P2PSIP Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 2, 2010

J. Maenpaa
G. Camarillo
J. Hautakorpi
Ericsson

March 1, 2010

# A Self-tuning Distributed Hash Table (DHT) for REsource LOcation And Discovery (RELOAD) draft-ietf-p2psip-self-tuning-01.txt

## Abstract

REsource LOcation And Discovery (RELOAD) is a peer-to-peer (P2P) signaling protocol that provides an overlay network service. Peers in a RELOAD overlay network collectively run an overlay algorithm to organize the overlay, and to store and retrieve data. This document describes how the default topology plugin of RELOAD can be extended to support self-tuning, that is, to adapt to changing operating conditions such as churn and network size.

## Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on September 2, 2010.

## Copyright Notice

Table of Contents

## 1. Introduction

REsource LOcation And Discovery (RELOAD) [I-D.ietf-p2psip-base] is a peer-to-peer signaling protocol that can be used to maintain an overlay network, and to store data in and retrieve data from the overlay. For interoperability reasons, RELOAD specifies one overlay algorithm, called chord-reload, that is mandatory to implement. This document extends the chord-reload algorithm by introducing self-tuning behavior.

DHT-based overlay networks are self-organizing, scalable and reliable. However, these features come at a cost: peers in the overlay network need to consume network bandwidth to maintain routing state. Most DHTs use a periodic stabilization routine to counter the undesirable effects of churn on routing. To configure the parameters of a DHT, some characteristics such as churn rate and network size need to be known in advance. These characteristics are then used to configure the DHT in a static fashion by using fixed values for parameters such as the size of the successor set, size of the routing table, and rate of maintenance messages. The problem with this approach is that it is not possible to achieve a low failure rate and a low communication overhead by using fixed parameters. Instead, a better approach is to allow the system to take into account the evolution of network conditions and adapt to them. This document extends the mandatory-to-implement chord-reload algorithm by making it self-tuning. Two main advantages of self-tuning are that users no longer need to tune every DHT parameter correctly for a given operating environment and that the system adapts to changing operating conditions.

The remainder of this document is structured as follows: Section 2 provides definitions of terms used in this document. Section 3 discusses alternative approaches to stabilization operations in DHTs, including reactive stabilization, periodic stabilization, and adaptive stabilization. Section 4 gives an introduction to the Chord DHT algorithm. Section 5 describes how this document extends the stabilization routine of the chord-reload algorithm. Section 6 describes how the stabilization rate and routing table size are calculated in an adaptive fashion.

## 2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

This document uses the terminology and definitions from the Concepts

and Terminology for Peer to Peer SIP [I-D.ietf-p2psip-concepts] draft.

Chord Ring: The Chord DHT orders identifiers on an identifier circle of size 2^numBitsInNodeId (numBitsInNodeId is the number of bits in node identifiers). This identifier circle is called the Chord ring.

DHT: Distributed Hash Tables (DHTs) are a class of decentralized distributed systems that provide a lookup service similar to a hash table. Given a key, any participating peer can retrieve the value associated with that key. The responsibility for maintaining the mapping from keys to values is distributed among the peers.

Finger Table: A data structure with up to numBitsInNodeId entries maintained by each peer in a Chord-based overlay. The ith entry in the finger table of peer n contains the identity of the first peer that succeeds n by at least 2^(numBitsInNodeId-i) on the Chord ring. This peer is called the ith finger of peer n. As an example, the first entry in the finger table of peer n contains a peer half-way around the Chord ring from peer n. The purpose of the finger table is to accelerate lookups.

log2(N): Logarithm of N with base 2.

n.id: Peer-ID of peer n.

Neighborhood Set: Consists of successor and predecessor lists.

numBitsInNodeId: Number of bits in a Node-ID.

O(g(n)): Informally, saying that some equation f(n) = O(g(n)) means that f(n) is less than some constant multiple of g(n).

Omega(g(n)): Informally, saying that some equation f(n) = Omega(g(n)) means that f(n) is more than some constant multiple of g(n).

Predecessor List: A data structure containing the predecessors of a peer on the Chord ring.

Routing Table: The set of peers that a node can use to route overlay messages. The routing table consists of the finger table, successor list and predecessor list.

Successor List: A data structure containing the first r successors of a peer on the Chord ring.

## 3. Introduction to Stabilization in DHTs

DHTs use stabilization routines to counter the undesirable effects of churn on routing. The purpose of stabilization is to keep the routing information of each peer in the overlay consistent with the constantly changing overlay topology. There are two alternative approaches to stabilization: periodic and reactive [rhea2004]. Periodic stabilization can either use a fixed stabilization rate or calculate the stabilization rate in an adaptive fashion.

### 3.1. Reactive vs. Periodic Stabilization

In reactive stabilization, a peer reacts to the loss of a peer in its neighborhood set or to the appearance of a new peer that should be added to its neighborhood set by sending a copy of its neighbor table to all peers in the neighborhood set. Periodic recovery, in contrast, takes place independently of changes in the neighborhood set. In periodic recovery, a peer periodically shares its neighborhood set with each or a subset of the members of that set.

The chord-reload algorithm [I-D.ietf-p2psip-base] supports both reactive and periodic stabilization. It has been shown in [rhea2004] that reactive stabilization works well for small neighborhood sets (i.e., small overlays) and moderate churn. However, in large-scale (e.g., 1000 peers or more [rhea2004]) or high-churn overlays, reactive stabilization runs the risk of creating a positive feedback cycle, which can eventually result in congestion collapse. In [rhea2004], it is shown that a 1000-peer overlay under churn uses significantly less bandwidth and has lower latencies when periodic stabilization is used than when reactive stabilization is used. Although in the experiments carried out in [rhea2004], reactive stabilization performed well when there was no churn, its bandwidth use was observed to jump dramatically under churn. At higher churn rates and larger scale overlays, periodic stabilization uses less bandwidth and the resulting lower contention for the network leads to lower latencies. For this reason, most DHTs such as CAN [CAN], Chord [Chord], Pastry [Pastry], Bamboo [rhea2004], etc. use periodic stabilization [ghinita2006]. As an example, the first version of Bamboo used reactive stabilization, which caused Bamboo to suffer from degradation in performance under churn. To fix this problem, Bamboo was modified to use periodic stabilization.

In Chord, periodic stabilization is typically done both for successors and fingers. An alternative strategy is analyzed in [krishnamurthy2008]. In this strategy, called the correction-on-change maintenance strategy, a peer periodically stabilizes its successors but does not do so for its fingers. Instead, finger pointers are stabilized in a reactive fashion. The results obtained in [krishnamurthy2008] imply that although the correction-on-change strategy works well when churn is low, periodic stabilization outperforms the correction-on-change strategy when churn is high.

## 3.2. Configuring Periodic Stabilization

When periodic stabilization is used, one faces the problem of selecting an appropriate execution rate for the stabilization procedure. If the execution rate of periodic stabilization is high, changes in the system can be quickly detected, but at the disadvantage of increased communication overhead. Alternatively, if the stabilization rate is low and the churn rate is high, routing tables become inaccurate and DHT performance deteriorates. Thus, the problem is setting the parameters so that the overlay achieves the desired reliability and performance even in challenging conditions, such as under heavy churn. This naturally results in high cost during periods when the churn level is lower than expected, or alternatively, poor performance or even network partitioning in worse than expected conditions.

In addition to selecting an appropriate stabilization interval, regardless of whether periodic stabilization is used or not, an appropriate size needs to be selected for the neighborhood set and for the finger table.

The current approach is to configure overlays statically. This works in situations where perfect information about the future is available. In situations where the operating conditions of the network are known in advance and remain static throughout the lifetime of the system, it is possible to choose fixed optimal values for parameters such as stabilization rate, neighborhood set size and routing table size. However, if the operating conditions (e.g., the size of the overlay and its churn rate) do not remain static but evolve with time, it is not possible to achieve both a low lookup failure rate and a low communication overhead by using fixed parameters [ghinita2006].

As an example, to configure the Chord DHT algorithm, one needs to select values for the following parameters: size of successor list, stabilization interval, and size of the finger table. To select an appropriate value for the stabilization interval, one needs to know the expected churn rate and overlay size. According to

[liben-nowell2002], a Chord network in a ring-like state remains in a ring-like state as long as peers send Omega(log2^2(N)) messages before N new peers join or N/2 peers fail. Thus, in a 500-peer overlay churning at a rate such that one peer joins and one peer leaves the network every 30 seconds, an appropriate stabilization interval would be on the order of 93s. According to [Chord], the size of the successor list and finger table should be on the order of log2(N). Having a successor list of size O(log2(N)) makes it unlikely that a peer will lose all of its successors, which would cause the Chord ring to become disconnected. Thus, in a 500-peer network each peer should maintain on the order of nine successors and fingers. However, if the churn rate doubles and the network size remains unchanged, the stabilization rate should double as well. That is, the appropriate maintenance interval would now be on the order of 46s. On the other hand, if the churn rate becomes e.g. six-fold and the size of the network grows to 2000 peers, on the order of eleven fingers and successors should be maintained and the stabilization interval should be on the order of 42s. If one continued using the old values, this could result in inaccurate routing tables, network partitioning, and deteriorating performance.

### 3.3. Adaptive Stabilization

A self-tuning DHT takes into consideration the continuous evolution of network conditions and adapts to them. In a self-tuning DHT, each peer collects statistical data about the network and dynamically adjusts its stabilization rate, neighborhood set size, and finger table size based on the analysis of the data [ghinita2006]. Reference [mahajan2003] shows that by using self-tuning, it is possible to achieve high reliability and performance even in adverse conditions with low maintenance cost. Adaptive stabilization has been shown to outperform periodic stabilization in terms of both lookup failures and communication overhead [ghinita2006].

## 4. Introduction to Chord

Chord [Chord] is a structured P2P algorithm that uses consistent hashing to build a DHT out of several independent peers. Consistent hashing assigns each peer and resource a fixed-length identifier. Peers use SHA-1 as the base hash fuction to generate the identifiers. As specified in RELOAD base, the length of the identifiers is numBitsInNodeId=128 bits. The identifiers are ordered on an identifier circle of size 2^numBitsInNodeId. On the identifier circle, key k is assigned to the first peer whose identifier equals or follows the identifier of k in the identifier space. The identifier circle is called the Chord ring.

Different DHTs differ significantly in performance when bandwidth is limited. It has been shown that when compared to other DHTs, the advantages of Chord include that it uses bandwidth efficiently and can achieve low lookup latencies at little cost [li2004].

A simple lookup mechanism could be implemented on a Chord ring by requiring each peer to only know how to contact its current successor on the identifier circle. Queries for a given identifier could then be passed around the circle via the successor pointers until they encounter the first peer whose identifier is equal to or larger than the desired identifier. Such a lookup scheme uses a number of messages that grows linearly with the number of peers. To reduce the cost of lookups, Chord maintains also additional routing information; each peer n maintains a data structure with up to numBitsInNodeId entries, called the finger table. The first entry in the finger table of peer n contains the peer half-way around the ring from peer n. The second entry contains the peer that is 1/4th of the way around, the third entry the peer that is 1/8th of the way around, etc. In other words, the ith entry in the finger table at peer n contains the identity of the first peer s that succeeds n by at least 2^(numBitsInNodeId-i) on the Chord ring. This peer is called the ith finger of peer n. The interval between two consecutive fingers is called a finger interval. The ith finger interval of peer n covers the range [n.id + 2^(numBitsInNodeId-i), n.id + 2^(numBitsInNodeId-i+1)) on the Chord ring. In an N-peer network, each peer maintains information about O(log2(N)) other peers in its finger table. As an example, if N=100000, it is sufficient to maintain 17 fingers.

Chord needs all peers' successor pointers to be up to date in order to ensure that lookups produce correct results as the set of participating peers changes. To achieve this, peers run a stabilization protocol periodically in the background. The stabilization protocol of the original Chord algorithm uses two operations: successor stabilization and finger stabilization. However, the Chord algorithm of RELOAD base defines two additional stabilization components, as will be discussed below.

To increase robustness in the event of peer failures, each Chord peer maintains a successor list of size r, containing the peer's first r successors. The benefit of successor lists is that if each peer fails independently with probability p, the probability that all r successors fail simultaneously is only p^r.

The original Chord algorithm maintains only a single predecessor pointer. However, multiple predecessor pointers (i.e., a predecessor list) can be maintained to speed up recovery from predecessor failures. The routing table of a peer consists of the successor

list, finger table, and predecessor list.

## 5. Extending Chord-reload to Support Self-tuning

This section describes how the mandatory-to-implement chord-reload algorithm defined in RELOAD base [I-D.ietf-p2psip-base] can be extended to support self-tuning.

The chord-reload algorithm supports both reactive and periodic recovery strategies. When the self-tuning mechanisms defined in this document are used, the periodic recovery strategy MUST be used. Further, chord-reload specifies that at least three predecessors and three successors need to be maintained. When the self-tuning mechanisms are used, the appropriate sizes of the successor list and predecessor list are determined in an adaptive fashion based on the estimated network size, as will be described in Section 6.

As specified in RELOAD base, each peer MUST maintain a stabilization timer. When the stabilization timer fires, the peer MUST restart the timer and carry out the overlay stabilization routine. Overlay stabilization has four components in chord-reload:

1. Update the neighbor table. We refer to this as neighbor stabilization.
2. Refreshing the finger table. We refer to this as finger stabilization.
3. Adjusting finger table size.
4. Detecting partitioning. We refer to this as strong stabilization.

As specified in RELOAD base [I-D.ietf-p2psip-base], a peer sends periodic messages as part of the neighbor stabilization, finger stabilization, and strong stabilization routines. In neighbor stabilization, a peer periodically sends an Update request to every peer in its Connection Table. The default time is every ten minutes. In finger stabilization, a peer periodically searches for new peers to include in its finger table. This time defaults to one hour. This document specifies how the neighbor stabilization and finger stabilization intervals can be determined in an adaptive fashion based on the operating conditions of the overlay. The subsections below describe how this document extends the four components of stabilization.

### 5.1. Update Requests

As described in RELOAD base [I-D.ietf-p2psip-base], the neighbor and finger stabilization procedures are implemented using Update

requests. RELOAD base defines three types of Update requests: 'peer_ready', 'neighbors', and 'full'. Regardless of the type, all Update requests include an 'uptime' field. Since the self-tuning extensions require information on the uptimes of peers in the routing table, the sender of an Update request MUST include its current uptime in seconds in the 'uptime' field.

When self-tuning is used, each peer decides independently the appropriate size for the successor list, predecessor list and finger table. Thus, the 'predecessors', 'successors', and 'fingers' fields included in RELOAD Update requests are of variable length. As specified in RELOAD [I-D.ietf-p2psip-base], variable length fields are on the wire preceded by length bytes. In the case of the successor list, predecessor list, and finger table, there are two length bytes (allowing lengths up to 2^16-1). The number of NodeId structures included in each field can be calculated based on the length bytes since the size of a single NodeId structure is 16 bytes. If a peer receives more entries than fit into its successor list, predecessor list or finger table, the peer MUST ignore the extra entries. If a peer receives less entries than it currently has in its own data structure, the peer MUST NOT drop the extra entries from its data structure.

## 5.2. Neighbor Stabilization

In the neighbor stabilization operation of chord-reload, a peer periodically sends an Update request to every peer in its Connection Table. In a small, low-churn overlay, the amount of traffic this process generates is typically acceptable. However, in a large-scale overlay churning at a moderate or high churn rate, the traffic load may no longer be acceptable since the size of the connection table is large and the stabilization interval relatively short. The self-tuning mechanisms described in this document are especially designed for overlays of the latter type. Therefore, when the self-tuning mechanisms are used, each peer MUST send a periodic Update request only to its first predecessor and first successor on the Chord ring.

The neighbor stabilization routine MUST be executed when the stabilization timer fires. To begin the neighbor stabilization routine, a peer MUST send an Update request to its first successor and its first predecessor. The type of the Update request MUST be 'neighbors'. The Update request MUST include the successor and predecessor lists of the sender. If a peer receiving such an Update request learns from the predecessor and successor lists included in the request that new peers can be included in its neighborhood set, it MUST send Attach requests to the new peers.

After a new peer has been added to the predecessor or successor list,

an Update request of type 'peer_ready' MUST be sent to the new peer. This allows the new peer to insert the sender into its neighborhood set.

### 5.3. Finger Stabilization

In the finger stabilization routine of chord-reload, a peer periodically searches for new peers to replace invalid (that is, repeated or failed) entries in the finger table. Chord-reload provides two possible methods for searching for new peers to include in the finger table. In alternative 1, a peer selects one entry from among the invalid entries in its finger table each time the stabilization timer fires and sends a Ping request to the selected entry. In alternative 2, a peer sends a RouteQuery request to all invalid entries in the finger table. After having received RouteQuery responses, the peer sends a Ping to those entries for which the RouteQuery response came from a peer not already present in the routing table. When the self-tuning mechanisms defined in this draft are being used in the overlay, alternative 1 MUST be used since the traffic load it generates is lower and thus more appropriate for large-scale overlays experiencing churn.

Immediately after a new peer has been added to the finger table, a Probe request MUST be sent to the new peer to fetch its uptime. The requested_info field of the Probe request MUST be set to contain the ProbeInformationType 'uptime' defined in RELOAD base [I-D.ietf-p2psip-base].

### 5.4. Adjusting Finger Table Size

The chord-reload algorithm defines how a peer can make sure that the finger table is appropriately sized to allow for efficient routing. Since the self-tuning mechanisms specified in this document produce a network size estimate, this estimate can be directly used to calculate the optimal size for the finger table. This mechanism MUST be used instead of the one specified by chord-reload. A peer MUST use the network size estimate to determine whether it needs to adjust the size of its finger table each time when the stabilization timer fires. The way this is done is explained in Section 6.2.

### 5.5. Detecting Partitioning

This document does not require any changes to the mechanism chord-reload uses to detect network partitioning.

### 5.6. Leaving the Overlay

As specified in RELOAD base [I-D.ietf-p2psip-base], a leaving peer SHOULD send a Leave request to all members of its neighbor table prior to leaving the overlay. The overlay_specific_data field MUST contain the ChordLeaveData structure. The Leave requests that are sent to successors MUST contain the predecessor list of the leaving peer. The Leave requests that are sent to the predecessors MUST contain the successor list of the leaving peer. If a given successor can identify better predecessors than are already included in its predecessor lists by investigating the predecessor list it receives from the leaving peer, it MUST send Attach requests to them. Similarly, if a given predecessor identifies better successors by investigating the successor list it receives from the leaving peer, it MUST send Attach requests to them.

## 6. Self-tuning Chord Parameters

This section specifies how to determine an appropriate stabilization rate and routing table size in an adaptive fashion. The proposed mechanism is based on [mahajan2003], [liben-nowell2002], and [ghinita2006]. To calculate an appropriate stabilization rate, the values of three parameters MUST be estimated: overlay size N, failure rate U, and join rate L. To calculate an appropriate routing table size, the estimated network size N can be used. Peers in the overlay MUST re-calculate the values of the parameters to self-tune the chord-reload algorithm at the end of each stabilization period before re-starting the stabilization timer.

### 6.1. Estimating Overlay Size

Techniques for estimating the size of an overlay network have been proposed for instance in [mahajan2003], [horowitz2003], [kostoulas2005], [binzenhofer2006], and [ghinita2006]. In Chord, the density of peer identifiers in the neighborhood set can be used to produce an estimate of the size of the overlay, N [mahajan2003]. Since peer identifiers are picked randomly with uniform probability from the numBitsInNodeId-bit identifier space, the average distance between peer identifiers in the successor set is (2^numBitsInNodeId)/N.

To estimate the overlay network size, a peer MUST compute the average inter-peer distance d between the successive peers starting from the most distant predecessor and ending to the most distant successor in the successor list. The estimated network size MUST be calculated as:

$$N = \frac{2^{numBitsInNodeId}}{d}$$

This estimate has been found to be accurate within 15% of the real network size [ghinita2006]. Of course, the size of the neighborhood set affects the accuracy of the estimate.

During the join process, a joining peer fills its routing table by sending a series of Ping and Attach requests, as specified in RELOAD base [I-D.ietf-p2psip-base]. Thus, a joining peer immediately has enough information at its disposal to calculate an estimate of the network size.

## 6.2. Determining Routing Table Size

As specified in RELOAD base, the finger table must contain at least 16 entries. When the self-tuning mechanisms are used, the size of the finger table MUST be set to max(log2(N), 16) using the estimated network size N.

The size of the successor list MUST be set to log2(N). An implementation MAY place a lower limit on the size of the successor list. As an example, the implementation might require the size of the successor list to be always at least three.

A peer MAY choose to maintain a fixed-size predecessor list with only three entries as specified in RELOAD base. However, it is RECOMMENDED that a peer maintains log2(N) predecessors.

## 6.3. Estimating Failure Rate

A typical approach is to assume that peers join the overlay according to a Poisson process with rate L and leave according to a Poisson process with rate parameter U [mahajan2003]. The value of U can be estimated using peer failures in the finger table and neighborhood set [mahajan2003]. If peers fail with rate U, a peer with M unique peer identifiers in its routing table should observe K failures in time K/(M*U). Every peer in the overlay MUST maintain a history of the last K failures. The current time MUST be inserted into the history when the peer joins the overlay. The estimate of U MUST be calculated as:

$$U = \frac{k}{M * Tk},$$

where M is the number of unique peer identifiers in the routing

table, Tk is the time between the first and the last failure in the history, and k is the number of failures in the history. If k is smaller than K, the estimate MUST be computed as if there was a failure at the current time. It has been shown that an estimate calculated in a similar manner is accurate within 17% of the real value of U [ghinita2006].

The size of the failure history K affects the accuracy of the estimate of U. One can increase the accuracy by increasing K. However, this has the side effect of decreasing responsiveness to changes in the failure rate. On the other hand, a small history size may cause a peer to overreact each time a new failure occurs. In [ghinita2006], K is set 25% of the routing table size. Use of this approach is RECOMMENDED.

### 6.3.1. Detecting Failures

A new failure MUST be inserted to the failure history in the following cases:

1. A Leave request is received from a neigbhor.
2. A peer fails to reply to a Ping request sent in the situation explained below. If no packets have been received on a connection during the past 2*Tr seconds (where Tr is the inactivity timer defined by ICE [I-D.ietf-mmusic-ice]), a RELOAD Ping request MUST be sent to the remote peer. RELOAD mandates the use of STUN [RFC5389] for keepalives. STUN keepalives take the form of STUN Binding Indication transactions. As specified in ICE [I-D.ietf-mmusic-ice], a peer sends a STUN Binding Indication if there has been no packet sent on a connection for Tr seconds. Tr is configurable and has a default of 15 seconds. Although STUN Binding Indications do not generate a response, the fact that a peer has failed can be learned from the lack of packets (Binding Indications or application protocol packets) received from the peer. If the remote peer fails to reply to the Ping request, the sender MUST consider the remote peer to have failed.

As an alternative to relying on STUN keepalives to detect peer failure, a peer could send additional, frequent RELOAD messages to every peer in its Connection Table. These messages could be Update requests, in which case they would serve two purposes: detecting peer failure and stabilization. However, as the cost of this approach can be very high in terms of bandwidth consumption and traffic load, especially in large-scale overlays experiencing churn, its use is NOT RECOMMENDED.

### 6.4. Estimating Join Rate

Reference [ghinita2006] proposes that a peer can estimate the join rate based on the uptime of the peers in its routing table. An increase in peer join rate will be reflected by a decrease in the average age of peers in the routing table. Thus, each peer MUST maintain an array of the ages of the peers in its routing table sorted in increasing order. Using this information, an estimate of the global peer join rate L MUST be calculated as:

$$L = \frac{N}{4} * \frac{1}{Ages[rsize/4]},$$

where Ages is an array containing the ages of the peers in the routing table sorted in increasing order and rsize is the size of the routing table. It is RECOMMENDED that only the ages of the 25% of the youngest peers in the routing table (i.e., the 25th percentile) are used to reduce the bias that a small number of peers with very old ages can cause [ghinita2006]. It has been shown that the estimate obtained by using this method is accurate within 22% of the real join rate [ghinita2006]. Of course, the size of the routing table affects the accuracy.

In order for this mechanism to work, peers need to exchange information about the time they have been present in the overlay. Peers receive the uptimes of their successors and predecessors during the stabilization operations since all Update requests carry uptime values. A joining peer learns the uptime of the admitting peer since it receives an Update from the admitting peer during the join procedure. Peers learn the uptimes of new fingers since they can fetch the uptime using a Probe request after having attached to the new finger.

### 6.5. Estimate Sharing

To improve the accuracy of network size, join rate, and leave rate estimates, peers MUST share their estimates. When the stabilization timer fires, a peer MUST select number-of-peers-to-probe random peers from its finger table and send each of them a Probe request. The targets of Probe requests are selected from the finger table rather than from the neighbor table since neighbors are likely to make similar errors when calculating their estimates. number-of-peers-to-probe is a new element in the overlay configuration document. It is defined in Section 7 and has a default value of 4. Both the Probe request and the answer returned by the target peer MUST contain a new message extension whose MessageExtensionType is 'self_tuning_data'. This extension type is defined in Section 9.1. The

extension_contents field of the MessageExtension structure MUST contain a SelfTuningData structure:

```
struct {
  uint32                  network_size;
  uint32                  join_rate;
  uint32                  leave_rate;
} SelfTuningData;
```

The contents of the SelfTuningData structure are as follows:

network_size
The latest network size estimate calculated by the sender.

join_rate
The latest join rate estimate calculated by the sender.

leave_rate
The latest leave rate estimate calculated by the sender.

The join and leave rates are expressed as joins or failures per 24 hours. As an example, if the global join rate estimate a peer has calculated is 0.123 peers/s, it would include in the join_rate field the value 10627 (24*60*60*0.123 = 10627.2).

The 'type' field of the MessageExtension structure MUST be set to contain the value 'self_tuning_data'. The 'critical' field of the structure MUST be set to False.

A peer MUST store all estimates it receives in Probe requests and answers during a stabilization interval. When the stabilization timer fires, the peer MUST calculate the estimates to be used during the next stabilization interval by taking the 75th percentile of a data set containing its own estimate and the received estimates.

## 6.6. Calculating the Stabilization Interval

According to [liben-nowell2002], a Chord network in a ring-like state remains in a ring-like state as long as peers send Omega(log2^2(N)) messages before N new peers join or N/2 peers fail. We can use the estimate of peer failure rate, U, to calculate the time Tf in which N/2 peers fail:

$$Tf = \frac{1}{2*U}$$

Based on this estimate, a stabilization interval Tstab-1 MUST be calculated as:

$$\text{Tstab-1} = \frac{Tf}{\log2^2(N)}$$

On the other hand, the estimated join rate L can be used to calculate the time in which N new peers join the overlay. Based on the estimate of L, a stabilization interval Tstab-2 MUST be calculated as:

$$\text{Tstab-2} = \frac{N}{L * \log2^2(N)}$$

Finally, the actual stabilization interval Tstab that MUST be used can be obtained by taking the minimum of Tstab-1 and Tstab-2.

The results obtained in [maenpaa2009] indicate that making the stabilization interval too small has the effect of making the overlay less stable (e.g., in terms of detected loops and path failures). Thus, a lower limit should be used for the stabilization period. Based on the results in [maenpaa2009], a lower limit of 15s is RECOMMENDED, since using a stabilization period smaller than this will with a high probability cause too much traffic in the overlay.

## 7. Overlay Configuration Document Extension

This document extends the RELOAD overlay configuration document by adding one new element, "number-of-peers-to-probe", inside each "configuration" element.

number-of-peers-to-probe: The number of fingers to which Probe requests are sent to obtain their network size, join rate, and leave rate estimates. The default value is 4.

## 8. Security Considerations

There are no new security considerations introduced in this document. The security considerations of RELOAD [I-D.ietf-p2psip-base] apply.

## 9. IANA Considerations

### 9.1. Message Extensions

This document introduces one additional extension to the "RELOAD Extensions" Registry:

| Extension Name | Code | Specification |
|---|---|---|
| self_tuning_data | 1 | RFC-AAAA |

The contents of the extension are defined in Section 6.5.

## 10. References

### 10.1. Normative References

[I-D.ietf-mmusic-ice]
Rosenberg, J., "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols", draft-ietf-mmusic-ice-19 (work in progress), October 2007.

[I-D.ietf-p2psip-base]
Jennings, C., Lowekamp, B., Rescorla, E., Baset, S., and H. Schulzrinne, "REsource LOcation And Discovery (RELOAD) Base Protocol", draft-ietf-p2psip-base-07 (work in progress), February 2010.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5389] Rosenberg, J., Mahy, R., Matthews, P., and D. Wing, "Session Traversal Utilities for NAT (STUN)", RFC 5389, October 2008.

### 10.2. Informative References

[CAN] Ratnasamy, S., Francis, P., Handley, M., Karp, R., and S. Schenker, "A scalable content-addressable network", In Proc. of the 2001 Conference on Applications, Technologies, Architectures and Protocols for Computer Communications 2001, pp. 161.172.

[Chord] Stoica, I., Morris, R., Liben-Nowell, D., Karger, D., Kaashoek, M., Dabek, F., and H. Balakrishnan, "Chord: A

Scalable Peer-to-peer Lookup Service for Internet Applications", IEEE/ACM Transactions on Networking Volume 11, Issue 1, 17-32, Feb 2003.

[I-D.ietf-p2psip-concepts]
Bryan, D., Matthews, P., Shim, E., Willis, D., and S. Dawkins, "Concepts and Terminology for Peer to Peer SIP", draft-ietf-p2psip-concepts-02 (work in progress), July 2008.

[Pastry] Rowstron, A. and P. Druschel, "Pastry: Scalable, Decentralized Object Location and Routing for Large-Scale Peer-to-Peer Systems", In Proc. of the IFIP/ACM International Conference on Distribued Systems Platforms Nov. 2001, pp. 329-350.

[binzenhofer2006]
Binzenhofer, A., Kunzmann, G., and R. Henjes, "A scalable algorithm to monitor chord-based P2P systems at runtime", 20th International Parallel and Distributed Processing Symposium April 2006.

[ghinita2006]
Ghinita, G. and Y. Teo, "An adaptive stabilization framework for distributed hash tables", 20th International Parallel and Distributed Processing Symposium April 2006.

[horowitz2003]
Horowitz, K. and D. Malkhi, "Estimating network size from local information", Information Processing Letters Dec. 2003, Volume 88, Issue 5, pp. 237-243.

[kostoulas2005]
Kostoulas, D., Psaltoulis, D., Gupta, I., Birman, K., and A. Demers, "Decentralized schemes for size estimation in large and dynamic groups", Fourth IEEE International Symposium on Network Computing and Applications July 2005, pp. 41-48.

[krishnamurthy2008]
Krishnamurthy, S., El-Ansary, S., Aurell, E., and S. Haridi, "Comparing maintenance strategies for overlays", In Proc. of 16th Euromicro Conference on Parallel, Distributed and Network-Based Processing Feb. 2008, pp. 473-482.

[li2004] Li, J., Strinbling, J., Gil, T., and M. Kaashoek, "Comparing the performance of distributed hash tables

under churn", In Proc. of the 3rd International Workshop on Peer-to-Peer Systems 2004.

[liben-nowell2002]
Liben-Nowell, D., Balakrishnan, H., and D. Karger, "Observations on the dynamic evolution of peer-to-peer networks", In Proc. of the First International Workshop on Peer-to-Peer Systems March 2002.

[maenpaa2009]
Maenpaa, J. and G. Camarillo, "A study on maintenance operations in a Chord-based Peer-to-Peer Session Initiation Protocol overlay network", In Proc. of IPDPS 2009 May 2009.

[mahajan2003]
Mahajan, R., Castro, M., and A. Rowstron, "Controlling the cost of reliability in peer-to-peer overlays", In Proceedings of the 2nd International Workshop on Peer-to-Peer Systems Feb. 2003.

[rhea2004]
Rhea, S., Geels, D., Roscoe, T., and J. Kubiatowicz, "Handling churn in a DHT", In Proc. of the USENIX Annual Techincal Conference June 2004.

## Authors' Addresses

Jouni Maenpaa
Ericsson
Hirsalantie 11
Jorvas 02420
Finland

Email: Jouni.Maenpaa@ericsson.com

Gonzalo Camarillo
Ericsson
Hirsalantie 11
Jorvas 02420
Finland

Email: Gonzalo.Camarillo@ericsson.com

Jani Hautakorpi
Ericsson
Hirsalantie 11
Jorvas 02420
Finland

Email: Jani.Hautakorpi@ericsson.com