Analysis of Link Lifetimes and Neighbor Selection in Switching DHTs

Zhongmei Yao, *Member, IEEE,* and Dmitri Loguinov, *Senior Member, IEEE*

Abstract—Several models of user churn, resilience, and link lifetime have recently appeared in the literature [13], [14], [36], [37]; however, these results do not directly apply to classical Distributed Hash Tables (DHTs) in which neighbor replacement occurs not only when current users die, but also when new users arrive into the system, and where replacement choices are often restricted to the successor of the failed zone in the DHT space. To understand neighbor churn in such networks, which we call *switching DHTs*, this paper proposes a simple, yet accurate, model for capturing link dynamics in structured P2P systems and obtains the distribution of link lifetimes for fairly generic DHTs. Similar to [9], our results show that deterministic networks (e.g., Chord [30], CAN [25]) unfortunately do not extract much benefit from heavy-tailed user lifetimes since link durations are dominated by small remaining lifetimes of newly arriving users that replace the more reliable existing neighbors. We also examine link lifetimes in randomized DHTs equipped with multiple choices for each link and show that selecting the best neighbor in these scenarios is rather complicated as it depends on the desired load-balancing, link resilience, and overhead. We offer insight into the various selection algorithms, their performance, and possibilities for improvement.

Index Terms—Distributed Hash Tables, Link Lifetimes, Neighbor Churn.

1 Introduction

Resilience of distributed peer-to-peer (P2P) networks under user churn has recently attracted significant attention and has become an important research area [2], [5], [12], [13], [15], [16], [17], [18], [27], [31], [36]. Traditional metrics of performance in this analysis have been the ability of the graph to stay connected during user departure [14], [18], [24], behavior of immediate neighbors during churn [12], data delivery ratio [32], evolution of out-degree [13] and in-degree [36], and churn rate in the set of participating nodes [9]. All metrics above depend on one fundamental parameter of churn – *link lifetime*, which is defined as the delay between formation of a link and its disconnection due to a sudden departure of the adjacent neighbor.

In many P2P networks, each user v creates k links to other peers when joining the system, where k may be a constant or a function of system size [19], and detects/repairs failed links in order to remain connected and perform P2P tasks (e.g., routing and key lookups) [25], [27], [28], [30]. Under fairly general conditions on user lifetimes [13], [36], link behavior is often modeled as an ON/OFF process in which each link is either ON at time t, which means that the corresponding user is currently alive, or OFF, which means that the user adjacent to the link has departed from the system and its failure is in the process of being detected and repaired. ON durations of links are commonly called *link lifetimes* and their OFF durations are called *repair delays*.

With this setup, it is not hard to see that link lifetimes play a key role in the study of resilience, performance, and reliability of P2P networks. For instance, longer average link lifetime means that users must repair failed links less frequently, which leads to smaller churn rates in the terminology of [9], and that queries are less likely to encounter dead neighbors during routing [12], which yields larger data delivery ratios [32] and higher lookup success rates.

If links do not switch to other users during each ON duration (i.e., keep connecting to the same neighbors until they fail), then link durations are simply *residual lifetimes* of original neighbors. We call this model *never-switching* and note that it applies to certain unstructured P2P networks [8] and some DHTs [28]. Link lifetimes for never-switching systems have been studied in fair detail under both age-independent [13], [36] and age-biased [32], [37] selection. However, many DHTs actively switch links to new neighbors before the current neighbor dies in order to balance the load and ensure DHT consistency. We call such systems *switching* and note that their link lifetimes require entirely different modeling techniques, which we present below.

1.1 Analysis of Existing DHTs

We start by introducing a stochastic process that keeps track of the changes in the identity of neighbors adjacent to the i-th link of a given user v as they become the current owner of this link under churn. We show that this process is a regular semi-Markov chain whose first hitting time to

- *Z. Yao is with the Department of Computer Science, the University of Dayton, 300 College Park, Dayton, OH 45469-2160. Email: zyao@udayton.edu.*
- *D. Loguinov is with the Department of Computer Science and Engineering, Texas A&M University, TAMU 3112, College Station, TX 77843. Email: dmitri@cse.tamu.edu.*

Manuscript received 16 Jan. 2009; revised 11 May 2010; accepted 16 Jan. 2011; published online 17 Mar. 2011.
Recommended for acceptance by J.C.S. Lui.
For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org, and reference IEEECS Log Number TPDS-2009-01-0026. Digital Object Identifier no. 10.1109/TPDS.2011.101.

the absorbing state (which corresponds to the failure of the last neighbor) is link lifetime R. Using this model, we find that the distribution of R is determined not only by lifetimes of attached users, but also by the zone size of the original neighbor holding the link. We thus additionally derive the distribution of zone size during the various phases of link ownership (i.e., for the initial neighbor and those obtained after each stabilization).

We next obtain the Laplace transform of the distribution of R and derive its expected value $E[R]$ for general user lifetimes L, including heavy-tailed cases. We use this result to show that under heavy-tailed peer lifetimes (e.g., Pareto) observed in many real P2P networks [4], [29], [33], link lifetime R is stochastically *smaller* than the residual lifetime Z of the initial neighbor holding the link. Consistent with simulations in [10], our results also show that $E[R]$ is very close to $E[L]$, which is in stark contrast to the results of [13], where $E[R]$ was several times larger than $E[L]$ depending on Pareto shape α of the lifetime distribution.

This phenomenon occurs because older (i.e., more reliable) neighbors in DHTs are replaced with new arrivals that exhibit much shorter remaining lifetimes. As a result, classical DHTs unfortunately do not extract any benefits from heavy-tailed user lifetimes and suffer much higher link churn rates than the corresponding unstructured systems [13]. A similar conclusion was obtained in [9] for query failure rates in Chord.

1.2 Improvements

One method of overcoming the problem identified above is to utilize randomized DHTs (e.g., randomized Chord [11], randomized hypercube [21], and Symphony [20]) in which the i-th finger pointer of a given user v is randomly selected from some set S_i of possible locations in the DHT space. By trying multiple options in S_i and linking to the user with the best characteristics, the hope is to improve link lifetime and reduce the impact of churn on system performance. While freedom of neighbor choice allows randomized DHTs to operate under never-switching, where link lifetime is understood pretty well [13], [32], [36], [37], we next explore their performance under switching.

The first obvious randomized technique, which we call *switching max-age* (SMA), selects $m \geq 1$ points in S_i uniformly randomly, places the finger into such generated point that its successor has the largest current age, and maintains a neighboring connection to whoever is the current successor (i.e., owner) of that finger. While quite effective in never-switching scenarios, this strategy has minimal impact in switching DHTs since link lifetime is determined by the remaining session length of not the *first*, but the *last* neighbor holding the link. To overcome this limitation, we examine several alternative randomized strategies that stem from our model of link lifetime R and discuss the various performance tradeoffs that arise in each case.

We finish the paper by examining an orthogonal approach that restricts DHT users to some minimal age before any links or objects are assigned to them and discussing how the developed models apply to these situations. Specifically, we study the *delayed-join* strategy of widely deployed unstructured P2P systems, in which only special nodes with enough uptime (e.g., ultra-peers in Gnutella) are allowed to route queries and hold keys. The remaining users (called *leaves*) can only initiate and answer queries to/from the system. As nothing prevents a similar approach from being deployed in a DHT, we show that for Pareto lifetimes with $E[L] = 0.5$ hours and 21% of the graph delegated to support DHT routing, delaying each join by just 6 minutes increases link lifetime by a factor of 4.4, which is quite significant in practice. More examples are discussed later in this paper.

2 General DHT Model

We start by formulating assumptions on the churn model, DHT space, and link switching in DHTs. Due to limited space in the printed edition, discussion of related work can be found below in Section 6, omitted simulations in Section 7, and all proofs in Section 8.

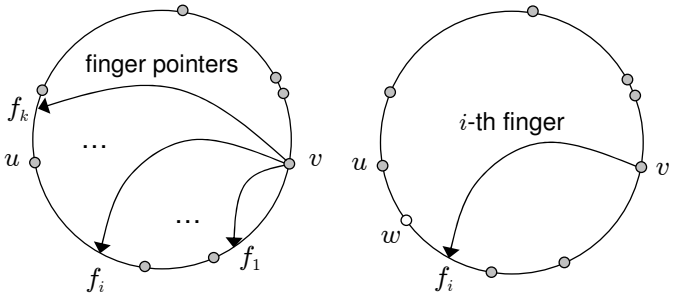
2.1 Churn Model

For user churn, we adopt the recently introduced [36] framework of n alternating renewal processes representing periodic online/offline behavior of users observed in real P2P systems [9], [33]. In this model, each user i is viewed as alternating between online and offline states, where the duration of each state is random and has some user-specific distribution.

While the total number of users n is fixed in this model, the number of *currently alive* peers N_t at time t is a random process that fluctuates over time. Once stationarity is reached, we usually replace N_t with its limiting version $N = \lim_{t\to\infty} N_t$. As a consequence of this churn model [36, Theorem 5], user arrivals into the system follow a Poisson process with a constant rate $\lambda = E[N]/E[L]$, where $E[N]$ is the average number of users in the steady state and $E[L]$ is the mean user lifetime.

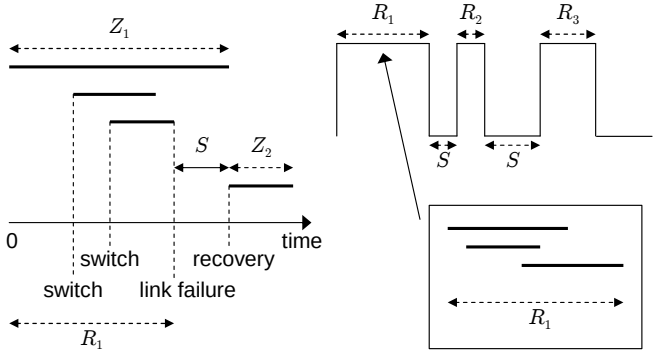
2.2 DHT Classification

Many traditional DHTs, including those with d-dimensional number spaces, can be mapped to a 1D ring by treating node IDs as some large integers. Depending on the DHT and the mapping applied, each node may hold a single contiguous or several non-contiguous zones on the ring. Due to limited space, we explicitly deal only with Chord-like systems; however, we believe that our neighbor-dynamics model introduced in the next section is general enough to apply to a variety of other underlying graphs. Furthermore, while numerical results for link lifetime in non-Chord DHTs may somewhat differ from those shown below, the main qualitative conclusions of the paper (i.e., switching reduces link



(a) k fingers of user v (b) arriving node w

Fig. 1. User v's fingers and neighbors in the DHT.



(a) one recovery cycle (b) ON/OFF link behavior

Fig. 3. Switch and recovery of v's link i.

lifetime) should hold almost universally as long as users exhibit heavy-tailed session durations.

Assume that the network maps keys and users using a uniform hashing function into the same identifier (ID) space, which is a continuous ring in the interval $[0, 1)$ [23]. Each user v is responsible for a fraction of the DHT space from its predecessor to v's own hash, which we call the user's *zone*. As the network evolves, one of v's functions is to store objects that map to its zone and answer queries related to them.

To facilitate routing, each peer selects k finger pointers $f_1, f_2, \ldots, f_k$ in the DHT space and creates transport-layer (usually TCP) connections to users whose zones hold the corresponding finger. Define $owner(x)$ to be the nearest live peer in the clockwise direction from x. Then, v's out-link i is connected to user $owner(f_i)$. This is illustrated in Fig. 1(a), where live users are marked with circles and v's fingers are shown as arrows. Observe in the figure that currently $u = owner(f_i)$; however, this may change as the system experiences churn and additional users arrive in the interval $[f_i, u]$ as shown in Fig. 1(b).

One strategy [25], [30] for dealing with zone churn, which we call *switching* and study throughout this paper, is to maintain invariance of $neighbor_i = owner(f_i)$ at all times. As peers join, they split existing zones and inherit not only the objects, but also the in-links, that now belong to their zone. This provides new arrivals with their share of in-degree and routing load, as well as guarantees certain finger-consistency properties and system-wide routing bounds. As shown in Fig. 2, the finger rules of switching systems can be further classified as either *rigid*, which means f_i is a deterministic function of v's ID, or *flexible*, which means f_i is selected from a certain (often randomized) set of options.

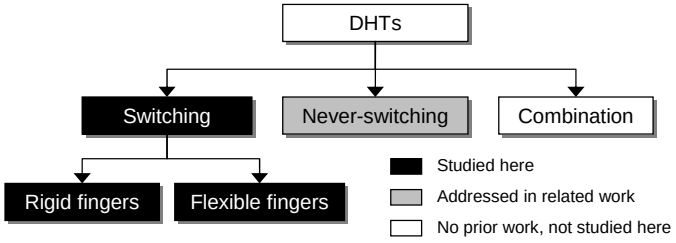


Fig. 2. DHT taxonomy and link lifetime analysis.

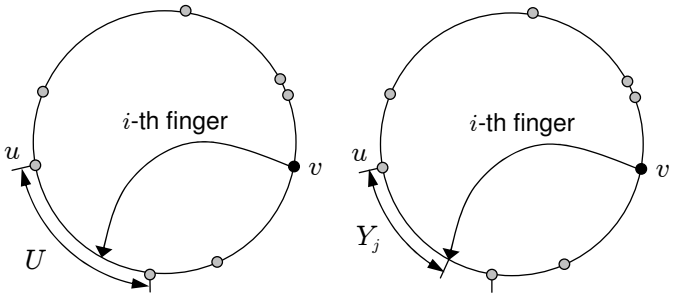
The second strategy [28], [38] for handling new arrivals, which we call *never-switching*, is commonly found in generalized hypercubes whose function $owner(f_i)$ treats *all* users within some fixed ϵ_i-proximity of f_i as equally suitable for neighboring. This allows v to retain the initial neighbor since property $||neighbor_i - f_i|| \leq \epsilon_i$ cannot be altered by new arrivals. Link lifetimes in never-switching DHTs are covered by prior analysis of unstructured P2P systems [13]. The last strategy [22], which we call *combination*, performs periodic switching based on various neighbor-quality metrics (e.g., ping delay, uptime, geographic proximity). However, exact modeling of its link lifetimes is far too involved to be included here.

2.3 Switching Neighbor Dynamics

We next formalize the link process in switching DHTs. Our discussion focuses on the behavior of one particular link i (other links are similar) and the lifetimes of neighbors adjacent to it during v's online session. As user v continues to stay in the system, the identity of its neighbors (i.e., finger owners/successors) may change over time as users join and leave the system. There are two types of changes in neighbor tables – graceful handoffs of existing zones to arriving users and node departures without explicit notification of v [30].

The former type, which we call a *switch*, occurs when a new arrival takes ownership of a link by becoming the new successor of the corresponding neighbor pointer. This is shown in Fig. 1(b) where a new arrival w splits the zone of an existing neighbor u and becomes v's new neighbor along link i since $w = owner(f_i)$. The latter type of neighbor change, which we call a *recovery*, happens when an existing neighbor dies and the successor of the failed neighbor takes over that zone to become the new neighbor of v.

We next define several additional metrics to facilitate explanation in later parts of the paper. Notice that one cycle in the life of a particular neighbor pointer is composed of several switches and one recovery as shown in Fig. 3(a). In the figure, thick horizonal lines represent online presence of peers that own v's neighbor pointer in the DHT space.



(a) zone size U of user u (b) remaining zone size Y_j

Fig. 4. Zone size U and remaining zone size Y_j of user u.

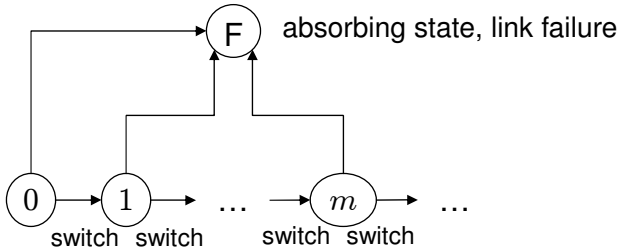


Fig. 5. State diagram for the process $\{A_\delta^j, \delta \geq 0\}$ of neighbor changes.

The topmost line represents the original neighbor (with residual lifetime Z_1) acquired by v during join. As peers split the zone of the current neighbor, the link switches to two additional users. Switch is complete after a new user performs all join tasks [30]. Once the last user dies at time R_1, the link is considered dead and a replacement process is initiated.[1] Recovery is finished after S time units when another node takes over the zone of the dead peer and is selected as v's new neighbor.

The second recovery cycle behaves identical to the first one (except the zone size of the initial neighbor is larger) and leads to link failure after R_2 time units. This ON/OFF nature of the link process is shown in Fig. 3(b) where we assume that all repair delays S are i.i.d. random variables, but distributions of link lifetimes $R_1, R_2, \ldots$ may depend on the cycle number (in fact they do in certain cases studied below).

The final note is that it is important to distinguish the residual lifetime of the first neighbor from that of a link. While in never-switching systems the former metric (e.g., variables $Z_1, Z_2, \ldots$) determines how long a link stays alive, this is no longer the case in switching networks. Instead, the latter metric formalized as $R_1, R_2, \ldots$ determines query performance and a user's ability to tolerate churn. Our next step is to understand the behavior of these random variables under general lifetime distributions.

3 Link Lifetime Model

In this section, we construct a semi-Markov model for the distribution of lifetimes $R_1, R_2, \ldots$ of a given link in a user's routing table.

3.1 Preliminaries

Recall that arriving users split zones of existing nodes based on a uniformly random hash function. Denote by U the random zone size of existing users in a stationary system as shown in Fig. 4(a). Further assume that during join or the current recovery step that starts cycle j, successor u takes over pointer i as shown in Fig. 4(b). Then, define Y_j to be the *remaining zone size* between this pointer and the hash index of u. Intuitively, if the remaining zone Y_j is large, then it is likely that a new arrival will soon split the zone and the ownership of the link will be transferred to another peer. Therefore, link lifetimes are determined not by the distribution of U, but rather by that of Y_j. We derive both metrics later in the paper and next show how they can be used to obtain $R_1, R_2, \ldots$

For simplicity of notation, define *conditional link lifetime* $R(y)$ as the duration of the link conditioned on the fact that the remaining zone size Y_j is $y > 0$. Then, observe that the CDF (cumulative distribution function) of link lifetimes R_j can be written as:

$$P(R_j < x) = \int_0^\infty P(R(y) < x) f_{Y_j}(y) dy, \tag{1}$$

where $f_{Y_j}(y)$ is the PDF (probability density function) of remaining zone size Y_j (note that the distribution of Y_j depends on cycle number j). Similarly, we can obtain the expectation of R_j as:

$$E[R_j] = \int_0^\infty E[R(y)] f_{Y_j}(y) dy. \tag{2}$$

Thus, the task of deriving link lifetime R_j is reduced to analyzing the properties of conditional link lifetime $R(y)$ and the distribution of remaining zone size Y_j. In the rest of this section, we construct a semi-Markov process for each $R(y)$ and leave the derivation of the distribution of Y_j to a later section.

3.2 Neighbor Dynamics

For each zone size y, let variable A_δ^y count the number of switches (i.e., replacements by new users) that have occurred along the link in the time interval $[0, \delta]$, where time 0 denotes the instance when user v finds the first neighbor at the beginning of the current cycle. Denote by $A_\delta^y = F$ a special absorbing state into which A_δ^y arrives if the current neighbor attached to the link is in the failed state at time δ.

Then, it is easy to see that $\{A_\delta^y; \delta \geq 0\}$ is a continuous-time stochastic process with state space $\{F, 0, 1, 2, \ldots\}$ whose state transitions are shown in Fig. 5. As depicted in this figure, for each state $i \geq 0$, the process can jump into either state $i + 1$, which means that a given zone is further split by a new arrival (i.e., the number of switches

1. Specifics of detecting failure are not essential to our results as repair delay is not studied in this paper.

increases by 1), or state F, which represents link failure. The initial state of the process at time 0 is always 0.

Using notation $\{A_\delta^y\}$, variable $R(y)$ can be described as the first-hitting time of process $\{A_\delta^y\}$ onto state F given that $A_0^y = 0$:

$$R(y) = \inf\{\delta > 0 : A_\delta^y = F | A_0^y = 0, Y_j = y\}. \quad (3)$$

The next theorem shows that $\{A_\delta^y; \delta \geq 0\}$ is a semi-Markov chain that describes the process of new users entering a given zone of initial length y and repeatedly splitting it.

Theorem 1. *Process $\{A_\delta^y, \delta \geq 0\}$ for a given remaining zone size $Y_j = y$ is a regular semi-Markov chain. The sojourn time τ_i in state i follows the following general distribution:*

$$P(\tau_i > x) = \begin{cases} P(W_0 > x)P(Z_j > x) & i = 0 \\ P(W_i > x)P(L > x) & i \geq 1 \end{cases}, \quad (4)$$

where Z_j is the residual lifetime of the first neighbor that starts the j-th cycle, L is user lifetime with CDF $F(x)$, W_i is an exponential random variable with rate λ_i:

$$\lambda_i = E[N]y/(E[L]2^i), \quad i \geq 0, \quad (5)$$

and $E[N]$ is the mean system size. Furthermore, transition probability $p_{i,i+1}$ from state i to $i+1$ is given by:

$$p_{i,i+1} = \begin{cases} P(W_0 < Z_j) & i = 0 \\ P(W_i < L) & i \geq 1 \end{cases}, \quad (6)$$

and the probability $p_{i,F}$ to absorb from state i is equal to $1 - p_{i,i+1}$.

This theorem shows in (5) that as the number of switches within a zone (i.e., variable i) increases, arrival rate λ_i of new users into the zone decreases exponentially fast (or alternatively, the mean waiting time $E[W_i]$ until the next arrival increases at the same rate). As $i \to \infty$, the likelihood of a new arrival into the zone diminishes and the delay in state i becomes simply the lifetime of the last user holding the edge. For small i, however, analysis is much more complex as shown in the next subsection.

3.3 Conditional Link Lifetimes

Next, we study the distribution and expectation of conditional link lifetime $R(y)$. To understand our next theorem, several definitions are necessary. First, denote the CDF of sojourn time τ_i in state i by $G_i(t) = P(\tau_i < t)$.

Second, observing from (4) that τ_i of chain $\{A_\delta^y\}$ is independent of the next state, define a semi-Markov kernel matrix $Q(t) = [q_{ik}(t)]$ using [6]:

$$q_{ik}(t) = p_{ik}G_i(t), \quad i, k \in \{F, 0, 1, \ldots\}, \quad (7)$$

where p_{ik} is the transition probability from state i to state k given in (6). The Laplace (Stieltjes) transform of $q_{ik}(t)$ is then simply:

$$\hat{q}_{ik}(s) = \int_0^\infty e^{-st} dq_{ik}(t) = p_{ik} \int_0^\infty e^{-st} dG_i(t). \quad (8)$$

Finally, define the Laplace transform of the first hitting time $R(y)$ from state 0 to F as $\hat{R}(s, y) = E[e^{-sR(y)}]$.

Although it is known that the Laplace transform of the first-hitting time of a semi-Markov chain can be computed using spectral properties of kernel $Q(t)$ [3], this approach hides the effect of system parameters on the resulting distribution. Due to the simplicity of state transitions of chain $\{A_\delta^y\}$, we next derive $\hat{R}(s, y)$ without involving matrix operations on $Q(t)$.

Theorem 2. *The Laplace transform $\hat{R}(s, y)$ of conditional link lifetime $R(y)$ is given by:*

$$\hat{R}(s, y) = \hat{q}_{0F}(s) + \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k-1} \hat{q}_{i,i+1}(s) \right) \hat{q}_{kF}(s), \quad (9)$$

where $\hat{q}_{ik}(s)$ are shown in (8).

With $\hat{R}(s, y)$ in hand, we can apply the inverse Laplace transform to retrieve the distribution of $R(y)$ and take the derivatives of $\hat{R}(s, y)$ to get its moments. Next, we use a simpler approach to obtain the mean $E[R(y)]$.

Theorem 3. *The expected conditional link lifetime is:*

$$E[R(y)] = E[\tau_0] + \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k-1} p_{i,i+1} \right) E[\tau_k], \quad (10)$$

where $E[\tau_k]$ is the expected sojourn time in state k shown in (4) *and $p_{i,i+1}$ are state transition probabilities in* (6).

Theorems 1–3 demonstrate that variable $R(y)$ is fully determined by user lifetimes L and residual neighbor lifetimes Z_j. Our remaining steps are to analyze the properties of Z_j and derive the distribution of remaining zone sizes Y_j for both deterministic and randomized DHTs.

4 Rigid Fingers

In DHTs with rigid (often called *deterministic*) finger rules, each neighbor pointer of user v is generated based on a fixed distance between the pointer and the user. We start this section by deriving a model for $R(y)$ under two types of user lifetimes and then analyze the distribution of residual zone size Y_j.

4.1 Residual Lifetimes of Neighbors

Under the user churn model assumed in this paper, the distribution of neighbor residual lifetime under *age-independent* selection converges to the following equilibrium CDF as system age $t \to \infty$ [36, Theorem 3]:

$$F_e(x) = \frac{1}{E[L]} \int_0^x (1 - F(u)) du, \quad (11)$$

where $F(x)$ is the user lifetime distribution. Since recovery in our DHT model is not biased with respect to user age, (11) is also the CDF of residual lifetime for users that are found during recovery, which we formally state in the next lemma.

Lemma 1. *For all $j \geq 1$, the CDF of residual lifetime Z_j of the initial neighbor that starts the j-th cycle converges to* (11) *as system age approaches infinity.*

Given Lemma 1, the mean residual lifetime $E[Z_j]$ can be expressed directly using the properties of L as [35]:

$$E[Z_j] = \frac{E[L^2]}{2E[L]}. \tag{12}$$

It is important to emphasize that Lemma 1 holds when switching occurs in DHTs in response to Poisson user arrivals into the system and may not hold otherwise. When a neighbor pointer switches to a new user, it loses track of which peer on the ring will be the neighbor that will start the next cycle in the link's ON/OFF process. Hence, neighbor selection during link recovery is essentially uniformly random among the existing neighbors (due to random hash indexes) and independent of the selected neighbor's age.

4.2 Exponential Lifetimes

If user lifetimes L are exponential with rate $\mu = 1/E[L]$, it is easy to obtain from Lemma 1 that Z_j of the initial neighbor, for all cycles $j \geq 1$, is exponential with the same rate μ. Due to the memoryless property of exponential distributions, the remainder of Z_j obtained at *any* random instant (i.e., when a switch occurs) is still exponential with rate μ. Therefore, it makes no difference whether the current neighbor is replaced by a new arrival or not. Interestingly, this result is valid not only for Poisson arrivals, but also for any arrival process independent of user lifetimes that results in non-explosive chain $\{A_\delta^y\}$.

Theorem 4. *For user lifetimes L with CDF $1 - e^{-\mu x}$, link lifetime R_j is independent of remaining zone size Y_j and has the same distribution as L:*

$$P(R_j < x) = 1 - e^{-\mu x}, \quad \textit{for all } j \geq 1, \tag{13}$$

where $\mu = 1/E[L]$.

Theorem 4 indicates that switching has no impact on link lifetimes in any DHT with exponential user lifetimes, which makes analysis of system performance in such systems very simple. However, we should note that this result does not hold for any non-exponential lifetime distribution. As recent measurements of P2P networks show that user lifetimes are often heavy-tailed [4], [33], we next use the Pareto distribution $P(L < x) = 1-(1+x/\beta)^{-\alpha}$ with shape parameter $\alpha > 1$ and scale parameter $\beta > 0$ to estimate the performance of real DHTs under churn.

4.3 Pareto Lifetimes

For Pareto L, it is clear from Lemma 1 that the residual lifetime Z_j of initial neighbors follows the CDF $P(Z_j < x) = 1 - (1 + x/\beta)^{-(\alpha-1)}$ for all $j \geq 1$, which shows that Z_j are also Pareto-distributed but more heavy-tailed. Next, we apply Theorem 2 to obtain the Laplace transform $\hat{R}(y, s)$ and Theorem 3 to obtain the mean of $R(y)$.

Theorem 5. *For Pareto lifetimes L, the mean conditional link lifetime $E[R(y)]$ is given by* (10) *with*

$$E[\tau_i] = \beta e^{\lambda_i \beta} E_{\alpha_i}(\lambda_i \beta), \quad p_{i,i+1} = \lambda_i E[\tau_i] \tag{14}$$

where arrival rate λ_i is given in (5), *$E_k(x) = \int_1^\infty e^{-xu}\, u^{-k} du$ is the generalized exponential integral, $\alpha_i = \alpha - 1$ for $i = 0$, and $\alpha_i = \alpha$ for $i \geq 1$. Furthermore, the Laplace transform $\hat{R}(y, s)$ is given by* (9) *with*

$$\hat{q}_{i,i+1}(s) = \lambda_i E[\tau_i] A, \quad \hat{q}_{iF}(s) = (1 - \lambda_i E[\tau_i]) A, \tag{15}$$

where $A = 1 + (1 - \lambda_i - s)\beta e^{(\lambda_i + s)\beta} E_{\alpha_i}((\lambda_i + s)\beta)$, and $E[\tau_i]$ is shown in (14).

We next derive the distribution of zone sizes in deterministic DHTs in order to obtain a computable model for R_j.

4.4 Zone Sizes

In order to determine the distribution of zone sizes U and Y_j in Fig. 4, we must decide on the zone splitting method. The derivations below only cover the random-split [34] mechanism (i.e., zones are split at hash indexes of arriving users) that is used in Chord [30] and only consider one-dimensional DHTs. A similar derivation can be carried out for the center-split [19], [25] strategy (i.e., zones are always split in the center) and multi-dimensional DHTs, but this analysis is much more tedious and is not shown here.

Since all arriving users are placed in the interval $[0, 1)$, the average zone size is approximately $1/E[N]$, where N is the random system size in the steady-state.[2] The next result states that in equilibrium DHTs, zone sizes no larger than $1/\sqrt{E[N]}$ are distributed approximately exponentially. Since most zone sizes do not deviate from the mean very far, this result directly applies to random variable U defined earlier.

Lemma 2. *As the mean system size tends to infinity, the distribution of small zones in the DHT becomes approximately exponential:*

$$\lim_{E[N]\to\infty} \frac{P(U > x)}{e^{-E[N]x}} = 1 \tag{16}$$

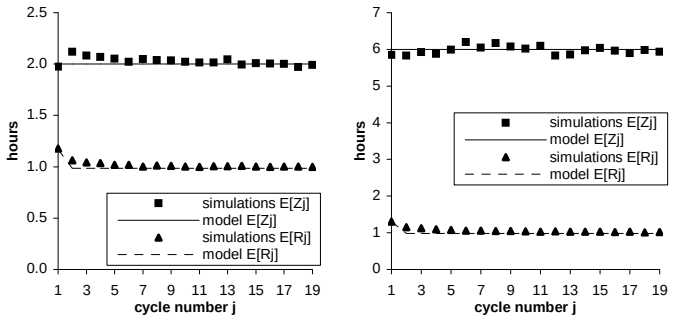
for all x such that $x^2 E[N] \to 0$.

Our next task is to obtain the distribution of remaining zone size Y_j in each cycle $j \geq 1$.

Lemma 3. *For a given zone size y, assume that $y^2 E[N] \to 0$ as $E[N] \to \infty$. Then, the PDF $f_{Y_j}(y)$ of remaining zone size Y_j is asymptotically:*

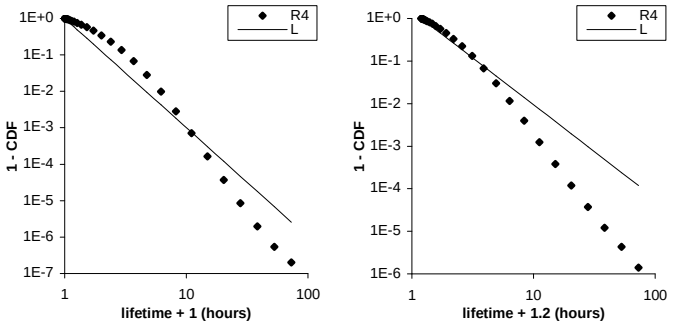
$$\begin{cases} \lim\limits_{E[N]\to\infty} \dfrac{f_{Y_1}(y)}{E[N]e^{-E[N]y}} = 1 & j = 1 \\ \lim\limits_{E[N]\to\infty} \dfrac{f_{Y_j}(y)}{E[N]^2 y e^{-E[N]y}} = 1 & j \geq 2 \end{cases}, \tag{17}$$

2. Approximation $E[1/N] = 1/E[N]$ is asymptotically accurate as system size tends to infinity for the ON/OFF churn model of [36]. This follows from the fact that $N/E[N]$ converges to 1 in probability.



(a) $\alpha = 3$ (b) $\alpha = 2.2$

Fig. 6. Comparison of $E[R_j]$ to $E[Z_j]$ in a deterministic DHT with mean size $E[N] = 2,500$ users, Pareto lifetimes with mean $E[L] = 1$ hour, and $\beta = E[L](\alpha - 1)$.



(a) $\alpha = 3$ (b) $\alpha = 2.2$

Fig. 7. Link lifetimes R_4 are less heavy-tailed than Pareto user lifetimes L in a deterministic DHT with mean size $E[N] = 2,500$ peers, $E[L] = 1$ hour, and $\beta = (\alpha - 1)E[L]$.

where $E[N]$ is the mean system size in equilibrium.

Lemma 3 shows that the distribution of Y_1 is exponential and that of Y_j for $j \geq 2$ is Erlang-2 (i.e., the distribution of the sum of two exponentials).

4.5 Putting the Pieces Together

The final step is to apply (1) and (2) to uncondition the distribution of link lifetime R_j and its mean $E[R_j]$ using the distribution of initial zone size Y_j given in (17). To this end, substituting $E[R(y)]$ shown in Theorem 5 and the PDF of Y_j in (17) into (2) leads to the final result on the mean link lifetime $E[R_j]$. Similarly, to get the distribution of R_j, we first retrieve the distribution of $R(y)$ from $\hat{R}(s, y)$ in Theorem 5 by applying an existing inverse Laplace transform software package [1]. Then substituting the distribution of $R(y)$ and (17) into (1) leads to the final model of the distribution of R_j.

Fig. 6 shows simulations results and the model of the mean link lifetime $E[R_j]$ and the average residual lifetime $E[Z_j]$ of the initial neighbor that starts the j-th cycle. The model of $E[Z_j]$ is obtained using (12) and the general solution to $E[R_j]$ is given in (2). As shown in the figure, both models match simulation results very well. Furthermore, as α becomes smaller, the difference between $E[R_j]$ and $E[Z_j]$ increases as expected.[3] The above results also show that the process of switching to new users can significantly reduce the lifetime of a link and that deterministic DHT systems with Pareto L can exhibit $E[R_j]$ very close to $E[L]$. This is in contrast to unstructured P2P systems where $E[R_j]$ can be $11 - 16$ times higher than $E[L]$ depending on shape parameter α [4], [33].

Further observe from the model and Fig. 6 that link lifetimes are completely characterized by two random variables R_1 and R_2 since R_j for $j \geq 3$ has the same distribution as R_2. This arises from the fact that zone size Y_1 is different from Y_2, while Y_j for $j \geq 3$ are all distributed as Y_2. Since Y_1 is stochastically smaller than Y_2 (see Lemma 3), it follows that R_1 is stochastically larger than R_2. Furthermore, from the analysis of the Markov chain in previous sections, it becomes clear that selecting neighbors with *smaller* initial residual zone sizes leads to larger link lifetimes since such neighbors are less likely to be replaced by newly arriving users and the link's $E[R_j]$ will be closer to $E[Z_j]$.

The most intriguing result shown in Fig. 6 is that $E[R_j]$ for all $j \geq 2$ is very close to the mean user lifetime $E[L]$ under different values of α (e.g., $E[R_4] = 0.986$ hours for $\alpha = 3$ and 1.096 for $\alpha = 2.2$). However, from the model of the tail distribution of link lifetime R_4 shown in Fig. 7, observe that the distribution of R_j for $j \geq 2$ is actually different from that of lifetime L and is *less* heavy-tailed than the original distribution. A similar result holds for other values of α and other distributions, which we do not show for brevity.

Given this disappointing performance of classical (i.e., rigid) DHTs, a natural question arises as to whether flexible (often called *randomized*) fingers can improve link lifetimes. In such systems, one obvious choice is never-switching, which retains the initial neighbor along each link i until it dies. Such algorithms have been covered in related work [13], [32], [37] and are not addressed here. Instead, in Sections 9 and 10 below we study link dynamics of switching DHTs under flexible finger rules and dissect the impact of delayed joins (i.e., age-based decisions to promote peers and/or increase their responsibility) on link lifetime.

5 Conclusion

This paper formalized the notion of "link lifetimes" in certain types of DHTs where link pointers switch to new neighbors in response to arriving peers. We introduced a semi-Markov process to model random replacement of neighbors along a given link and showed that lifetimes of deterministic links are much worse than those in unstructured P2P networks with heavy-tailed user lifetimes. For randomized DHTs, our results showed that finger placement based on both node age and zone size was the most general approach. We also demonstrated that if none of the approaches above were viable, simply

3. Recall that smaller α leads to stochastically larger Z_j and thus increases reliability of never-switching systems [13].

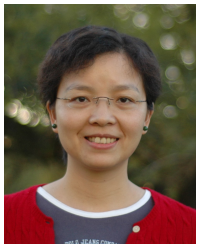
delaying assignment of responsibility to arriving users by just several minutes could yield significant improvements.

ACKNOWLEDGMENT

The authors are grateful to P. Brighten Godfrey for bringing this problem to our attention and offering insightful discussion. They are also indebted to the anonymous *IEEE TPDS* reviewers for helping greatly improve this paper. An earlier version of this paper appeared in IEEE INFOCOM 2008. This work is supported by US National Science Foundation (NSF) grant CNS-0720571.

REFERENCES

[1] J. Abate and P. P. Valkó, "Multi-Precision Laplace Transform Inversion," *Int. J. Numer. Meth. Engng*, vol. 60, pp. 979–993, 2004.
[2] R. Bhagwan, S. Savage, and G. M. Voelker, "Understanding Availability," in *Proc. IPTPS*, Feb. 2003, pp. 256–267.
[3] J. T. Bradley, N. J. Dingle, P. G. Harrison, and W. J. Knottenbelt, "Distributed Computation of Passage Time Quantiles and Transient State Distributions in Large Semi-Markov Models," in *Proc. IPDPS*, Apr. 2003.
[4] F. E. Bustamante and Y. Qiao, "Friendships that Last: Peer Lifespan and its Role in P2P Protocols," in *Proc. Intl. Workshop on Web Content Caching and Distribution*, Sep. 2003.
[5] M. Castro, M. Costa, and A. Rowstron, "Performance and Dependability of Structured Peer-to-Peer Overlays," in *Proc. DSN*, Jun. 2004.
[6] E. Çinlar, *Introduction to Stochastic Processes*. Englewood Cliffs, NJ: Prentice Hall, 1997.
[7] L. Devroye, "Law of the Iterated Logarithm for Order Statistics of Uniform Spacings," *Annals of Probability*, vol. 9, pp. 860–867, 1981.
[8] Gnutella. [Online]. Available: http://www.gnutella.com/.
[9] P. B. Godfrey, S. Shenker, and I. Stoica, "Minimizing Churn in Distributed Systems," in *Proc. ACM SIGCOMM*, Sep. 2006.
[10] P. B. Godfrey, Personal Communication, 2006.
[11] K. Gummadi, R. Gummadi, S. Gribble, S. Ratnasamy, S. Shenker, and I. Stoica, "The Impact of DHT Routing Geometry on Resilience and Proximity," in *Proc. ACM SIGCOMM*, Aug. 2003, pp. 381–394.
[12] S. Krishnamurthy, S. El-Ansary, E. Aurell, and S. Haridi, "A Statistical Theory of Chord under Churn," in *Proc. IPTPS*, Feb. 2005, pp. 93–103.
[13] D. Leonard, V. Rai, and D. Loguinov, "On Lifetime-Based Node Failure and Stochastic Resilience of Decentralized Peer-to-Peer Networks," in *Proc. ACM SIGMETRICS*, Jun. 2005, pp. 26–37.
[14] D. Leonard, Z. Yao, X. Wang, and D. Loguinov, "On Static and Dynamic Partitioning Behavior of Large-Scale Networks," in *Proc. IEEE ICNP*, Nov. 2005, pp. 345–357.
[15] J. Li, J. Stribling, T. M. Gil, R. Morris, and M. F. Kaashoek, "Comparing the Performance of Distributed Hash Tables under Churn," in *Proc. IPTPS*, Feb. 2004, pp. 87–99.
[16] J. Li, J. Stribling, R. Morris, and M. F. Kaashoek, "Bandwidth-Efficient Management of DHT Routing Tables," in *Proc. USENIX NSDI*, May 2005, pp. 1–11.
[17] J. Li, J. Stribling, R. Morris, M. F. Kaashoek, and T. M. Gil, "A Performance vs. Cost Framework for Evaluating DHT Design Tradeoffs under Churn," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 225–236.
[18] D. Liben-Nowell, H. Balakrishnan, and D. Karger, "Analysis of the Evolution of the Peer-to-Peer Systems," in *Proc. ACM PODC*, Jul. 2002, pp. 233–242.
[19] D. Loguinov, A. Kumar, V. Rai, and S. Ganesh, "Graph-Theoretic Analysis of Structured Peer-to-Peer Systems: Routing Distances and Fault Resilience," in *Proc. ACM SIGCOMM*, Aug. 2003, pp. 395–406.
[20] G. Manku, M. Bawa, and P. Raghavan, "Symphony: Distributed Hashing in a Small World," in *Proc. USITS*, Mar. 2003, pp. 127–140.
[21] G. S. Manku, M. Naor, and U. Weider, "Know thy Neighbor's Neighbor: the Power of Lookahead in Randomized P2P Networks," in *Proc. ACM STOC*, Jun. 2004, pp. 54–63.
[22] P. Maymounkov and D. Mazieres, "Kademlia: A Peer-to-Peer Information System Based on the XOR Metric," in *Proc. IPTPS*, Mar. 2002, pp. 53–65.
[23] M. Naor and U. Wieder, "Novel Architectures for P2P Applications: The Continuous-Discrete Approach," in *Proc. ACM SPAA*, Jun. 2003, pp. 50–59.
[24] G. Pandurangan, P. Raghavan, and E. Upfal, "Building Low-Diameter Peer-to-Peer Networks," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 6, pp. 995–1002, Aug. 2003.
[25] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A Scalable Content-Addressable Network," in *Proc. ACM SIGCOMM*, Aug. 2001, pp. 161–172.
[26] S. I. Resnick, *Adventures in Stochastic Processes*. Boston, MA: Birkhäuser, 2002.
[27] S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz, "Handling Churn in a DHT," in *Proc. USENIX Ann. Tech. Conf.*, Jun. 2004, pp. 127–140.
[28] A. Rowstron and P. Druschel, "Pastry: Scalable, Decentralized Object Location and Routing for Large-Scale Peer-to-Peer Systems," in *Proc. IFIP/ACM International Conference on Distributed Systems Platforms (Middleware)*, Nov. 2001, pp. 329–350.
[29] S. Saroiu, P. K. Gummadi, and S. D. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems," in *Proc. SPIE/ACM Multimedia Computing and Networking*, vol. 4673, Jan. 2002, pp. 156–170.
[30] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications," in *Proc. ACM SIGCOMM*, Aug. 2001, pp. 149–160.
[31] D. Stutzbach and R. Rejaie, "Understanding Churn in Peer-to-Peer Networks," in *Proc. ACM IMC*, Oct. 2006, pp. 189–202.
[32] G. Tan and S. Jarvis, "Stochastic Analysis and Improvement of the Reliability of DHT-based Multicast," in *Proc. IEEE INFOCOM*, May 2007, pp. 2198–2206.
[33] X. Wang, Z. Yao, and D. Loguinov, "Residual-Based Measurement of Peer and Link Lifetimes in Gnutella Networks," in *Proc. IEEE INFOCOM*, May 2007, pp. 391–399.
[34] X. Wang, Y. Zhang, X. Li, and D. Loguinov, "On Zone-Balancing of Peer-to-Peer Networks: Analysis of Random Node Join," in *Proc. ACM SIGMETRICS*, Jun. 2004, pp. 211–222.
[35] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice Hall, 1989.
[36] Z. Yao, D. Leonard, X. Wang, and D. Loguinov, "Modeling Heterogeneous User Churn and Local Resilience of Unstructured P2P Networks," in *Proc. IEEE ICNP*, Nov. 2006, pp. 32–41.
[37] Z. Yao, X. Wang, D. Leonard, and D. Loguinov, "On Node Isolation under Churn in Unstructured P2P Networks with Heavy-Tailed Lifetimes," in *Proc. IEEE INFOCOM*, May 2007, pp. 2126–2134.
[38] B. Y. Zhao, L. Huang, J. Stribling, S. C. Rhea, A. D. Joseph, and J. Kubiatowicz, "Tapestry: A Resilient Global-Scale Overlay for Service Deployment," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 1, pp. 41–53, Jan. 2004.



Zhongmei Yao (S'06/ACM S'06) received the B.S. degree in engineering from Donghua University, Shanghai, China, in 1997, the M.S. degree in computer science from Louisiana Tech University, Ruston, in 2004, and the PhD degree in computer science from Texas A&M University, College Station, in 2009. She is currently an Assistant Professor of Computer Science at the University of Dayton, Dayton, OH. Her research interests include P2P networks, stochastic modeling, and performance analysis.

Dmitri Loguinov (S'99–M'03–SM'08) received the B.S. degree (with honors) in computer science from Moscow State University, Russia, in 1995 and the Ph.D. degree in computer science from the City University of New York, New York, in 2002. He is currently an Associate Professor in the Department of Computer Science and Engineering at Texas A&M University, College Station. His research interests include P2P networks, information retrieval, congestion control, Internet measurement and modeling.

SUPPLEMENTAL MATERIAL

6 RELATED WORK

Among the recent studies of link lifetimes, one direction focuses on never-switching P2P systems. Leonard *et al.* [13] show that heavy-tailed lifetimes allow link lifetime $E[R]$ to be significantly larger than user lifetime $E[L]$. Additional results of this model and its application to unstructured networks are available in [14], [36], [37]. Another recent study [32] examines DHTs without switching with a focus on the *delivery ratio*, which is the fraction of time that all forwarding nodes between each source and destination are alive. Their results show that the delivery ratio is a function of link lifetime R for all examined neighbor-selection techniques.

Another direction covers switching networks exemplified by traditional DHTs. Godfrey *et al.* [9] study the impact of node-selection techniques on the churn rate and observe that switching DHTs exhibit dramatically smaller link lifetimes than never-switching networks. In their notation, switching/never-switching are agnostic neighbor replacement strategies, where the former is called Active Preference List (APL) and the latter encompasses both Passive Preference List (PPL) and Random Replacement (RR). Krishnamurthy *et al.* [12] compute the probability that neighbors in Chord are in one of three states (alive, failed, or incorrect) and use this model to predict lookup consistency and query latency.

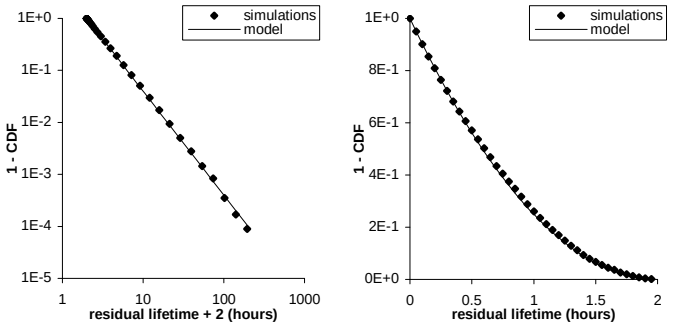
Additional work [2], [5], [15], [16], [17], [27], [31] focuses on measurement and simulation of structured P2P systems under churn.

7 SIMULATIONS

Before we show experimental results of discrete-event simulations, we define rules for generating DHTs under churn. In simulations, user arrivals follow a Poisson process with a constant rate $E[N]/E[L]$, where the mean system size $E[N]$ and the average user lifetime $E[L]$ are determined a-priori. Each user departs at the end of its lifetime L, which is drawn from a given distribution $F(x)$. In addition, each joining user obtains a uniformly random hash index in $[0, 1)$, follows the random-split algorithm during join, and performs recovery when its successors die. After the system has evolved for enough time, we compare simulation results to the derived models to assess their accuracy in finite graphs and systems with age $t < \infty$.

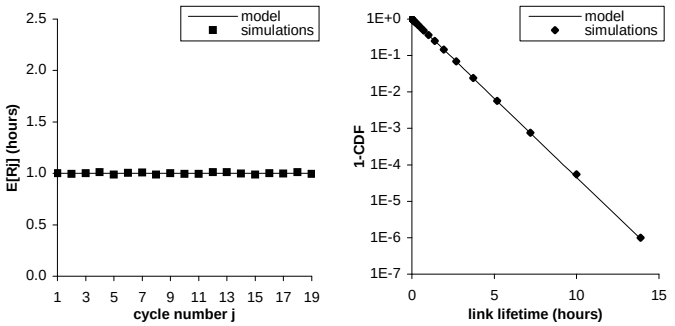
7.1 Residual Lifetimes of Initial Neighbors

Simulations of residual lifetimes Z_j of initial neighbors for $j = 2, 3$ and two lifetime distributions are shown in Fig. 8. As demonstrated by the figure, Lemma 1 correctly predicts that recovery obtains neighbors whose residuals can be considered drawn uniformly randomly from the system and whose residual lifetimes are given by (11). This result holds for both heavy-tailed (e.g., Pareto) and light-tailed (e.g., uniform) user lifetimes. Additional simulations for larger j and other lifetime distributions confirming (11) are not shown here for brevity.



(a) Z_2 for Pareto L with $\alpha = 3$ (b) Z_3 for uniform L

Fig. 8. Comparison of simulation results to model (11) in a deterministic DHT with $E[N] = 1,000$. In both cases, $E[L] = 1$ hour.



(a) $E[R_j]$ (b) distribution of R_3

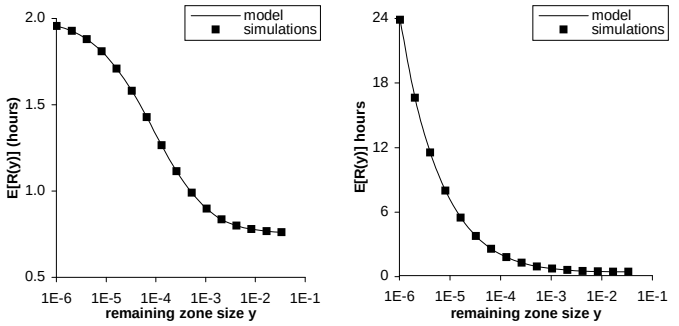
Fig. 9. Comparison of model (13) to simulations in a deterministic DHT with $E[N] = 2,000$ and exponential user lifetimes with $E[L] = 1$ hour.

7.2 Link Lifetime

We next verify the model of R_j for both exponential and Pareto cases. In the former case, the accuracy of (13) is shown in Fig. 9. Notice from the left subfigure that $E[R_j]$ is equal to the mean user lifetime $E[L]$ and from the right subfigure that the distribution of R_j is indeed exponential, which holds for any $j \geq 1$ (only R_3 is shown in the figure). In the latter case, Fig. 10 shows simulation results of $E[R(y)]$ for several values of remaining zone sizes y and plots the corresponding model from Theorem 5. Besides the accuracy of the model, notice from this figure that as remaining zone size y reduces, $E[R(y)]$ increases and converges to $E[Z_1]$, where the distribution of neighbor residual lifetime Z_1 is given in (11).

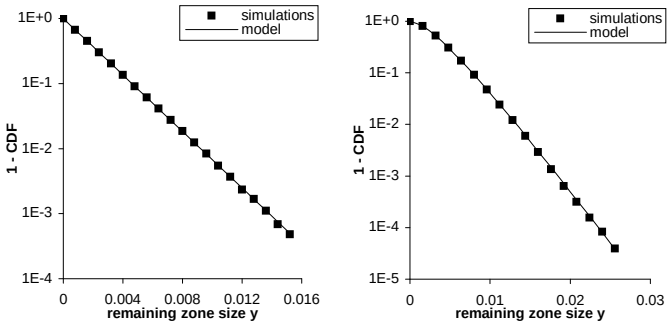
7.3 Zone Size

As demonstrated in Fig. 11, model (17) is very accurate even for small average system size $E[N] = 500$ users. Additional simulation results confirming (17) for larger $E[N]$ and different j are not shown for brevity.



(a) $\alpha = 3$ (b) $\alpha = 1.5$

Fig. 10. Comparison of model $E[R(y)]$ in Theorem 5 to simulation results in a deterministic DHT with mean size $E[N] = 2,000$ and Pareto user lifetimes L with mean $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.



(a) distribution of Y_1 (b) distribution of Y_2

Fig. 11. Comparison of simulation results of Y_j to model (17) in a deterministic DHT with mean size $E[N] = 500$ under churn produced by Pareto L with $\alpha = 3$ and $E[L] = 1$ hour.

8 PROOFS

8.1 Theorem 1

Proof: For the heterogeneous churn model of [36] used in this work, new user arrivals into the DHT space approach a Poisson process with constant rate [36, Theorem 5]:

$$\lambda = \frac{E[N]}{E[L]}, \tag{18}$$

where $E[N]$ is the mean number of users in an equilibrium system and $E[L]$ is the mean user lifetime. Then from the Marked Poisson theorem [26], the arrival process into any fixed zone with size y is Poisson with average rate:

$$\lambda_0 = \lambda q_0, \tag{19}$$

where $q_0 = y$ is the probability that a given zone of length y is selected from the DHT space $[0, 1)$. This indicates that the wait time W_0 to transition from state 0 to state 1 (i.e., the delay before the next arrival into the remaining zone of size y between the neighbor pointer and the current neighbor) is exponentially distributed with rate λ_0.

Next, as the given zone is successively divided by new arrivals over time, its length is reduced over time, which in turn reduces the user arrival rate into the zone. Since a given zone of length y is uniformly divided under random split by a new arrival, the expected length of the new zone is simply $y/2$. Recalling the technique used in (19), we obtain that the wait time W_i to transition from state i to state $i + 1$ is exponential with rate:

$$\lambda_i = \lambda q_i = \frac{E[N]}{E[L]} \cdot \frac{y}{2^i}, \quad i \geq 0, \tag{20}$$

where the probability of selecting the new zone is $q_i = y/2^i$, which depends not only on state i, but also the initial zone size y.

We now consider transitions to state F. Given $A_\delta = i$, $i \geq 1$, a jump to state F is triggered by the departure of the current user, which happens L time units after the chain arrives to state i, where L is the random user lifetime. For state $i = 0$, the delay before the jump to state F is slightly different and equals the original user's remaining lifetime Z_j where j is the cycle number of R_j. It then follows that due to the independence among user departures and arrivals in a sufficiently large system, the sojourn time τ_i in state i is simply:

$$\tau_i = \begin{cases} \min(W_0, Z_j) & i = 0 \\ \min(W_i, L) & i \geq 1 \end{cases}, \tag{21}$$

where $W_i \sim \exp(\lambda_i)$ and is independent of Z_j and L. Since Z_j and L may follow general distributions, respectively, sojourn time τ_i may have a non-exponential distribution.

Finally, transition probability $p_{i,i+1}$ from state i to $i + 1$ is given by:

$$p_{i,i+1} = \begin{cases} P(W_0 < Z_j) & i = 0 \\ P(W_i < L) & i \geq 1 \end{cases}, \tag{22}$$

and the probability $p_{i,F}$ to absorb from state i is equal to $1 - p_{i,i+1}$. Note that due to $W_i \to \infty$ for $i \to \infty$, it is clear that $p_{i,i+1} \to 0$ as $i \to \infty$ and the decay rate is exponentially fast. Thus, $\{A^y_\delta\}$ is regular.

Recognizing that these transitions behave like a discrete-time Markov chain and sojourn times in states depend only on their current states and follow general distributions, we immediately conclude that $\{A^i_\delta\}$ is a regular semi-Markov chain (SMC). □

8.2 Theorem 2

Proof: Generalize the first hitting time from any starting state $i \geq 0$ to state F as:

$$T_{iF} = \inf\{\delta > 0 : A^y_\delta = F | A^y_0 = i, Y_j = y\} \tag{23}$$

and define the following Laplace transform for T_{iF}:

$$\hat{T}_{iF}(s) = E[e^{-sT_{iF}}] = \int_0^\infty e^{-st} dF_{T_{iF}}(t), \tag{24}$$

where $F_{T_{iF}}(t)$ is the CDF of T_{iF}. Then, from first-step analysis, (24) can be transformed into:

$$E[e^{-sT_{iF}}] = p_{iF} E[e^{-s\tau_i}] + p_{i,i+1} E[e^{-s(\tau_i + T_{i+1,F})}], \tag{25}$$

where p_{ik} is the transition probability from state i to k shown in (6). Noting that τ_i is independent of $T_{i+1,F}$ and conditioning on the current state being i, (25) reduces to:

$$E[e^{-sT_{iF}}] = p_{iF}E[e^{-s\tau_i}] + p_{i,i+1}E[e^{-s\tau_i}]E[e^{-sT_{i+1,F}}] = \hat{q}_{iF}(s) + \hat{q}_{i,i+1}(s)E[e^{-sT_{i+1,F}}], \quad (26)$$

where $\hat{q}_{i,k}(s)$ is defined in (8). Using the above recurrent functions and observing that $\hat{q}_{i,i+1}(s) \to 0$ for $i \to \infty$ (due to transition probability $p_{i,i+1} \to 0$ in this case), we readily obtain:

$$E[e^{-sT_{0F}}] = \hat{q}_{0F}(s) + \sum_{k=1}^{\infty}\left(\prod_{i=0}^{k-1}\hat{q}_{i,i+1}(s)\right)\hat{q}_{kF}(s), \quad (27)$$

which establishes (9) upon recalling that $R(y)$ is defined as T_{0F}. □

8.3 Theorem 3

Proof: Given that the chain currently is in state $i \geq 0$, it can jump either to state F or $i+1$. Then by conditioning on the first jump, it is not hard to see that:

$$E[T_{iF}] = E[\tau_i] + p_{i,i+1}E[T_{i+1,F}], \quad (28)$$

where T_{iF} is defined in (23). Using the above recurrence functions, we easily obtain:

$$E[R(y)] = E[T_{0F}] = E[\tau_0] + p_{01}E[T_{1F}] = E[\tau_0] + p_{01}\left(E[\tau_1] + p_{12}E[T_{2F}]\right) = E[\tau_0] + \sum_{k=1}^{\infty}\left(\prod_{i=0}^{k-1}p_{i,i+1}\right)E[\tau_k], \quad (29)$$

where the last step is obtained by induction and recalling that $p_{i,i+1} \to 0$ for $i \to \infty$. □

8.4 Theorem 5

Proof: Since $Z_j \sim \text{Pareto}(\alpha - 1, \beta)$ for all $j \geq 1$, we obtain the distribution of sojourn time τ_0 in state 0 from (4):

$$P(\tau_0 > t) = P(W_0 > t)P(Z_j > t) = e^{-\lambda_0 t}\Big(1 + \frac{t}{\beta}\Big)^{-(\alpha-1)}, \quad (30)$$

where λ_0 is given in (5). Then, we easily get the PDF of τ_0:

$$f_{\tau_0}(t) = -\frac{dP(\tau_0 > t)}{dt} = \lambda_0 e^{-\lambda_0 t}\Big(1 + \frac{t}{\beta}\Big)^{-(\alpha-1)} + \frac{\alpha - 1}{\beta}e^{-\lambda_0 t}\Big(1 + \frac{t}{\beta}\Big)^{-\alpha}, \quad (31)$$

and its mean:

$$E[\tau_0] = \int_0^{\infty} P(\tau_0 > t)dt = \int_0^{\infty} e^{-\lambda_0 t}\big(1 + \frac{t}{\beta}\big)^{-\alpha+1}dt = \beta e^{\lambda_0\beta}E_{\alpha-1}(\lambda_0\beta), \quad (32)$$

where $E_k(x) = \int_1^{\infty} e^{-xu}\ u^{-k}du$ is the generalized exponential integral. Next, the transition probability p_{01} from state 0 to 1 can be computed from (6) as:

$$p_{01} = P(W_0 < Z_j) = \int_0^{\infty} P(W_0 < t)f_Z(t)dt = \int_0^{\infty}\left(1 - e^{-\lambda_0 t}\right)\frac{\alpha - 1}{\beta}\left(1 + \frac{t}{\beta}\right)^{-\alpha}dt = 1 - (\alpha - 1)e^{\lambda_0\beta}E_{\alpha}(\lambda_0\beta) = \lambda_0\beta e^{\lambda_0\beta}E_{\alpha-1}(\lambda_0\beta) = \lambda_0 E[\tau_0], \quad (33)$$

where the last step is established upon recalling (32). Substituting (31) and (33) into (8) and doing certain algebra, we obtain the Laplace transforms of the semi-Markov kernel starting from state 0:

$$\hat{q}_{01}(s) = p_{01}\int_0^{\infty} e^{-st}f_{\tau_0}(t)dt = \lambda_0 E[\tau_0] \times [1 + (1 - \lambda_0 - s)\beta e^{(\lambda_0+s)\beta}E_{\alpha-1}((\lambda_0 + s)\beta)], \quad (34)$$

$$\hat{q}_{0F}(s) = (1 - \lambda_0 E[\tau_0])\left[1 + (1 - \lambda_0 - s)\beta \times e^{(\lambda_0+s)\beta}E_{\alpha-1}((\lambda_0 + s)\beta)\right]. \quad (35)$$

Laplace transforms $\hat{q}_{i,i+1}(s)$ and $\hat{q}_{iF}(s)$, $i \geq 1$ can be obtained by replacing λ_0 with λ_i and $\alpha - 1$ with α in the above equations. Invoking Theorems 2-3, we have the desired result. □

8.5 Lemma 2

Proof: We assume that the probability that a user of any given zone size departs is equally likely (i.e., zone sizes do not depend on user lifetimes and vice versa). Then, given that hash index X_i of any user i is uniformly random in $[0, 1)$ at any time t, it is well-known that zone sizes U are uniformly distributed on the simplex $\{(x_1, \cdots, x_N)|x_i \geq 0; \sum x_i = 1\}$ [7]. It follows that conditioning on $N = z$, the probability that a zone of size x from a given point X_i of user i is unoccupied by the remaining $z - 1$ users is simply:

$$P(U > x|N = z) = (1 - x)^{z-1}. \quad (36)$$

Note that $(1 - x)^{z-1}$ can be transformed into:

$$(1 - x)^{z-1} = e^{(z-1)\log(1-x)} = e^{-x(z-1)+O(x^2)(z-1)}, \quad (37)$$

where the expansion uses the Taylor approximation of $\log(1 - x)$. Substituting (37) into (36) and keeping in mind that $x = o(1/\sqrt{E[N]})$, we obtain:

$$\frac{P(U > x|N = z)}{e^{-xz}} = e^{x+O(x^2)(z-1)} \to 1, \quad (38)$$

as $E[N] \to \infty$.

For the heterogeneous user churn model, recall from [36, Lemma 1] that N is a Gaussian variable with PDF $f_N(z)$. The distribution $P(U > x)$ can then be computed by integrating $P(U > x|N = z)$ with respect to z:

$$\lim_{E[N]\to\infty}\frac{P(U > x)}{e^{-E[N]x}} = \frac{\int_0^{\infty} e^{-xz}f_N(z)dz}{e^{-E[N]x}}, \quad (39)$$

where the last step is obtained by using (38). It then follows from (39) that:

$$\lim_{E[N]\to\infty} \frac{P(U>x)}{e^{-E[N]x}} = \frac{e^{-E[N]x+Var[N]x^2/2}}{e^{-E[N]x}}, \quad (40)$$

since e^{-xN} is a lognormal random variable. Recalling $Var[N] < E[N]$ [36, Lemma 1] and $x^2E[N] \to 0$ as $E[N] \to \infty$, (40) yields:

$$\lim_{E[N]\to\infty} \frac{P(U>x)}{e^{-E[N]x}} = 1, \quad (41)$$

which is the desired result. Finally, note that the requirement of $x^2E[N] \to 0$ is tight and cannot be relaxed for computing the distribution of U. □

8.6 Lemma 3

Proof: Due to the memoryless property of the exponential limiting distribution of U shown in (16), the remaining zone size Y_1 from a neighbor pointer, which randomly splits the zone of some neighbor u, to the hash index of u follows the same distribution of U.

Next, note that Y_j, $j \geq 2$, is the initial zone size of a replacement neighbor u obtained by user v during each recovery. At this time, replacement neighbor u covers its own zone as well as that of the failed user. Thus, it is clear that $Y_j = Y_1 + U$, which has the same distribution as $U + U$. It then immediately follows that Y_j, $j \geq 2$, has the Erlang-2 distribution since it is a sum of two exponentials. □

9 Flexible Fingers

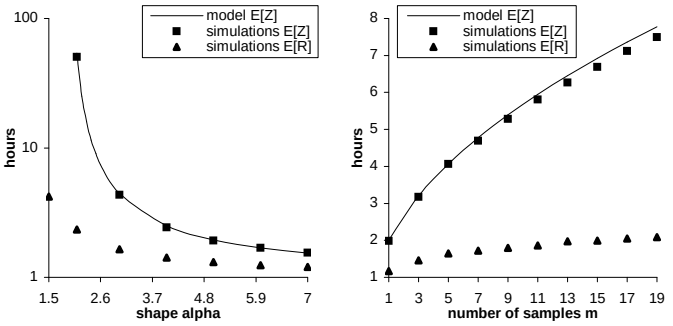
We start by formalizing the well-known principle of *multiple choices* in flexible DHTs. Assume user v has a set S_i of possible values for finger i. It then selects m random points $f_i^1, \ldots, f_i^m$ from S_i and applies a certain function $best(.)$ to chose the location of its finger, i.e.,

$$f_i = best(f_i^1, \ldots, f_i^m). \quad (42)$$

As the system churns, switching DHTs maintain $neighbor_i = owner(f_i)$ at all times. Prior literature [32], [37] has suggested that in never-switching scenarios v should choose the zone whose owner has the largest age. In our context, we call this method *switching max-age* (SMA). Intuition suggests that yet another strategy is possible if v prefers neighbors with the smallest residual zone size, which we call *switching min-zone (SMZ)*.

9.1 Analysis of SMA

It is clear that link lifetimes R_j for all cycles $j \geq 1$ have the same distribution since the neighbor pointer in each cycle is uniformly randomly generated within a certain range of users. Simulation results of SMA and the model of $E[Z_j]$ from [37] are shown in Fig. 12. Notice from part (a) that for a fixed number of samples $m = 6$, as shape α decreases, the mean link lifetime $E[R_j]$ increases much slower than the mean residual lifetime $E[Z_j]$ of the initial neighbor as α decreases.



(a) $m = 6$ (b) $\alpha = 3$

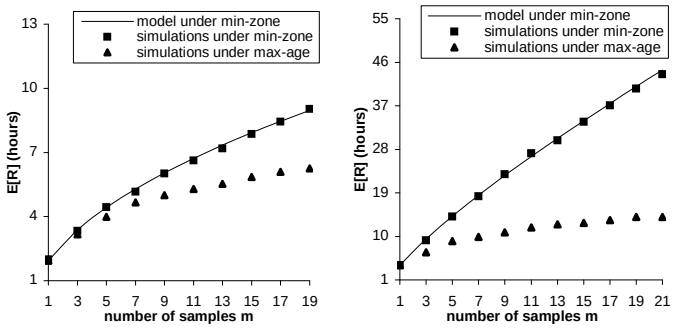
Fig. 12. Impact of shape α and number of samples m on mean link lifetime $E[R_j]$ under SMA in randomized Chord with mean size $E[N] = 2,000$ for Pareto lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

A similar phenomenon appears in part (b) where $E[Z_j]$ increases at the rate of $\sqrt{m}$ for $\alpha = 3$ (see [37, Lemma 5]), while $E[R_j]$ rises from 1.17 hours to only 2.09 hours as m increases from 1 to 19. This demonstrates that the improvement in terms of the mean link lifetime $E[R_j]$ under age-dependent selection is generally very small since new arrivals sooner or later split initial neighbors to take ownership of the link and hence ages or residual lifetimes of original neighbors do not affect link churn rate very much.

9.2 Analysis of SMZ

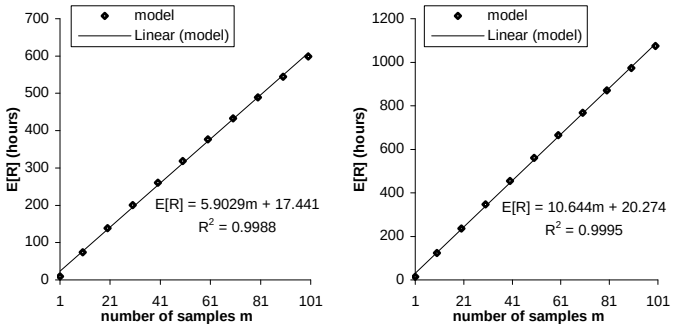
To obtain a model for $E[R_j]$ under SMZ, first note that residual lifetime Z_j of the initial neighbor starting the j-th cycle follows the distribution given in (11) since all m samples are uniformly random and zone sizes are independent of user ages or lifetimes. It is then clear that for a fixed remaining size $Y_j = y$, the Laplace transform and the mean conditional link lifetime given in Theorem 5 are both still valid. Next, recall from Lemma 3 that the residual zone size Y_1 of each sample $f_i^1, \ldots, f_i^m$ is exponential with rate $1/E[N]$. It then follows that their minimum is also exponential with rate $m/E[N]$. The final step is to combine Theorem 5 and the new distribution of Y_1 to obtain the distribution of R_j and its mean under SMZ.

As shown in Fig. 13, the model of $E[R_j]$ matches simulation results very well. Most interestingly, the figure demonstrates that the mean link lifetime $E[R_j]$ under SMZ is significantly larger than that under SMA for both choices of α and that the difference between the two metrics becomes more pronounced as the number of samples m increases or shape α decreases. Furthermore, this figure suggests that as $m \to \infty$, $E[R_j]$ for SMZ and $\alpha < 2$ goes to infinity, while $E[R_j]$ for SMA converges to some fixed number regardless of α. The following theorem confirms this result.



(a) $\alpha = 1.5$ (b) $\alpha = 1.2$

Fig. 13. Comparison of mean link lifetime $E[R_j]$ under SMZ to that under SMA in randomized Chord with mean size $E[N] = 2,000$ for Pareto user lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.



(a) $\alpha = 1.09$ (b) $\alpha = 1.06$

Fig. 14. Approximation of $E[R_j]$ as a linear function of number of samples m under SMZ for Pareto user lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

Theorem 6. *For Pareto user lifetimes with $1 < \alpha \leq 2$, the expected link lifetime under SMZ approaches infinity for sufficiently large system population and random sample size:*

$$\lim_{E[N]\to\infty} \lim_{m\to\infty} E[R_j] = \infty. \tag{43}$$

For SMA and any α, the mean link lifetime converges to a constant:

$$\lim_{E[N]\to\infty} \lim_{m\to\infty} E[R_j] < \infty. \tag{44}$$

Proof: To obtain $E[R_j]$ under SMZ for $m \to \infty$, first note that $P(Y_1 > y) \approx e^{-my/E[N]} \to 0$ as $m \to \infty$ for all fixed $y > 0$. This indicates that $Y_1 \to 0$ in probability. It is then clear that the probability that a new arrival splits a given zone with size Y_1 also approaches 0, and hence in the limit R_j is simply residual lifetime Z_j of the initial neighbor. Recalling from (11) that $E[Z_j] = \infty$ for $\alpha \leq 2$, we immediately obtain $E[R_j] \to E[Z_j] = \infty$ as $m \to \infty$. The condition $E[N] \to \infty$ is required for $m \to \infty$.

When SMA is used, it is shown in [37, Theorem 5] that residual lifetimes $Z_j \to \infty$ with probability 1 as $m \to \infty$ for Pareto lifetimes. It is then easy to obtain using the semi-Markov chain $\{A^y_\delta\}$ in Theorem 1 that sojourn time τ_0 in state 0 is $\min(Z_j, W_0) \to W_0$ as $m \to \infty$, where W_0 is exponential with rate λ_0 given in (5), and transition probability $p_{0,1} = P(W_0 < Z_j) \to 1$. After the chain jumps into state 1, sojourn times are $\min(L, W_i)$, which are no longer affected by the number of samples m. Hence, $E[R_j]$ is finite since the mean sojourn time in each state i is finite and the probability that the chain jumps into the failed state increases exponentially fast. □

Assuming multiple-selection model (42), the above analysis indicates that SMZ is significantly better than SMA for very heavy-tailed user lifetimes. Since real systems have been observed to exhibit $\alpha \approx 1.06$ in [4] and $\alpha = 1.09$ in [33], this result paves a simple way for building better DHTs in practice. The amount of actual improvement in $E[R_j]$ for these two values of α is shown in Fig. 14, where the growth rate in both curves is approximately linear in m. The figures also show the corresponding linear fits to the model, which can be used to predict how m affects link lifetime $E[R_j]$ in these two cases. For instance, with $\alpha = 1.09$, users can obtain $E[R_j] \approx 76$ hours by sampling $m = 10$ points for each suitable link in a randomized DHT. For $\alpha = 1.06$, the corresponding average link lifetime is 127 hours. Comparing these numbers to $E[R_j] \approx E[L] = 1$ hour in deterministic DHTs, the extent of improvement is undoubtedly dramatic.

9.3 Hybrid Age-Zone (HAZ)

If the DHT is not restricted to placing fingers into one of the m generated points as in (42), both age-based and zone-based selection, as well as switching and never-switching DHTs, can be covered by a more general technique we call *Hybrid Age-Zone* (HAZ). In this method, v obtains m random points $f_i^1, \ldots, f_i^m$ for link i and chooses the neighbor with the largest age, but actual finger placement emulates SMZ. Specifically, assume $l \geq 1$ is some integer and s_i is the zone size of the chosen max-age neighbor. Then, define $\{\varsigma_j\}_{j=1}^l$ to be l i.i.d. uniform random variables in $[0, s_i]$ and $Q_i = \min(\varsigma_1, \ldots, \varsigma_l)$ to be their minimum, i.e., $P(Q_i > x) = (1 - x/s_i)^l$. Then, the finger is placed at distance Q_i from the chosen neighbor:

$$f_i = succ(best(f_i^1, \ldots, f_i^m)) - Q_i, \tag{45}$$

where function $best(.)$ selects the point whose owner has the maximum age and $succ(x)$ is the successor of x. During v's lifetime, f_i stays fixed and v remains connected along link i to $owner(f_i)$ regardless of churn.

This method has two tuning knobs. Parameter m controls the tradeoff between join overhead and resilience of the initial neighbor. Parameter l controls the tradeoff between frequency of switching (and thus routing load through new users) and "longevity" given to the initial neighbor. It should be noted that the main model of the paper, fed with proper zone-size distributions and neighbor residuals, covers link lifetimes of $HAZ(m, l)$; however, exploring the myriad of options along the entire 2D plane of (m, l) is beyond our scope. Instead, we make several simple observations.

Notice that combination $HAZ(1,1)$ corresponds to rigid fingers, $HAZ(m,1)$ to SMA, $HAZ(1,l)$ to SMZ, but without the extra overhead, and $HAZ(m,\infty)$ to never-switching max-age (NSMA). As shown above, SMA without the SMZ component increases overhead, but provides very little resilience improvement and is thus not advisable. Therefore, m should be increased only if l is sufficiently large, where the actual thresholds depend on the lifetime distribution, desired load-balancing, and constraints on join overhead.

10 Delayed Joins

If link lifetime is the primary objective of a DHT design, then clearly flexible fingers and never-switching (i.e., HAZ with $l = \infty$) is the optimal solution. However, it is unclear if any fundamental improvement can be achieved for *rigid* fingers and, if so, how this impacts flexible DHTs. We next explore the possibility of changing the lifetime distribution of joining users and making it more heavy-tailed.

Given Pareto lifetimes, one possibility is to prevent young (i.e., unreliable) peers from holding objects and links until their age increases beyond a certain threshold $\eta \geq 0$. In other words, a peer arrives in the DHT and establishes its outgoing finger pointers as usual; however, it is not allowed to receive in-links or hold any objects until its age $A(t)$ exceeds η. Having out-fingers allows all users to search the system and store their own keys on remote peers, but there is no responsibility (e.g., storing of other peers' keys or routing of their queries) until these new arrivals have statistically "proven" their resilience to the community. While this concept has found application in unstructured P2P systems (e.g., Gnutella requires a certain amount of uptime and bandwidth before a node can become an ultra-peer), there is nothing stopping it from being deployed in DHTs.

In order to understand this system, notice that one can apply earlier analysis in the paper by substituting $F(x)$ with a new user lifetime distribution:

$$F'(x) = P(L' < x) = P(L < x + \eta | L > \eta), \qquad (46)$$

where L' is the random variable describing the residual lifetime of users with age $A(t) = \eta$.

The amount of improvement in $E[L']$ vs $E[L]$, as well as residuals $E[Z'_j]$ vs $E[Z_j]$, depends on the tail of the original lifetime distribution. Assuming Pareto $F(x)$ with parameters (α, β), we get from [37, page 5] that L' is Pareto with parameters $(\alpha, \beta + \eta)$, which leads to:

$$E[L'] = E[L] + \frac{\eta}{\alpha - 1}, \alpha > 1 \qquad (47)$$

and

$$E[Z'_j] = E[Z_j] + \frac{\eta}{\alpha - 2}, \alpha > 2. \qquad (48)$$

Both rigid and flexible fingers obtain resilience benefits. In the former case, $E[R'_j] \approx E[L']$ and, in the latter case, $E[R'_j] = E[Z'_j]$ assuming the simplest case of NSMA with $m = 1$. Interestingly, both improve by the same factor:

$$r = \frac{E[L']}{E[L]} = \frac{E[Z'_j]}{E[Z_j]} = 1 + \frac{\eta}{\beta}. \qquad (49)$$

Also observe that the fraction of users that survive at least η time units is simply $r^{-\alpha}$. Thus, for a given r, smaller α allows more users to participate in the system and thus achieve better load distribution.

TABLE 1
Link Lifetimes under Delayed Joins

Delay η (min)	$E[L']$ (hours)	Ratio r	Reliable fraction
0	0.5	1	100%
6	2.2	4.4	21%
12	3.8	7.6	12%
18	5.5	11	7.8%
30	8.8	17.6	4.8%
45	13	26	3.1%
60	17	34	2.4%

Several examples are shown in Table 1 for $E[L] = 0.5$ hours and $\alpha = 1.06$. Observe in the table, that 21% of the users can deliver a DHT with $r = 4.4$ times larger link lifetimes under both rigid and flexible fingers. This requires delaying each join by just 6 minutes. With 4.8% of the graph, the DHT can offer an 17.6-fold improvement, while the top 2.4% of the peers can increase link lifetimes by a factor of 34, all of which is quite significant in practice. Additional benefits, including avoiding frequent key transfers to arriving users, key loss during departure, and disruption in routing, make delayed-join systems significantly more robust in practice.