Scalable Computing: Practice and Experience
Volume 11, Number 3, pp. 239–249. http://www.scpe.org
ISSN 1895-1767
© 2010 SCPE

DESIGN AND ANALYSIS OF A SCALABLE ALGORITHM TO MONITOR CHORD-BASED P2P SYSTEMS AT RUNTIME*

ANDREAS BINZENHÖFER[†], GERALD KUNZMANN[‡], AND ROBERT HENJES[†]

Abstract. Peer-to-peer (p2p) systems are a highly decentralized, fault tolerant, and cost effective alternative to the classic client-server architecture. Yet companies hesitate to use p2p algorithms to build new applications. Due to the decentralized nature of such a p2p system the carrier does not know anything about the current size, performance, and stability of its application. In this paper we present an entirely distributed and scalable algorithm to monitor a running p2p network. The snapshot of the system enables a telecommunication carrier to gather information about the current performance parameters of the running system as well as to react to discovered errors.

1. Introduction. In recent years peer-to-peer (p2p) algorithms have widely been used throughout the Internet. So far, the success of the p2p paradigm was mainly driven by file sharing applications. However, despite their reputation p2p mechanisms offer the solution to many problems faced by telecommunication carriers today [8]. Compared to the classic client-server architecture they are decentralized, fault tolerant, and cost effective alternatives. Those systems are highly scalable, do not suffer from a single point of failure, and require less administration overhead than existing solutions. In fact, there are more and more successful p2p based applications like Skype [14], a distributed VoIP solution, Oceanstore [4], a global persistent data store, and even p2p-based network management [10].

One of the main reasons why telecommunication carriers are still hesitant to build p2p applications is the lack of control a provider has over the running system. At first, the system appears as a black box to its operator. The carrier does not know anything about the current size, performance, and stability of its application. The decentralized nature of such a system makes it hard to find a scalable way to gather information about the running system at a central unit. Operators, however, do not want to lose control over their systems. They want to know what their systems look like right now and where problems occur at the moment. The first problems already occur when testing and debugging a distributed application. Finding implementation errors in a highly distributed system is a very complex and time consuming process [9]. A provider also needs to know whether his newly deployed application can truly handle the task it was designed for.

The latest generation of p2p algorithms is based on distributed hash tables (DHTs). The algorithm that currently attracts the most attention is Chord, which uses a ring topology to realize the underlying DHT [12]. DHTs are theoretically understood in depth and proved to be a scalable and robust basis for distributed applications [7]. However, the problem of monitoring such a system from a central location is far from being solved. [11] gives a good overview of different approaches to monitor and debug distributed systems in general. In the field of p2p, the process of measuring and monitoring a running system was so far limited to unstructured overlays. [13], e.g., introduces a crawling-based approach to query Gnutella-like networks.

In this paper, however, we exploit the special features of structured p2p overlays and present an entirely novel and scalable approach to create a snapshot of a running Chord-based network. Using our algorithm a provider can either monitor the entire system or just survey a specific part of the system. This way, he is able to react to errors more quickly and can verify if the taken countermeasures are successful. On the basis of the gathered information it is, e.g., possible to take appropriate action to relief a hotspot or to pinpoint the cause of a loss of the overlay ring structure. The overhead involved in creating the snapshot is evenly distributed to the participating peers so that each peer only has to contribute a negligible amount of bandwidth. Most importantly, the snapshot algorithm is very easy to use for a provider. It only takes one parameter to adjust the trade off between duration of the snapshot and bandwidth needed at the central unit which collects the measurements.

The remainder of this paper is structured as follows. Section 2 gives a brief overview of Chord with a focus on aspects relevant to this paper. The snapshot algorithm as well as some areas of application are described in Section 3. The functionality of the algorithm is verified analytically in Section 4 and by simulation in Section 5. Section 6 concludes this paper.

*Corresponding mail: binzenhoefer@informatik.uni-wuerzburg.de
[†]University of Würzburg, Institute of Computer Science, Germany.
[‡]Technical University of Munich, Institute of Communication Networks, Germany.

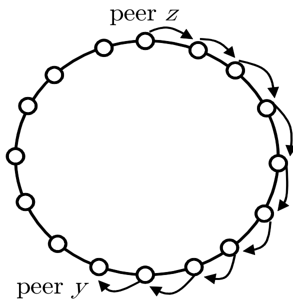

Fig. 2.1. *A simple search.*

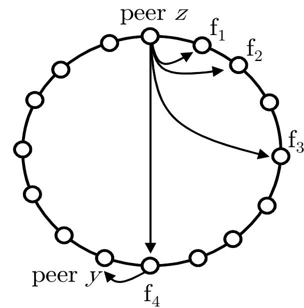

Fig. 2.2. *Search using the fingers.*

2. Chord Basics. This section gives a brief overview of Chord with a focus on aspects relevant to this paper. A more detailed description can be found in [12]. The main purpose of p2p networks is to store data in a decentralized overlay network. Participating peers will then be able to retrieve this data using some sort of search algorithm. The Chord algorithm solves this problem by arranging the participating peers on a ring topology. The position id_z of a peer z on this overlay ring is determined by an m-bit identifier generated by a hash function such as SHA-1 or MD5. In a Chord ring each peer knows at least the id of its immediate successor in a clockwise direction on the ring. This way, a peer looking up another peer or a resource is able to pass the query around the circle using its successor pointers. Figure 2.1 illustrates a simple search of peer z for another peer y using only the immediate successor. The search has to be forwarded half-way around the ring. Obviously, the average search would require $\frac{n}{2}$ overlay hops, where n is the current size of the Chord ring. To speed up searches a peer z in a Chord ring also maintains pointers to other peers, which are used as shortcuts through the ring. Those pointers are called fingers, whereby the *i-th* finger in a peer's finger table contains the identity of the first peer that succeeds z's own id by at least 2^{i-1} on the Chord ring. That is, peer z with hash value id_z has its fingers pointing to the first peers that succeed $\left(id_z + 2^{i-1}\right) \bmod 2^m$ for $i = 1$ to m, where 2^m is the size of the identifier space.

Figure 2.2 shows fingers f_1 to f_4 for peer z. Using this finger pointers, the same search does only take two overlay hops. For the first hop peer z uses its finger f_4. Peer y can then directly be reached using the successor of f_4 as indicated by the small arrow. This way, a search only requires $\frac{1}{2}\log_2(n)$ overlay hops on average. A detailed mathematical analysis of the search delay in Chord rings can be found in [3]. The snapshot algorithm presented in Section 3 makes use of the finger tables of the peers.

3. Design of the Snapshot Algorithm. In this section we introduce a scalable and distributed algorithm to create a snapshot of a running Chord system. The algorithm is based on a very simple two step approach. In step one, the overlay is recursively divided into subparts of a predefined size. In step two, the desired measurement is done for each of these subparts and sent back to a central collecting point (CP). In the following, we describe both steps in detail.

3.1. Step 1: Divide Overlay into Subparts. The algorithm *snapshot(R_s, R_e, S_{min}, CP)* divides a specific region of the overlay into subparts. This function is called at an arbitrary peer p with id_p. The peer then tries to divide the region from $R_s = id_p$ to R_e into contiguous subparts using its fingers. The exact procedure is illustrated in Figure 3.1. In this example peer p has four fingers f_1 to f_4. It sends a request to the finger closest to R_e within $[R_s; R_e]$. At first, finger f_4 is disregarded since it does not fall into the region between R_s and R_e (cf. a). This makes f_3 the closest finger to R_e in our example. If this finger does not respond to the request, as illustrated by the bolt (cf. b), it is removed from the peer's finger list and the peer tries to contact the next closest finger f_2 (cf. c). If this finger acknowledges the request, peer p recursively tries to divide the region from $R_s = id_p$ to $\widehat{R}_e = id_{f_2} - 1$ into contiguous subparts. Finger f_2 partitions the region from $\widehat{R}_s = id_{f_2}$ to R_e accordingly.

As soon as a peer does not know any more fingers in the region between the current R_s and the current R_e, the recursion is stopped. The peer changes into step two of the algorithm and starts a measurement of this specific region. In this context, the parameter S_{min} can be used to determine the minimum size of the regions, which will be measured in step two. Taking into account S_{min}, a peer will already start the measurement if it

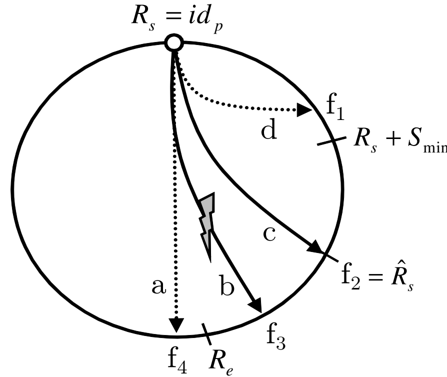

FIG. 3.1. *Visualization of the algorithm.*

does not know any more fingers in the region between the current $R_s + S_{min}$ and the current R_e. In this case finger f_1 would be disregarded, as illustrated by the dotted line (cf. d in Figure 3.1), since it points into the minimum measurement region. Parameter S_{min} is designed to adjust the trade off between the duration of the snapshot and the bandwidth needed at the collecting point. The larger the regions in step two, the longer the measurement will take. The smaller the regions, the more results are sent back to the CP.

Algorithm 2
The snapshot algorithm (first call $R_s = id_p$)

snapshot(R_s, R_e, S_{min}, CP)
send acknowledgment to the sender of the request
$id_{fm} = max(\{id_f | id_f \in \text{fingerlist} \wedge id_f < R_e\})$
while $id_{fm} > R_s + S_{min}$ **do**
  send snapshot(id_{fm}, R_e, S_{min}, CP) request to peer id_{fm}
  if acknowledgment from id_{fm} **then**
    call snapshot(id_p, $id_{fm} - 1$, S_{min}, CP) at local peer
    return //exit the function
  else
    remove id_{fm} from fingerlist
    $id_{fm} = max(\{id_f | id_f \in \text{fingerlist} \wedge id_f < R_e\})$
  end if
end while
$\widehat{S} = \frac{R_e - R_s}{\lceil \frac{R_e - R_s}{S_{min}} \rceil}$ //explanation see step two
call countingtoken(id_p, R_e, S_{min}, CP, $\emptyset$) at local peer

A detailed technical description of the procedure is given in Algorithm 2. Peer p will contact the closest finger to R_e until it does not know any more fingers in between $R_s + S_{min}$ and R_e. If so, it changes into step two and starts a measurement of this region calling the function countingtoken(id_p, R_e, S_{min}, CP, *result*) at the local peer.

3.2. Step 2: Measure a Specific Subpart. The basic idea behind the measurement of a specific subpart from R_s to R_e is very simple. The first peer creates a token, adds its local statistics, and passes the token to its immediate successor. The successor proceeds recursively until the first peer with an $id > R_e$ is reached. This peer sends the token back to the collecting point, whose IP is given in the parameter CP.

Ideally, each of the regions measured in step two would be of size S_{min} as specified by the user. The problem, however, is that the region from R_s to R_e is slightly larger than S_{min} according to step one of the algorithm. In fact, if the responsible peer did not know enough fingers, the region might even be significantly larger than S_{min}. The solution to this problem is to introduce checkpoints with a distance of S_{min} in the corresponding region. Results are sent to the CP every time the token passes a checkpoint instead of sending only one answer

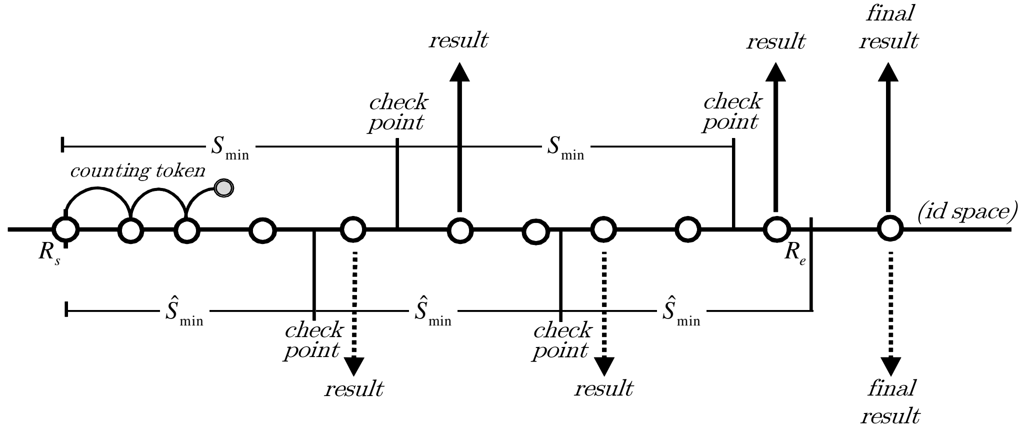

FIG. 3.2. *Results sent after each checkpoint.*

at the end of the region. This is illustrated in the upper part of Figure 3.2. The counting token is started at R_s. The first peer behind each checkpoint sends a *result* back to the CP as illustrated by the large solid arrows. The final *result* is still sent by the first peer with $id > R_e$.

A drawback of this solution is that the checkpoints might not be equally distributed in the region. In particular, the last two checkpoints might be very close to each other. We therefore recalculate the positions of the checkpoints according to the following equation:

$$\widehat{S}_{min} = \frac{R_e - R_s}{\left\lceil \frac{R_e - R_s}{S_{min}} \right\rceil}.$$

The new checkpoints can be seen in the lower part of Figure 3.2. The number of checkpoints remains the same, while their positions are moved in such a way, that the results are now sent at equal distance.

As can be seen at the end of Algorithm 2, the recalculation of S_{min} is already done in the first step, just before the counting token is started. A detailed description of the counting token mechanism is given in Algorithm 3. If a peer p receives a counting token it makes sure that its identifier is still within the measured region, i.e. $R_s \leq id_p \leq R_e$. If not, it sends a *result* back to the CP and stops the token. Otherwise it adds its local measurement to the token and tries to pass the token to its immediate successor. If it is the first peer behind one of the checkpoints, it sends an intermediate result back to the CP and resets the token.

As mentioned above the parameter S_{min} roughly determines the minimum size of the regions measured in step two. If S_{id} is the total size of the identifier space, there will be N_c counting tokens arriving at the CP, whereas:

$$2 \cdot \left\lceil \frac{S_{id}}{S_{min}} \right\rceil \geq N_c \geq \left\lceil \frac{S_{id}}{S_{min}} \right\rceil.$$

A more detailed analysis of the snapshot algorithm is given in Section 4 as well as in [1].

3.3. Collect Statistics. Generally speaking, there are two different kinds of statistics, which can be collected using the counting tokens. Either a simple mean value or a more detailed histogram. In the first case the counting token memorizes two variables, V_a for the accumulated value and V_n for the number of values. Each peer receiving the counting token adds its measured value to V_a and increases V_n by one. The sample mean can then be calculated at the CP as $\frac{\sum V_a}{\sum V_n}$. In case of a histogram, the counting token maintains a specific number of bins and their corresponding limits. Each peer simply increases the bin matching its measured value by one. If the measured value is outside the limits of the bins it simply increases the first or the last bin respectively.

There are numerous things that can be measured using the above mentioned methods. Table 3.1 summarizes some exemplary statistics and the kind of information which can be gained from them. The most obvious application is to count the number of hops for each counting token. On the one hand, this is a direct measure for the size of the overlay network. On the other hand, it also shows the distribution of the identifiers in the

Algorithm 3
The countingtoken algorithm (first call $R_s = id_p$)

countingtoken(R_s, R_e, S_{min}, CP, $result$)
send acknowledgment to the sender of the request
if $R_s \leq id_p \leq R_e$ **then**
  if $id_p > R_s + S_{min}$ **then**
    send $result$ to CP
    $result = 0$
    $R_s = R_s + S_{min}$
  end if
  add local measurement to $result$
  $id_s = id$ of direct successor
  while 1 **do**
    send countingtoken(R_s, R_e, S_{min}, CP, $result$) request to direct successor id_s
    if acknowledgment **then**
      break
    else
      remove id_s from successor list
      $id_s = id$ of new direct successor
    end if
  end while
else
  send $result$ to CP
end if

TABLE 3.1
Possible statistics gathered during snapshot

Statistic	Information gained
Number of hops per token	Size of the network, Distribution of the identifiers
Mean search delay	Performance of the algorithm
Sender $\stackrel{?}{==}$ predecessor	Overlay stability
Number of timeouts per token	Churn rate
Number of resources per peer	Fairness of the algorithm
Number of searches answered	User behavior
Bandwidth used per time unit	Maintenance overhead
Missing resources	Data integrity

identifier space. To gain information about the performance of the Chord algorithm, the mean search delay or a histogram for the search time distribution can be calculated and compared to expected values. Furthermore, Chord's stability can only be guaranteed as long as the successor and predecessor pointers of the individual peers match each other correspondingly. This invariant can be checked by counting the percentage of hops, where the sender of the counting token did not match the predecessor of the receiving peer. Additionally, the number of timeouts per token can be used to measure the current churn rate in the overlay network. The more churn there is, the more timeouts are going to occur due to outdated successor pointers. Similarly, the number of resources stored at each peer is a sign of the fairness of the Chord algorithm. The number of searches answered at each peer can likewise be used to get an idea of the search behavior of the end users. Finally, a peer can keep track of the number of missing resources to verify the integrity of the stored data. This can, e.g., be done counting the number of search requests which could not be answered by the peer.

All of the above statistics can be collected periodically to survey the time dependent status of the overlay. Note, that it is also possible to monitor a specific part of the overlay network by setting R_s and R_e accordingly. This can, e.g., be helpful if there are problems in a certain region of the overlay network and the operator needs to verify that his countermeasures have been successful.

4. Analysis and Optimizations. To analyze our algorithm we derive the duration of a snapshot (cf. Subsection 4.1) and the temporal distribution of the token arrival times at the CP (cf. Subsection 4.2).

4.1. Duration of a Snapshot. To calculate an estimate of the duration of a snapshot, we assume a scenario without any peers joining or leaving the network. It is quite straightforward to estimate the duration of step one, the signaling step. The last counting token which will be started is the one covering the region directly following the initiating peer. This is due to the fact, that the initiating peer will start its counting token no sooner than it divided the ring into separate regions. Before it initiates the counting token, it contacts its fingers until the first finger is closer to itself than S_{min}. The initiating peer has at most $\log_2(n)$ fingers and each of the fingers sends an acknowledgment, before the peer can go on with the algorithm. If T_O is the random variable describing one overlay hop, then the duration of step one of the algorithm is at most

$$D_{step1} = 2 \cdot \log_2(n) \cdot E[T_O]. \quad (4.1)$$

The worst case for step two would be that the initiating peer does not know any fingers and directly sends the counting token. This would take $n \cdot E[T_O]$, but is very unlikely to happen. In general, if there are n peers in the overlay, there are roughly $P_r = \frac{n \cdot S_{min}}{S_{id}}$ peers per region. Furthermore, in the worst case S_{min} is slightly larger than a power of two and the region covered by a counting token may become almost twice as large as S_{min}. Therefore a good estimate for the duration of the counting step of the algorithm is:

$$D_{step2} = 2 \cdot P_r \cdot E[T_O]. \quad (4.2)$$

This results in the following total duration of a snapshot:

$$D = \left(\log_2(n) + \frac{n \cdot S_{min}}{S_{id}}\right) \cdot 2 \cdot E[T_O]. \quad (4.3)$$

4.2. Token Arrival Time Distribution. To get a rough estimate for the distribution of the arrival times of the counting tokens at the CP, we consider the special case where the size of the overlay $n = 2^g$ is a power of two and S_{min} is such that $N_r = 2^h$ with $h < g$. Furthermore, we assume that the individual peers are located at equal distances on the ring as shown in Figure 4.1.

It can be shown, that in this case $h = \log_2(N_r)$ is the number of overlay hops it takes until the first counting token is started during a snapshot. Similarly, it takes $2 \cdot h$ hops until the last counting token is started according to our assumptions. The probability p_i that a counting token is started after exactly i hops for $i = h, h+1, ..., 2 \cdot h$ can be calculated as:

$$p_i = \frac{\binom{h}{i-h}}{\sum_{x=h}^{2 \cdot h} \binom{h}{x-h}}. \quad (4.4)$$

The above considerations are nontrivial, but can nicely be explained using the simple example shown in Figure 4.1, where $g = 4$, $h = 2$, and therefore $n = 2^4$ and $N_r = 2^2$. The solid arrows in the figure show the messages of the signaling step, the dotted arrows the corresponding acknowledgments. The numbers next to the arrows represent the number of overlay hops, which have passed since the beginning of the snapshot.

In the example, peer A starts a snapshot of the entire ring. It sends a request to B to cover the region between B and A. Peer B sends an acknowledgment back to A and a simultaneous request to C to cover the region from C to A. C has no fingers outside its minimum measurement region and starts the first counting token after $h = 2$ overlay hops. Simultaneously, it sends an acknowledgment back to B. Peer B then starts its counting token after 3 overlay hops. In the meantime A signals D to cover the region from D to B. Peer D immediately starts its counting token after a total of 3 overlay hops. Peer A waits for the final acknowledgment and starts its counting token after $4 = 2 \cdot h$ overlay hops. Summarizing the above, there are four counting tokens started after 2, 3, 3, and 4 overlay hops respectively.

According to our assumptions, each counting token needs exactly $P_r = 4$ hops to travel the corresponding region and one final hop to arrive at the CP. A rough estimate for the distribution of the arrival times of the counting tokens at the CP is therefore given by the phase diagram shown in Figure 4.3. It indicates that the signaling step takes i overlay hops with a probability p_i for $i = h, h+1, ..., 2 \cdot h$, which is followed by P_r hops of the counting token and the final hop to report the result back to the CP.

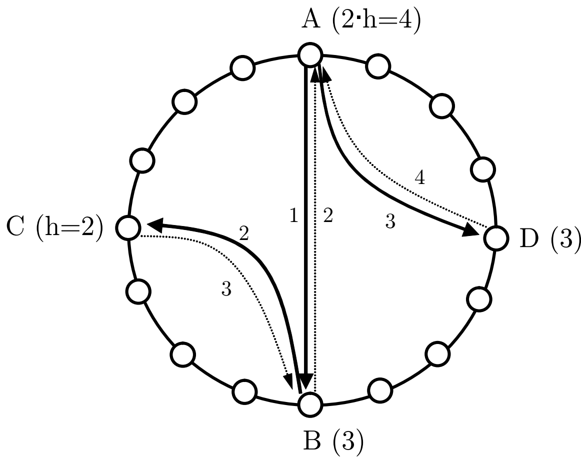

Fig. 4.1. *Starting times of the counting tokens for $N_r = 2^2$ and $n = 2^4$.*

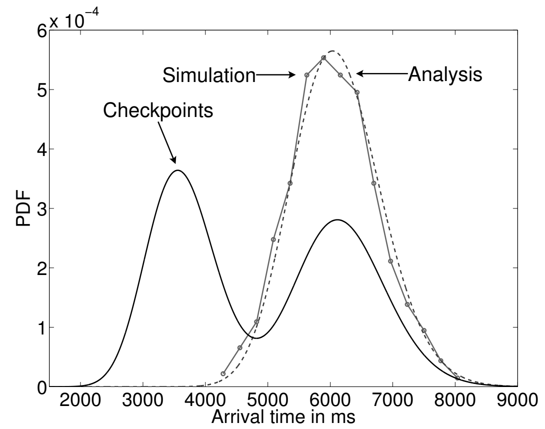

Fig. 4.2. *Probability density function of the token arrival time.*

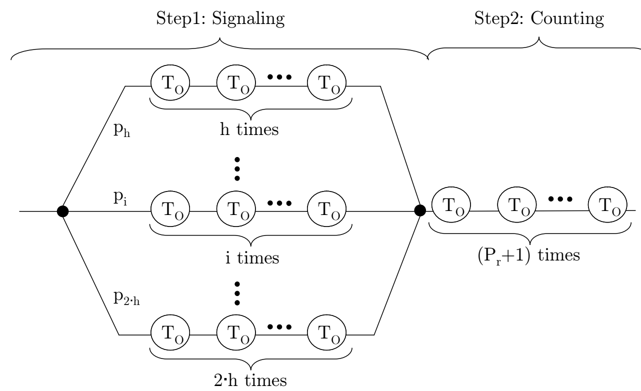

Fig. 4.3. *Phase diagram of the token arrival time distribution.*

To validate our analytical results, we simulated a Chord ring of size $n = 2^{15}$ with $S_{min} = 2^9$ according to the above assumptions. Figure 4.2 shows the probability density function of the token arrival times at the CP. Obviously, the curves match very well and the binomial distribution of the duration of step one can be recognized. So far, in our example each peer has a finger at an exact distance of S_{min}. In reality, however, the finger would sit at a slightly different position, which again would result in an additional checkpoint at the middle of the region. The curve labeled "Checkpoints" corresponds to a slightly modified phase diagram, which adds an intermediate result in the middle of the measurement region. The first rise of the probability density function (pdf) therefore represents the intermediate results sent back to the CP at the checkpoint. The second rise still represents the regular results at the end of the region. In the following section we will present simulations of more realistic scenarios including churn and timeouts.

5. Results. The results in this section were obtained using our ANSI-C simulator, which incorporates a detailed yet slightly modified Chord implementation. A good description of the general simulation model can be found in [5, 6]. In this work an overlay hop is modeled using an exponentially distributed random variable with a mean of 80ms. The results considering churn are generated using peers, which stay online and offline for an exponentially distributed period of time with a mean as indicated in the corresponding description of the figures.

The snapshot algorithm takes one single input argument S_{min} which directly translates into $N_r = \left\lceil \frac{S_{id}}{S_{min}} \right\rceil$, the number of areas the overlay will be divided into. This parameter influences the duration of the snapshot as well as the number of results arriving at the central collecting point. Figure 5.1 shows the distribution of the

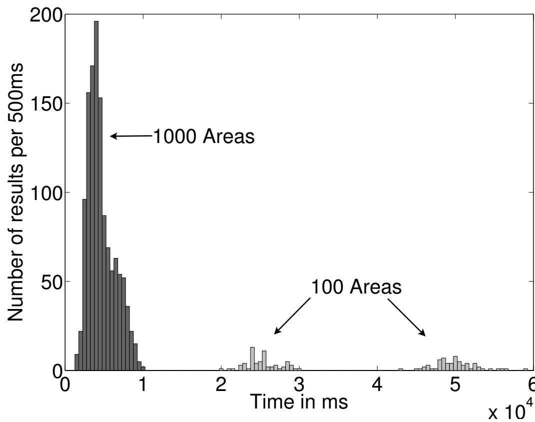

FIG. 5.1. *Arrival times of the results.*

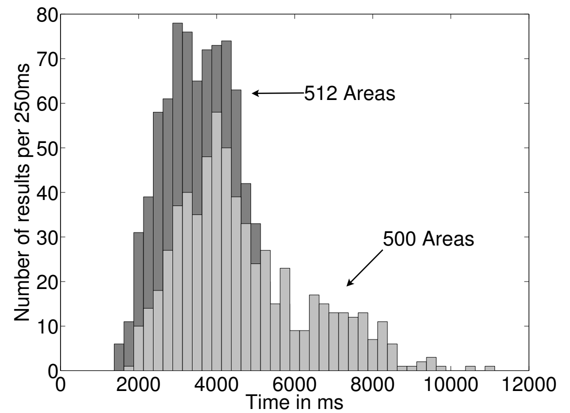

FIG. 5.2. *Influence of* N_r *for* 20000 *peers.*

arrival times of the results in an overlay of 40000 peers using $N_r = 1000$ and $N_r = 100$ areas in times of no churn. Obviously, the more areas the overlay is divided into, the faster the snapshot is completed. While the snapshot using 1000 areas was finished after about ten seconds, the snapshot with 100 areas took about one minute. In exchange the latter snapshot produces significantly smaller bandwidth spikes at the CP. The two elevations of the second histogram correspond to the intermediate results (first elevation) and the final results at the end of the measured subpart (second elevation). Note that the final results arrive about twice as late as the intermediate results. The first step of the algorithm uses the fingers to divide the ring into subparts. Since the distance between a peer and its fingers is always slightly larger than a power of two it is usually not possible to divide the ring exactly into the desired number of areas. In fact it is very likely, that a peer stops the recursion and starts its measurement once it contacted its xth finger, where $2^{x-1} < S_{min} = \frac{S_{id}}{N_r} \leq 2^x$. That is, the recursion stops at finger x with id_{f_x}, whereas the distance between the peer and this specific finger might almost be twice as large as the desired S_{min}. It is therefore advisable to choose N_r as a power of two itself in order to ensure that id_{f_x} is only slightly larger than $id_p + S_{min}$. Figure 5.2 shows the different arrival times of the results for $N_r = 512$ and $N_r = 500$ in an overlay of 20000 peers without churn. The skewed shape of the histogram in the foreground results from the fact that 500 is slightly smaller than a power of two, which in turn makes S_{min} slightly larger than a power of two. In this case it is likely that the peer has a finger just before the end of the minimum measurement region $id_p + S_{min}$. Thus, finger x sits at a distance of about twice S_{min} from the peer. The resulting counting token will therefore travel a distance of about twice S_{min} as well.

A more detailed analysis of the influence of N_r can be found in Figure 5.3, which shows the number of results received at the CP in dependence of N_r. As shown in [1], N_c, the number of counting tokens sent to the CP, is limited by $2 \cdot N_r > N_c \geq N_r$. The straight lines in the figure show the corresponding limits. The solid and dotted curves represent the results obtained for 20000 and 10000 peers, respectively. The number of results sent to the CP is within the calculated limits and independent of the overlay size. The curves roughly resemble the shape of a staircase, whereas the steps are located at powers of two. There are two main reasons for this behavior. First of all, the average counting token sends two results back to the CP, one intermediate result and the final result at the end of the measurement region. Hence, the smaller the region covered by the average counting token, the more results arrive at the CP. As explained above, the closer N_r gets to a power of two, the smaller the region covered by the average counting token. This accounts for the first part of the rise of the number of results received at the CP.

The distribution of the arrival times of the results is also influenced by the current size of the network. The larger the network, the more peers are within one region. However, the more peers are within one region, the more hops each counting token has to make, before it can send its results back to the CP. Figure 5.4 shows the token arrival time distribution for three different overlay sizes of 10000, 20000, and 40000 peers, respectively. We did not generate any churn in this scenario and set $N_r = 512$ areas. As expected, the larger the overlay network, the longer the snapshot is going to take. However, the curves are not only shifted to the right, but also differ in shape. This can again be explained by the increasing number of hops per counting token.

As mentioned above, the average counting token sends two results back to the CP, whereas the checkpoints are equally spaced. Thus, the final result takes twice as many hops as the intermediate result. In a network of

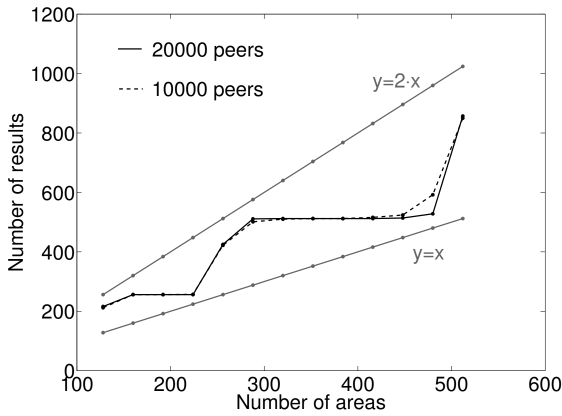

Fig. 5.3. *Number of results received at the CP.*

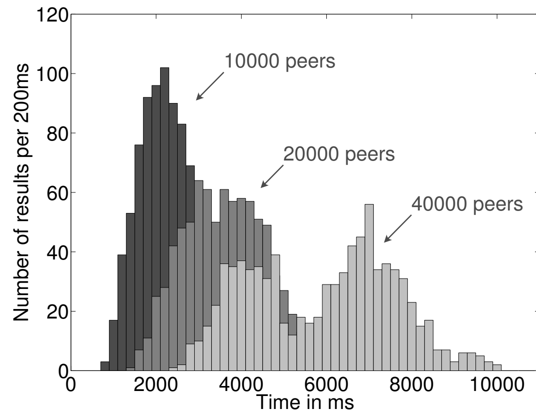

Fig. 5.4. *Arrival times of the results at the CP.*

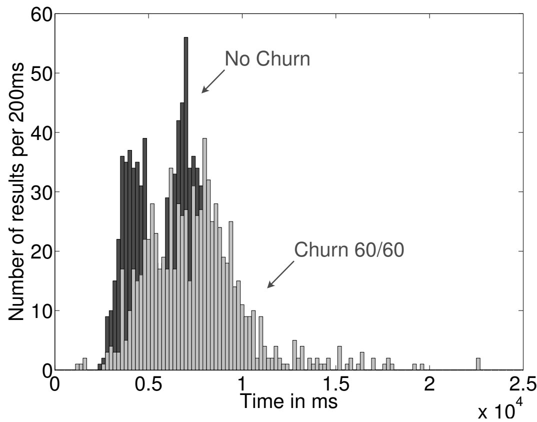

Fig. 5.5. *Influence of churn on the traffic pattern at the CP.*

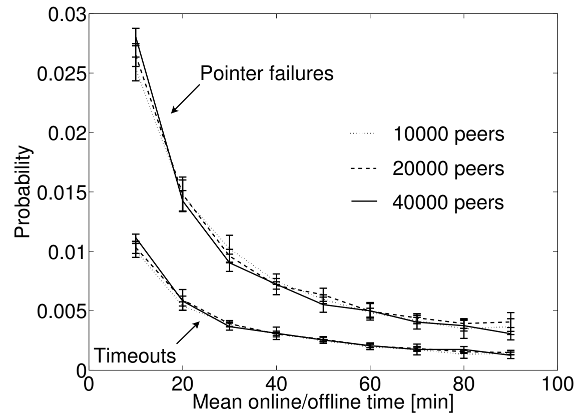

Fig. 5.6. *Relative frequency of timeouts and pointer failures.*

10000 peers there are approximately 20 peers in each of the 512 regions. The intermediate results are therefore sent after about 10 hops, the final results after about 20 hops, respectively. The two corresponding elevations in the histogram overlap in such a way, that they build a single elevation. In a network of 40000 peers, however, there are approximately 78 peers in each of the 512 regions. The intermediate results are therefore sent after about 39 hops, the final results after about 78 hops, respectively. The difference between these two numbers is large enough to account for the two elevations of the histogram in the foreground of Figure 5.4. Note, that all curves are shifted to the right as compared to the mere hop count since it takes some time for the signaling step until the counting tokens can be started. In practice the current size of the overlay can be estimated to be able to choose an appropriate value for N_r as suggested in [2].

The arrival time of the results at the CP is also affected by the online/offline behavior of the individual peers. To study the influence of churn we consider 80000 peers with an exponentially distributed online and offline time, each with a mean of 60 minutes. This way, there are 40000 peers online on average, which makes it possible to compare the results to those obtained using 40000 peers without churn. Figure 5.5 shows the corresponding histograms.

As a result of churn in the system, the two elevations of the original histogram become noticeably blurred and the snapshot is slightly delayed. This is due to the inconsistencies in the successor and finger lists of the peer as well as the timeouts which occur during the forwarding of the counting tokens. In return the spike in the diagram and thus the required bandwidth at the CP becomes smaller.

It is easy to show, that the probability to lose a token is almost negligible [1]. Therefore, a more meaningful method to measure the influence of churn is to regard the number of timeouts which occur during a snapshot. Furthermore, the influence of churn on the stability of the overlay network can be studied looking at the

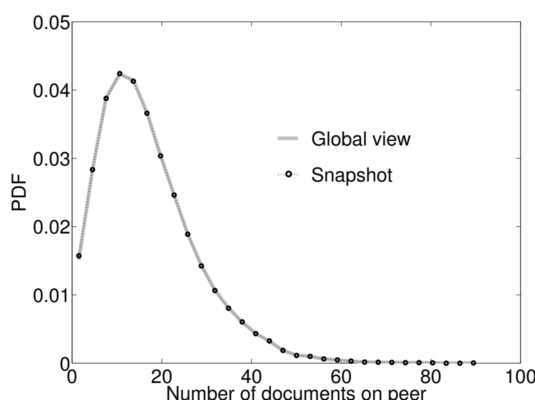

Fig. 5.7. *Results of a snapshot compared to the global view.*

frequency at which the predecessor pointer of a peer's successor does not match the peer itself. Figure 5.6 plots the relative frequency of timeouts and pointer failures against the mean online/offline time of a peer. The smaller the online/offline time of a peer, the more churn is in the system.

The results represent the mean of several simulation runs, whereas the error bars show the 95 percent confidence intervals. The relatively small percentage of both timeouts and failures is to some extent implementation specific. More interesting, however, is the exponential rise of the curves under higher churn rates. The shape of both curves is independent of the size of the overlay and only affected by the current churn rate. The curve can therefore be used to map the frequency of timeouts or failures measured in a running system to a specific churn rate.

Until now, we only regarded the traffic pattern at the central collecting point. From an operator's point of view, however, it is more important to know, whether the snapshot itself is meaningful. To validate the accuracy of the snapshot algorithm, we again simulated an overlay network with 80000 peers, each with a mean online/offline time of 60 minutes. Due to the properties of the hash function and the churn behavior in the system the number of documents on a single peer can be regarded as a random variable. The measurement we are interested in is the corresponding pdf in order to see whether the distribution of the documents among the peers is fair or not. The pdf was measured using our snapshot algorithm as explained in Section 3.3. The result of the snapshot is compared to the actual pdf obtained using the global view of our discrete event simulator (c.f. Figure 5.7). The two curves are almost indistinguishable from each other. The same is true for all the other statistics shown in Table 3.1, like the current size of the system or the average bandwidth used per time unit. That is, the snapshot provides the operator with a very accurate picture of the current state of its system. This nicely demonstrates that the results obtained by the snapshot can be used to better understand the performance of the running p2p system. The multiple possibilities to interpret the collected data are well beyond the scope of this paper.

6. Conclusion. One of the main reasons that telecommunication carriers are still hesitant to build p2p applications is the lack of control a provider has over the running system. In this paper we introduced an entirely distributed and scalable algorithm to monitor a Chord based p2p network at runtime. The load generated during the snapshot is evenly distributed among the peers of the overlay and the algorithm itself is easy to configure. It only takes one input parameter, which influences the trade-off between the duration of the snapshot and the bandwidth required at the central server which collects the results. In general it takes less than one minute to create a snapshot of a Chord ring consisting of 40000 peers. We performed a mathematical analysis of the basic mechanisms and provided a simulative study considering realistic user behavior.

The algorithm is resistant to instabilities in the overlay network (churn) and provides the operator with a very accurate picture of the current state of its system. It offers the possibility to monitor the entire overlay network or to concentrate on a specific part of the system. The latter is especially useful if a problem occurred in a specific part of the system and the operator wants to assure that his countermeasures have been successful.

REFERENCES

[1] A. Binzenhöfer, G. Kunzmann, and R. Henjes, *A Scalable Algorithm to Monitor Chord-based P2P Systems at Runtime*, Tech. Report 373, University of Würzburg, November 2005.
[2] A. Binzenhöfer, D. Staehle, and R. Henjes, *On the Fly Estimation of the Peer Population in a Chord-based P2P System*, in ITC19, Beijing, China, September 2005.
[3] A. Binzenhöfer and P. Tran-Gia, *Delay Analysis of a Chord-based Peer-to-Peer File-Sharing System*, in ATNAC 2004, Sydney, Australia, December 2004.
[4] U. B. C. S. Division, *The oceanstore project.* URL: http://oceanstore.cs.berkeley.edu/.
[5] G. Kunzmann, A. Binzenhöfer, and R. Henjes, *Analysis of the Stability of the Chord protocol under high Churn Rates*, in 6th Malaysia International Conference on Communications (MICC) icw International Conference on Networks (ICON), Kuala Lumpur, Malaysia, November 2005.
[6] G. Kunzmann, R. Nagel, and J. Eberspächer, *Increasing the reliability of structured p2p networks*, in 5th International Workshop on Design of Reliable Communication Networks, Island of Ischia, Italy, October 2005.
[7] J. Li, J. Stribling, T. M. Gil, R. Morris, and M. F. Kaashoek, *Comparing the performance of distributed hash tables under churn*, in Proceedings of the 3rd International Workshop on Peer-to-Peer Systems (IPTPS04), San Diego, CA, February 2004.
[8] D. S. Milojicic, V. Kalogeraki, R. Lukose, K. Nagaraja, J. Pruyne, B. Richard, S. Rollins, and Z. Xu., *P2P Computing*, Tech. Report HPL-2002-57, Hewlett Packard Lab, 2002.
[9] D. L. Oppenheimer, V. Vatkovskiy, H. Weatherspoon, J. Lee, D. A. Patterson, and J. Kubiatowicz, *Monitoring, analyzing, and controlling internet-scale systems with acme*, CoRR, cs.DC/0408035 (2004).
[10] V. N. Padmanabhan, S. Ramabhadran, and J. Padhye, *Netprofiler: Profiling wide-area networks using peer cooperation*, in Fourth International Workshop on Peer-to-Peer Systems (IPTPS), Ithaca, NY, USA, February 2005.
[11] A. Singh, P. Maniatis, T. Roscoe, and P. Druschel, *Using queries for distributed monitoring and forensics*, in 1st Eurosys, Leuven, Belgium, March 2006.
[12] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, *Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications*, in SIGCOMM 2001, San Diego, USA, August 2001.
[13] D. Stutzbach and R. Rejaie, *Capturing accurate snapshots of the gnutella network*, in INFOCOM 2005, Miami, USA, March 2005, pp. 2825–2830.
[14] S. Technologies, *Skype.* URL: http://www.skype.com.

Edited by: Pasqua D'Ambra, Daniela di Serafino, Mario Rosario Guarracino, Francesca Perla
Received: June 2007
Accepted: November 2008