured P2P file sharing service. KaZaA designates the more available and powerful peers as *supernodes*. In KaZaA, when a new peer wants to join, it bins itself with the existing supernodes, and establishes an overlay connection with the supernode that has the shortest RTT. The supernodes are connected through a top-level overlay network. A similar architecture has been proposed in CAP [10], a two-tier unstructured P2P network. Our design is a blend of the supernode/hierarchy/heterogeniety of KaZaA with the lookup services in the structured DHTs.

Brocade [11] proposes to organize the peers in a two-level overlay. All peers form a *single* overlay O_L. Supernodes are typically well connected and situated near network access points, forming another overlay O_H. Brocade is not truly hierarchical since *all* peers are part of O_L.

Finally, Castro et al. present in [12] a topology-aware version of Pastry [3]. At each hop Pastry presents multiple equivalent choices to route a request. By choosing the closest (smallest network delay) peer at each hop, they try to minimize network delay. However, at each step the possibilities decrease exponentially, so delay is mainly determined by the last hop, usually the longest. We propose large hops to first get to a group, and then shorter local hops inside the group, leading to a more natural caching scheme, as shown later in Section 3.5.

3. Hierarchical Framework

We begin by presenting a general framework for a hierarchical DHT. Although we focus on a two-tier hierarchy, the framework can be extended to a general n-tier hierarchy.

Let P denote the set of peers participating in the system. Each peer has a node identifier (ID). Each peer also has an IP address (dynamic or not). The peers are interconnected through a network of links and switching equipment (routers, bridges, etc.) The peers send lookup query messages to each other using a hierarchical overlay network, as described below.

The peers are organized into groups (see group management in Section 3.3). The groups may or may not be such that the peers in the same group are topologically close to each other, depending on the application needs: for example, the groups could be named according to a common discussion topic in an IRC P2P application, independently of the peers' location. Each group has a unique group id. Let I be the number of groups, G_i the peers in group i, and g_i the ID for group i.

The groups are organized into a *top-level overlay network* defined by a directed graph (X, U), where $X = \{g_1, \ldots, g_I\}$ is the set of all the groups and U is a given set of virtual edges between the nodes (that is, groups) in X. The graph (X, U) is required to be connected, that is, between any two nodes g and g' in X there is a directed path from g to g' that uses the edges in U. It is important to note that this overlay network defines directed edges among groups and not among specific peers in the groups.

Each group is required to have one or more *superpeers*. Let $S_i \subseteq G_i$ be the set of superpeers in group i. Our architecture allows for $S_i = G_i$ for all $i = 1, \ldots, I$, in which case all peers are superpeers. We refer to architectures for which all peers are superpeers as the *symmetric design*. Our architecture also allows $|S_i| = 1$ for all $i = 1, \ldots, I$, in which case each group has exactly one superpeer. Let $R_i = G_i - S_i$ be the set of all "regular peers" in group g_i. For non-symmetric designs ($S_i \neq G_i$), an attempt is made to designate the more powerful peers as superpeers. By "more powerful," we primarily mean the peers that are up and connected the most (and secondarily, those with high CPU power and/or network connection bandwidth). The superpeers are gateways between the groups: they are used for inter-group

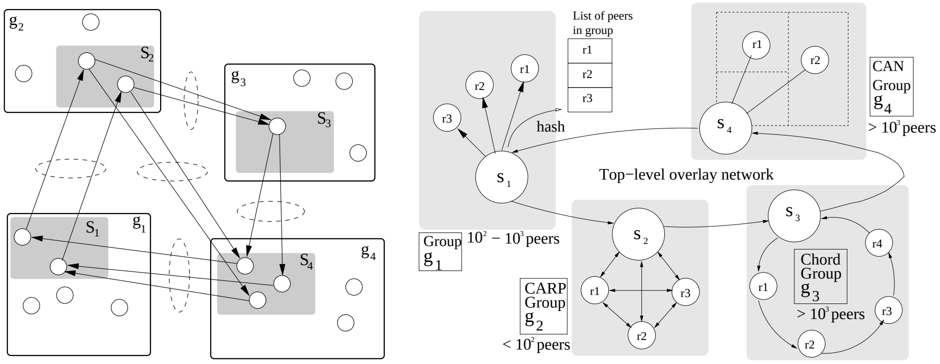

Fig. 1. Left: sample communication relationships between groups in the overlay network and superpeers in neighboring groups. Right: a ring-like overlay network with a single superpeer per group. Intra-group lookup is implemented using different lookup services (CARP, Chord, CAN).

query propagation. To this end, we require that if s_i is a superpeer in G_i, and (g_i, g_j) is an edge in the top-level overlay network (X, U), then s_i knows the name and the current IP address of at least one superpeer $s_j \in S_j$. With this knowledge, s_i can send query messages to s_j. If p is a regular peer, then p can only reach other groups through superpeers. Figure 1 (left) shows a top-level overlay network and possible communication relationships between the corresponding superpeers. Figure 1 (right) shows an example for which there is one superpeer in each group and the top-level overlay network is a ring. Within each group there is also an overlay network among the peers in the group.

3.1. Hierarchical Lookup Service

Consider a two-level lookup service. Given a key k, we say that group g_j is responsible for k if g_j is the "closest" group to k among all the groups. Here "closest" is defined by the specific top-level lookup service (e.g., Chord, CAN, Pastry, or Tapestry).

Our two-tier DHT operates as follows. Suppose a peer $p_i \in G_i$ wants to determine the peer that is responsible for a key k.

(i) Peer p_i sends a query message to one of the superpeers in S_i.
(ii) Once the query reaches a superpeer, the top-level lookup service routes the query through (X, U) to the group G_j that is responsible for the key k. During this phase, the query only passes through superpeers, hopping from one group to the next. Eventually, the query message arrives at some superpeer $s_j \in G_j$.
(iii) Using the overlay network in group j, the superpeer s_j routes the query to the peer $p_j \in G_j$ that is responsible for the key k.

This approach can be generalized to an arbitrary number of levels. A request is first routed through the top-most overlay network to some superpeer at the next level below, which in turn routes the request through its "local" overlay network, and so on until the request finally reaches some peer node at the bottom-most level.

The hierarchical architecture has several important advantages when compared to the flat overlay networks.

- *Exploiting heterogeneous peers:* By designating as superpeers the peers that are "up" the most, the top-level overlay network will be more stable than the corresponding flat overlay network.
- *Transparency:* When a key is moved from one peer to another within a group, the search for the peer holding the key is completely transparent to the top-level algorithm. Similarly, if a group changes its intra-group lookup algorithm, the change is completely transparent to the other groups and to the top-level lookup algorithm. Also, the failure of a regular peer $r_i \in G_i$ (or the appearance of a new peer) will be *local* to G_i; routing tables in peers outside of G_i are not affected.
- *Faster lookup time:* Because the number of groups will be typically orders of magnitude smaller than the total number of peers, queries travel over fewer hops.
- *Less messages in the wide-area:* If the most stable peers form the top-level DHT, most overlay reconstruction messages happen inside groups, which gather peers that are topologically close. Less hops per lookup means also less messages exchanged. Finally, content caching inside groups can further reduce the number of messages that need to get out of the group.

3.2. Intra-Group Lookup

The framework we just described is quite flexible, allowing for different independent intra-group overlays:

If a group has a small number of peers (say, in the tens), each peer could track all the other peers in its group (their IDs and IP addresses); CARP [13] or consistent hashing [14] could be used to assign and locate keys within the group. The number of steps necessary to perform such an intra-group lookup in the destination group is $O(1)$ (g_2 in Figure 1, right).

If the group is a little larger (say, in the hundreds), then the superpeers could track all the peers in the group. In this case, by forwarding a query to a local superpeer, a peer can do a local lookup in $O(1)$ steps (g_1 in Figure 1, right).

Finally, for larger groups, a DHT such as Chord, CAN, Pastry, or Tapestry can be used within the group (g_3 and g_4 in Figure 1, right). A local lookup typically takes $O(\log N)$ hops, for N peers in the group. Even for large groups, with the tools provided later we will see how we can maintain short lookup delays inside groups.

3.3. Hierarchy and Group Management

We now briefly describe the protocols used to manage groups: consider peer p joining the hierarchical DHT. We assume that p is able to get the ID g of the group it belongs to (e.g., g may correspond to the name of p's ISP or university campus). First, p contacts and asks another peer p' already part of the P2P network to look up key g. Following the first step of the hierarchical lookup, p' locates and returns the IP address of the superpeer(s) of the responsible group. If the group ID of the returned superpeer(s) is precisely g, then p joins the group using the regular join mechanisms of the underlying intra-group DHT; additionally, p notifies the superpeer(s) of its CPU and bandwidth resources. If the group ID is not g, then a new group is created with ID g and p as only (super)peer.

In a network with m superpeers per group, the first m peers to join a group g become the superpeers of that group. Because superpeers are expected to be the most stable nodes, we let superpeers monitor the peers that join a group and present "good" characteristics. Superpeers keep an ordered list of the superpeer candidates: the longer a peer remains connected and the higher its resources, the better a superpeer candidate it becomes. This list is sent periodically to the regular peers of the group. When a superpeer fails or disconnects, the first regular peer in the list becomes superpeer and joins the top-level overlay. It informs all peers in its group, as well as the superpeers of the neighboring groups. Although large groups should be avoided, in those cases it may not be cost-effective to broadcast that information to all peers. Instead, peers can learn lazily about changes in the network (like new superpeers), for example by letting superpeers embed membership updates in the lookup responses.

We are thus able to provide stability to the top-level overlay by using multiple superpeers, promoting the most stable peers as superpeers, and rapidly repairing the infrequent failures or departures of the superpeers.

3.4. Locality in a Hierarchical DHT

It is essential for an efficient hierarchical DHT to keep intra-group lookup delays low. While it is easy to provide fast communication between peers within a group confined in a high-speed LAN, some groups may span a full ISP or several ASes with heterogeneous WANs. In those cases, we must try to build the DHT in a such a way that intra-group delay is *as low as possible.* To that end, we must organize the peers in a general n-tier hierarchy of groups, such that each group at each tier gathers peers with good locality properties.

Assume that the IP address of a peer is used to assign that peer to a group, in such a way that all peers that share the same IP prefix are in the same group. The work of Wang [15] shows that the peers sharing the same existing IP network prefix are closer to each other than to other peers. We extend this idea to create a hierarchical DHT called TOPLUS [5], where groups are formed using IP prefixes obtained from BGP tables and other sources. Since the Internet is organized as a hierarchy of networks, using their IP prefixes to group nodes will create a hierarchy of groups. The groups at the bottom of this hierarchy contain nodes, while groups at higher levels contain nested groups. More formally, this structure can be described as follows:

Let I be the set of all 32-bit IP addresses. Let $\mathcal{G}$ be a collection of sets such that $G \subseteq I$ for each $G \in \mathcal{G}$. Thus, each set $G \in \mathcal{G}$ is a set of IP addresses. We refer to each such set G as a *group.* Any group $G \in \mathcal{G}$ that does not contain another group in $\mathcal{G}$ is said to be an *inner group.* We say that the collection $\mathcal{G}$ is a *proper nesting* if it satisfies all the following properties:

(i) $I \in \mathcal{G}$.
(ii) For any pair of groups in $\mathcal{G}$, the two groups are either disjoint, or one group is a proper subset of the other.
(iii) Each $G \in \mathcal{G}$ consists of a set of contiguous IP addresses that can be represented by an IP prefix of the form $w.x.y.z/n$ (for example, 123.13.78.0/23).

In TOPLUS, each node n is contained in a collection of telescoping sets in $\mathcal{G}$; denote these sets by $H_N(n), H_{N-1}(n), \cdots, H_0(n) = I$, where $H_N(n) \subset H_{N-1}(n) \subset \cdots \subset H_0(n)$ and $N \leq L$ is the tier depth of n's inner group. Except for $H_0(n)$, each of these telescoping sets has one or more siblings in the partial-order tree (see Figure 2). Let $\mathcal{S}_i(n)$ be the set of siblings groups of $H_i(n)$ at tier i. n knows a peer in each of these groups.

A lookup for a key k is done forwarding the request to the tier-1 group whose ID (network prefix) is "closest" to the key (using an XOR metric). Then the request is forwarded to the tier-2 group inside the tier-1 group which is closest to the key k, then to the corresponding tier-3 group, and so on. For a system spanning

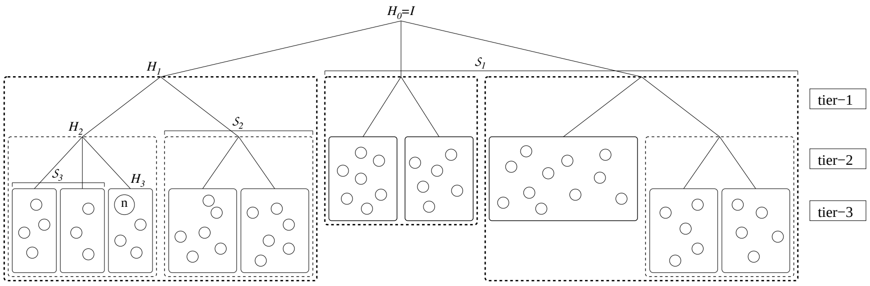

Fig. 2. A sample TOPLUS hierarchy.

the whole Internet, we find that at each step the lookup query gets closer to the destination. This is validated by measuring the stretch, or average ratio of delay between two peers using TOPLUS and using the direct IP path, which is close to 1. In Table 1 we show results for different trees built from IP prefixes. $Tree$ 1 is obtained directly from our IP prefix sources, and the other two trees were obtained by slightly rearranging tier-1 groups to reduce routing table size.

Table 1. Stretch obtained in each tree, depending on the tier of the destination peer.

	stretch TOPLUS vs. IP (± confidence interval)		
Tier	Tree 1	Tree 2	Tree 3
1	1.00 (±0.00)	1.00 (±0.00)	1.00 (±0.00)
2	1.29 (±0.15)	1.32 (±0.14)	1.56 (±0.23)
3	1.31 (±0.16)	1.30 (±0.17)	1.53 (±0.23)
4	1.57 (±0.50)	1.41 (±0.20)	1.56 (±0.50)
Mean	1.17 (±0.06)	1.19 (±0.08)	1.28 (±0.09)

We observe that the deeper the destination is in the hierarchy, the higher the stretch. This is unsurprising, since a lookup through multiple tiers makes additional hops from one group to another, increasing the possible divergence from IP routing paths. However, those hops do not increase too much the stretch. Thus, we conclude that the hops are made, at each tier, among groups that are close to each other.

It is straightforward to build a proper nesting of groups from IP network prefixes inside a top-level group, and TOPLUS has demonstrated that this is a valid method to obtain efficient communication inside groups. Peer groups based on IP prefixes is thus an effective mechanism to construct hierarchical P2P overlays, independently of the lookup algorithm used within groups. Figure 3 shows a mapping of TOPLUS groups to a 3-tier hierarchy of Chord rings. At each level we obtain the best locality achievable inside and among groups. This validates our assumption of a hierarchy of groups of local peers.

3.5. Content Caching

In P2P data sharing applications, once a peer p determines the peer p' that is

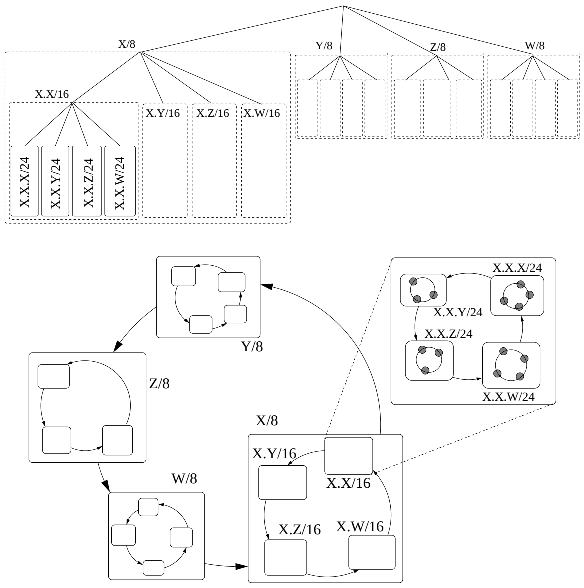

Fig. 3. Top: a sample TOPLUS hierarchy with several groups and their associated IP prefixes. Bottom: the mapping of those groups to a 3-tier hierarchical Chord (group with IP prefix $X.X/16$ detailed).

responsible for a key, p then asks p' for the data associated with the key. If the path from p' to p traverses a congested or low-speed link, the data transfer delay will be long.

In many hierarchical DHT setups, we expect the peers in a same group to be topologically close and to be interconnected by high-speed links (corporate or university campus). By frequently confining data transfers to intra-group transfers, we reduce traffic loads on the access links between the groups and higher-tier ISPs.

Such hierarchical setups can be naturally extended to implement cooperative caching: when a peer $p \in G_i$ wants to obtain the file associated with some key k, it first uses group G_i's intra-lookup algorithm to find the peer $p' \in G_i$ that would be responsible for k if G_i were the entire set of peers. If p' has a local copy of the file associated with k, it returns the file to p; otherwise, p' obtains the file (using the hierarchical DHT), caches a copy, and forwards the file to p. Files are cached in the groups where they have been previously requested. Standard analytical techniques to quantify the reduction in average file transfer time and load on access links can be found in [16].

4. Chord Instantiation

For the remainder of this paper we focus on a specific top-level DHT, namely, Chord. In Chord, each peer and each key has a m-bit ID. IDs are ordered on a circle modulo 2^m (see Figure 4, left). Key k is assigned to the first peer whose identifier is equal to or follows k in the identifier space. This peer is called the successor of key k. Each peer tracks its successor and predecessor peer in the ring. In addition, each peer tracks m other peers, called *fingers*; specifically, a peer with ID p tracks all the successors of the IDs $p + 2^{j-1}$ for each $j = 1, \ldots, m$ (note that p's first finger is in fact its successor). The successor, predecessor, and fingers make up the Chord routing table.

During a lookup, a peer forwards a query to the finger with the largest ID that precedes the key value. The process is repeated from peer to peer until the peer preceding the key is reached, which is the "closest" peer to the key. When there are N peers, the average number of hops needed to reach the destination is $O(\log N)$ [1].

4.1. Inter-Group Chord Ring

In the top-level overlay network, each "node" is actually a group of peers. This implies that the top-level lookup system must manage an overlay of groups, each of which is represented by a set of superpeers. Chord requires some adaptations to manage groups instead of nodes. We will refer to the modified version of Chord as "top-level Chord."

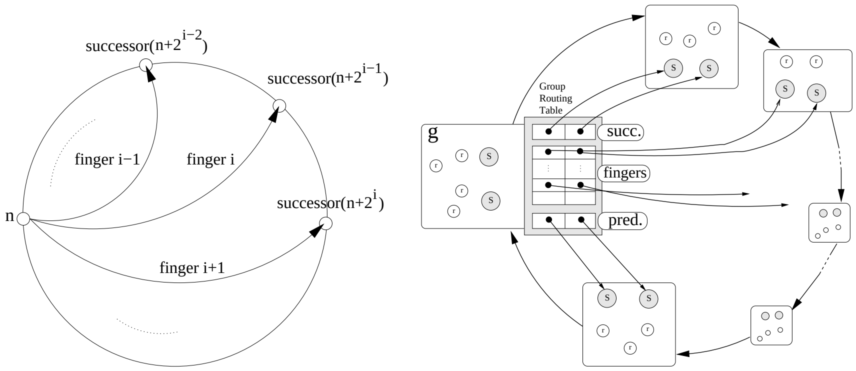

Fig. 4. Normal Chord routing (left) and hierarchical Chord routing (right).

Each node in top-level Chord has a predecessor and successor *vector*, holding the IP addresses of the superpeers of the predecessor and successor group in the ring, respectively. Each finger is also a vector. The routing table of a top-level Chord with two superpeers per group is shown in Figure 4 (right).

The population of groups in the top-level overlay network is expected to be rather stable. However, individual superpeers may fail and disconnect the top-level Chord ring. When the identity of the superpeers S_i of a group g_i changes, the new superpeers eagerly update the vectors of the predecessor and successor groups. This guarantees that each group has an up-to-date view of its neighboring groups and that the ring is never disconnected. Fingers improve the lookup performance, but are not necessary for successfully routing requests. We lazily update the finger tables when we detect that they contain invalid references (similarly to the lazy update of the fingers in regular Chord rings [1]). It is worth noticing that the regular Chord must perform a lookup operation to find a lost finger. Due to the redundancy that our multiple superpeer approach provides, we can choose without delay another superpeer in the finger vector for the same group.

To route a request to a group pointed to by a vector (successor or finger), we choose a random IP address from the vector and forward the request to that superpeer, thus balancing load among superpeers.

4.2. Lookup Latency with Hierarchical Chord

In this section, we quantify the improvement of lookup latency due to the hierarchical organization of the peers. To this end, we compare the lookup performance of the flat Chord and a two-tier hierarchical DHT in which Chord is used for the top level overlay, and arbitrary DHTs are used for the bottom level overlays. For each bottom level group, we only suppose that the peers in the group are topologically close so that intra-group lookup delays are negligible.

In order to make a fair comparison, we suppose that both the flat and hierarchical DHTs have the same number of peers, denoted by P. Let I be the number of groups in the hierarchical design. Because peers are joining and leaving the ring, the finger entries in the peers will not all be accurate. This is more than probable, since fingers are updated lazily. To capture the heterogeneity of the peers, we suppose that there are two categories of peers: *Stable peers*, for which each peer is down with probability p_s. *Unstable peers*, for which each peer is down with probability p_r, with $p_r \gg p_s$. We suppose that the vast majority of the peers are unstable peers. In real P2P networks, like Gnutella, most peers just remain connected the time of getting data from other peers [17]. For the hierarchical organization, we select superpeers from the set of stable peers, and we suppose there is at least one stable peer in each group. Because there are many more unstable peers than stable peers, the probability that a randomly chosen Chord node is down in the flat DHT is approximately p_r. In the hierarchical system, as all the superpeers are stable peers, the probability that a Chord node is down is p_s.

To compare the lookup delay for flat and hierarchical DHTs, we thus only need to consider a Chord ring with N peers, with each peer having the same probability p of being down. The flat DHT corresponds to $(N, p) = (P, p_r)$ and the hierarchical DHT corresponds to $(N, p) = (I, p_s)$. We now proceed to analyze the lookup of the Chord ring (N, p). To simplify the analysis, we assume the N peers are equally spaced on the ring, i.e., the distance between two adjacent peers is $\frac{2^m}{N}$. Our model implies that when a peer attempts to contact another in its finger table, the peer in the finger table will be down with probability p, except if this is the successor peer, for which we suppose that the finger entry is always correct (i.e., the successor is up or the peer is able to find the new successor, which guarantees the correct routing of lookup queries).

Given an initial peer and a randomly generated key, let the random variable H denote the number of Chord hops needed to reach the target peer, that is, to reach the peer responsible for the key. Let T be the random variable that is the clockwise distance in number of peers from the initial peer to the target peer. We want to compute the expectation $E[H]$. Clearly

$$E[H] = \sum_{n=0}^{N-1} P(T=n)E[H|T=n] = \frac{1}{N}\sum_{n=0}^{N-1} E[H|T=n] \tag{1}$$

From (1), it suffices to calculate $E[H|T=n]$ to compute $E[H]$. Let $h(n) = E[H|T=n]$. Note that $h(0)=0$ and $h(1)=1$. Let $j_n = \max\{j : 2^j \leq \frac{2^m n}{N}\}$. The value j_n represents the number of finger entries that precede the target peer, excluding finger 0, the successor. For each of the finger entries, the probability that the corresponding peer is up is p.

Starting at the initial peer, when hopping to the next peer, the query will advance $\lceil \frac{2^{j_n}}{2^m/N} \rceil$ peers if the j_nth finger peer is up; if this peer is down but the $(j_n - 1)$th finger peer is up, the query will advance $\lceil \frac{2^{j_n - 1}}{2^m/N} \rceil$; and so on. Let $q_n(i)$ denote the probability that the ith finger is used. We therefore have

$$h(n) = 1 + \sum_{i=0}^{j_n} q_n(i) h\Big(n - \Big\lceil \frac{2^i}{2^m/N} \Big\rceil\Big) \tag{2}$$

The probability that the ith finger is used is given by $q_n(i) = p^{j_n - i}(1-p)$ for $i = 1, \ldots, j_n$, and by $q_n(0) = p^{j_n}$. Combining (2) with the above expression for $q_n(i)$ we obtain

$$h(n) = 1 + p^{j_n} h(n-1) + (1-p)\sum_{i=1}^{j_n} p^{j_n - i} h\Big(n - \Big\lceil \frac{2^i}{2^m/N} \Big\rceil\Big) \tag{3}$$

Using this recursion, we can calculate all the $h(n)$'s beginning at $h(0)=0$. We then use (1) to obtain the expected number of hops, $E[H]$.

In Figure 5, we plot the expected number of hops in a lookup as a function of the availability of the peers in a Chord system, for different values of N. With smaller values of p, although peers in the ring are not totally reliable, we are still able to advance quite quickly on the ring. Indeed, while the best finger for a target peer is unavailable with probability p, the probability of the second best choice to be also down is p^2, which is far smaller than p.

Despite the good scalability of the Chord lookup algorithm in a flat configuration, the hierarchical architecture can yet significantly decrease the lookup delay. Table 2 gives the expected number of hops for the flat and hierarchical schemes, for different values of P, I, and p_r ($p_s = 0$). We suppose in all cases groups of $\frac{P}{I}$ peers. Since the number of steps is directly related to the lookup delay, we can conclude that the average lookup delay is significantly improved in the hierarchical DHT.

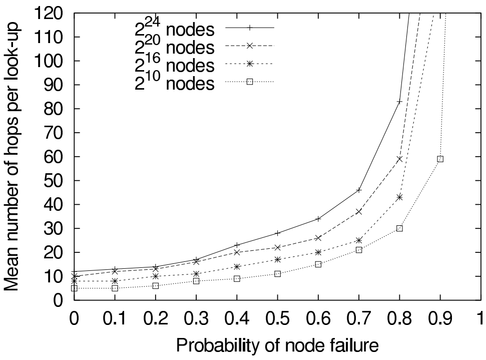

Fig. 5. Number of hops per lookup in Chord.

Table 2. Flat vs. hierarchical networks.

	Flat		*Hierarchical*
	$p_r = 0.5$	$p_r = 0.8$	$p_s = 0$
$P = 2^{16}$, $I = 2^{10}$	17	43	5
$P = 2^{20}$, $I = 2^{16}$	22	59	8
$P = 2^{24}$, $I = 2^{20}$	28	83	10
$P = 2^{24}$, $I = 2^{16}$	28	83	8

5. Conclusion

Hierarchical organizations in general improve overall system scalability. In this paper, we have proposed a generic framework for the hierarchical organization of peer-to-peer overlay networks, and we have demonstrated the various advantages it offers over a flat organization. A hierarchical design offers higher stability by using more "reliable" peers (superpeers) at the top levels. It can use various inter- and intra-group lookup algorithms simultaneously, and treats join/leave events and key migration as local events that affect a single group. By gathering peers into groups based on topological proximity, a hierarchical organization also generates less messages in the wide area and can significantly improve the lookup performance. Finally, our architecture is ideally suited for caching popular content in local groups.

We have presented an instantiation of our hierarchical peer organization using Chord at the top level. The Chord lookup algorithm required only minor adaptations to deal with groups instead of individual peers. When all peers are available, a hierarchical organization reduces the length of the lookup path by a factor of $\frac{\log P}{\log I}$, for I groups and P peers. A hierarchical organization reduces the length of the lookup path dramatically when superpeers are far more stable than regular peers.

Acknowledgments

This research has been partially funded by the VTHD++ (http://www.vthd.org) project from the French Government.

References

[1] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan, Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications, in *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, (San Diego, California, USA), ACM Press, 149–160, Aug. 2001.

[2] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, A Scalable Content-Addressable Network, in *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, (San Diego, California, USA), ACM Press, 161–172, Aug. 2001.

[3] A. Rowstron and P. Druschel, Pastry: Scalable, Distributed Object Location and Routing for Large-Scale Peer-to-Peer Systems, in *Proceedings of IFIP/ACM International Conference on Distributed Systems Platforms (Middleware)*, (Heidelberg, Germany), 329–350, Nov. 2001.

[4] B. Y. Zhao, J. Kubiatowicz, and A. D. Joseph, Tapestry: An Infrastructure for Fault-Tolerant Wide-Area Location and Routing, Tech. Rep. UCB/CSD-01-1141, Computer Science Division, University of California, Berkeley, Apr. 2001.

[5] L. Garces-Erice, K. W. Ross, E. W. Biersack, P. A. Felber, and G. Urvoy-Keller, Topology-Centric Look-Up Service, in *Proceedings of Fifth International Workshop on Networked Group Communications (NGC)*, (Munich, Germany), Springer Verlang, Sep. 2003.

[6] K. Aberer, P-grid: A Self-Organizing Access Structure for P2P Information Systems, in *Proceedings of the Sixth International Conference on Cooperative Information Systems (CoopIS 2001)*, (Trento, Italy), Sep. 2001.

[7] Gnutella. http://gnutella.wego.com

[8] Kazaa. http://www.kazaa.com

[9] S. Shenker, S. Ratnasamy, M. Handley, and R. Karp, Topologically-Aware Overlay Construction and Server Selection, in *Proceedings of IEEE Infocom'02*, (New York City, NY, USA), Jun. 2002.

[10] B. Krishnamurthy, J. Wang, and Y. Xie, Early Measurements of a Cluster-based Architecture for P2P Systems, in *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, (San Francisco, CA, USA), Nov. 2001.

[11] A. D. Joseph, B. Y. Zhao, Y. Duan, L. Huang, and J. D. Kubiatowicz, Brocade: Landmark Routing on Overlay Networks, in *Proceedings of IPTPS'02*, (Cambridge, MA), Mar. 2002.

[12] M. Castro, P. Druschel, Y. C. Hu and A. Rowstron, Topology-Aware Routing in Structure Peer-to-Peer Overlay Network, *International Workshop on Future Directions in Distributed Computing (FuDiCo)*, (Bertinoro, Italy), Jun. 2002.

[13] K. W. Ross, Hash-routing for Collections of Shared Web Caches, *IEEE Network Magazine*, vol. 11 (7) , 37–44, Nov.-Dec. 1997.

[14] D. Karger, A. Sherman, A. Berkhemier, B. Bogstad, R. Dhanidina, K. Iwamoto, B. Kim, L. Matkins, and Y. Yerushalmi, Web Caching with Consistent Hashing, in *Eighth International World Wide Web Conference*, May 1999.

[15] J. Wang, Network Aware Client Clustering and Applications. PhD thesis, Cornell University, May 2001.

[16] J. F. Kurose and K. W. Ross, *Computer Networks: A Top-Down Approach Featuring the Internet*, 2nd edition. Addison Wesley, 2002.

[17] E. Adar and B. A. Huberman, Free Riding on Gnutella, *First Monday*, vol. 5, Oct. 2000.