DEPARTMENT OF TELECOMMUNICATIONS AND MEDIA INFORMATICS
BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS

MODELING, OPTIMIZATION AND PERFORMANCE EVALUATION OF DISTRIBUTED HASH TABLE OVERLAYS

Collection of Ph.D. Theses
by
Péter Kersch

Research Supervisor:

Róbert Szabó, Ph.D.
Department of Telecommunications and Media Informatics

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
BUDAPEST, HUNGARY

2008

1 Introduction

Originally emerging as file sharing applications, the peer-to-peer principle has proven to be very successful in a number of different areas of application ranging from Internet telephony (e.g. Skype [1]) via video on demand and Internet television (e.g. SopCast [2]) to grid computing. Nowadays, peer-to-peer traffic accounts for more than half of the overall Internet traffic [3]. Among many advantages of the peer-to-peer principle, the elimination of bandwidth, processing and storage bottlenecks, savings on deployment and operation costs and the difficult traceability of peer-to-peer users all contributed to this success.

Generally speaking, a peer-to-peer network can be defined as a self-organizing system of equal, autonomous entities (peers), which aims for the shared usage of distributed resources in a networked environment avoiding central services [4]. One of the most basic functionalities of such a peer-to-peer system is the lookup of shared resources in the network. This functionality can be implemented in two fundamentally different ways. In unstructured peer-to-peer systems, peers form an unstructured low diameter overlay network, in which shared resources can be looked up using limited flooding or random walk search. In contrast, structured peer-to-peer systems, e.g. Distributed Hash Tables (DHTs), create and maintain a distributed indexing structure where resource location information can be found iteratively by a directed lookup process without requiring network-wide flooding.

In a DHT, similarly to ordinary hash tables, each data element is associated with a key. Key space is partitioned among all participating peers and each peer is storing data elements whose key lies within its respective key space partition. Keys and peer identifiers are mapped to a same ID space and a closeness metric is used to define distances in this ID space. Key space partitioning is based on this metric: data is stored at the node(s) whose ID is the closest to the key of the data. To access data belonging to key space partitions of remote peers, nodes forming a DHT are organized into a structured overlay that can be described by a graph. Data in a DHT is stored and retrieved using a routing algorithm on top of this overlay.

My dissertation investigates two main aspects of routing process in DHT overlays. In Theses 1, I have proposed a generic mathematical model to describe relationship between overlay structure and static routing performance for a large class of DHTs. Then, in Theses 2, I have investigated maintenance of DHT overlays and have proposed a novel asymptotically optimal stochastic overlay maintenance strategy.

2 Objectives

A huge number of Distributed Hash Tables variants have been proposed [5, 6, 7, 8, 9, 10, 11] since the publication of the very first structured peer-to-peer systems in 2001. Although architectural and algorithmic details of these DHT proposals can

differ significantly, the foundations of lookup mechanisms are very similar for most of them. There are several empirical studies (based on simulations) comparing static and dynamic performance of different DHT routing mechanisms using various parameter settings [12, 13]. There exist also detailed analytical models for some DHTs, however, these models are usually restricted to one specific DHT implementation. Finally, some aspects of DHT routing are covered by generic models, e.g., static resilience of DHT routing against failure [14] or the impact of lookup strategy, lookup parallelism and replication on DHT routing performance under churn [15]. However, to the best of my knowledge, there exist no generic analytical models capturing the relationship between overlay structure and routing performance of DHTs in static networks. *The objective of Theses 1 is to fill this gap proposing a generic stochastic model of DHT overlays and overlay routing which covers a large family of DHTs.*

Most P2P systems are inherently dynamic; peers join, leave and rejoin the network frequently. This process of permanent joining and leaving of peers is called *churn* [16] in peer-to-peer terminology. To cope with this dynamism and provide good lookup performance under churn, DHTs need to implement several architectural components, e.g.: replica maintenance, static resilience [11, 12, 14], churn-tolerant lookup strategies [15, 17], overlay maintenance [8, 10], etc. *The objective of These 2 is to investigate and propose a novel, stochastic approach for overlay maintenance in order to minimize maintenance overhead under churn.*

3 Methodology

In Theses 1, mathematical analysis has been the main approach. To provide a generic model of overlay routing in a large subclass of DHTs, I have applied a logarithmic transformation to distances in the DHT metric space. In the obtained transformed view, I have used stochastic analysis (mainly renewal theory) to describe both long-range connection distribution and routing progress. To derive closed form formulas for various distribution parameters, I have used the *Mathematica* software.

In Theses 2, I have applied the obtained mathematical model to design a novel stochastic long-range connection maintenance algorithm. To evaluate analytically performance of the proposed maintenance algorithm in a network under churn, I have used a linear equation system combined with a continuous time Markov-chain model. I have also validated analytical results by extensive simulations up to network sizes of 128k nodes. In order to be able to simulate very large networks and analyze scalability, I have implemented algorithms in a cycle-based simulator and ignored transport layer mechanisms.

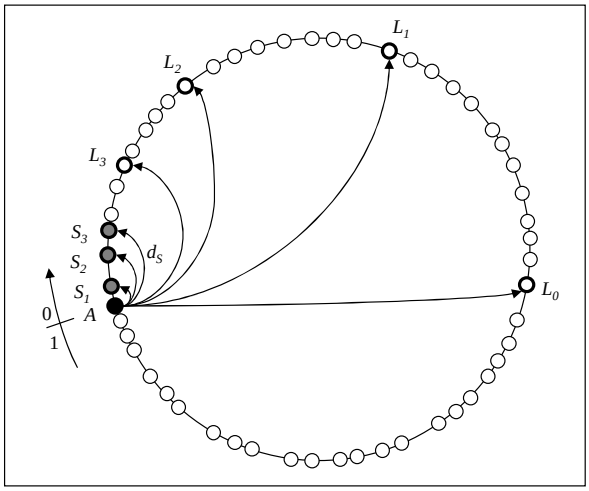

Figure 1: General model to describe DHT routing overlays (example)

3.1 Modeling assumptions and notations

To describe routing overlay of DHTs, I have used the terminology and reference model defined in [18]. The metric space – into which node identifers are mapped to in a DHT – plays a key role in the analysis of DHT routing. I have considered a one dimensional Euclidean metric space within the interval $[0, 1)$ that wraps around (this can be represented as a ring, see Figure 1). Distance between two nodes in this metric space is defined as their distance along the ring in clockwise direction[1], formally: $d(x, y) = y - x + I_{x>y}$.

Each node has two different types of connections to other nodes: short-range connections (called "local" connection in [18]) to a fixed number (N_S) of closest nodes (in clockwise direction) and long-range connections to some distant nodes. These nodes are called short-range and long-range peers of the node, respectively. Figure 1 shows short-range connections (S_1, S_2, S_3) and long-range connections (L_0, L_1, L_2, L_3) of node A.

Routing is assumed to be greedy: a node forwards a lookup request to its peer being the closest to the target node in the metric space of the DHT (without overshooting it).

Routing overlay of many DHT implementations (Chord [5], Pastry [6], Symphony [9], Accordion [10] etc.) can be described (or approximated) using the above system model (e.g. routing in Pastry is more complex but is based on the same greedy algorithm). However, there are a few exceptions, for example DHTs using multidimensional metric spaces (e.g., CAN [7]) or non-Euclidean metric spaces (e.g., Kademlia

[1]This definition implies that the metric space is in fact only a quasi-metric space, since it does not satisfy the symmetry requirements. Extending the model to bidirectional routing where distance is defined as the shortest path along the ring, a real metric space can be obtained.

[8]). In Theses 1, I analyze in more details two extreme overlay families, deterministic and probabilistic power-law routing overlays:

Definition 1 (Probabilistic power-law routing overlay (PPLRO))**.** A routing overlay is called probabilistic power-law routing overlay when the choice of long-range connections is not deterministic and they only have to satisfy the following requirement: the probability of having a long-range connection to another overlay node is inversely proportional to the d^{th} power of the distance between the two nodes in the d dimensional metric space[2] embedding node identifiers [19]. Join algorithm of probabilistic routing overlays create initial long-range connections of joining nodes according to this distance distribution.

Definition 2 (Deterministic power-law routing overlay (DPLRO))**.** A routing overlay over a one-dimensional metric space is called deterministic power-law routing overlay if long-range peers are determined by the power series of the distances $d_i = \frac{q}{c^i}$ where c and q are constant so that $c > 1$ and $0 < q \leq 1$. For unidirectional overlays, the i^{th} long-range connection is chosen as the first node whose distance exceeds d_i while for bidirectional overlays, the i^{th} connection is the node closest to the point at distance d_i.

A deterministic power-law routing overlay can be thought of as a generalization of the Chord [5] overlay (for Chord, $c = 2$ and $q = 1$).

The unidirectional overlay and routing model presented above can be easily extended to bidirectional overlays with bidirectional routing. In a bidirectional overlay, both short-range and long-range connections are bidirectional. The other important difference is the distance metric of the DHT metric space: using the ring representation, distance of two nodes is defined as their shortest distance along the ring, formally: $d_b(x, y) = min\,[d(x, y), d(y, x)]$. As a result – from the point of view of distances – each node can split the DHT metric space into two symmetrical partitions. Connections of a node are created independently in both of these partitions using the distance distribution of the unidirectional equivalent of the overlay. Greedy routing in bidirectional overlays also becomes bidirectional; requests can be forwarded in both directions depending on the position of the peer node being the closest to the target. Theses 1 present results for unidirectional routing overlays while Theses 2 extend theses results and apply them for bidirectional routing overlays.

The distribution of node identifiers in the metric space of the DHT also affects mathematical analysis of routing in the overlay. Throughout my dissertation, I have assumed that node identifiers are drawn independently at random according to a uniform distribution[3] (this is a reasonable assumption in most cases). This implies that distances between adjacent IDs on the ring will be exponentially distributed. In

[2]In my thesis work, analysis is restricted to one dimensional metric spaces.

[3]Strictly speaking, the ID space is discrete for any real system. However, the granularity of this discrete ID space (typically between $2^{128} - 2^{256}$) is so large that it can be considered continuous.

a few cases (explicitly noted), I have assumed that peer identifiers partition the metric space of the DHT deterministically in equal partitions. This is not a realistic scenario, but simplifies considerably mathematical analysis. In these cases, I've compared approximate analytical result with simulations using random uniform distribution of peer identifiers.

Finally, when modeling long-range connection selection, I have assumed that it is possible to find a peer node at any given distance in the metric space. This is not realistic in a real system composed of a finite number of nodes. In practice, the closest existing node to the given point is used instead. However, the resulting error between these theoretical and real distances is inversely proportional to the size of the network, hence this is negligible for large networks (which are the main scope of my dissertation).

4 New Results

4.1 Stochastic modeling of overlay structure and routing in distributed hash tables

Although the myriad of different DHT variants might seem significantly different at first sight, routing in most DHT overlays is based on the same foundations – being structurally similar to the "small-world" navigation model of Kleinberg [19]. To obtain a better understanding of these common foundations, I have investigated analytically overlay structure and routing process and proposed a general mathematical model that can be used to describe and compare a large subclass of DHTs.

THESES 1: ***[J2, J3, B1]*** *Using stochastic methods and a logarithmic distance transformation, I have analyzed the overlay structure and the routing process in DHTs. I have proposed a set of generic overlay parameters to characterize overlay structure in a large subclass of DHTs. Furthermore, I have derived closed form upper bounds on the expected number of routing hops in static networks as a function of network size and these overlay parameters.*

Analysis of long-range connection distribution

THESIS 1.1: ***[J2, J3, B1]*** *I have proposed a logarithmically transformed view where long-range connections of a node in most DHTs form a linear or quasi-linear sequence. I have identified a large subclass of DHTs where this sequence can be described (or approximated) as a random sample from an infinite renewal process. To characterize long-range connection distribution in these DHTs, I have introduced a λ long-range connection density and a c_v long-range connection density coefficient of variation parameter. For $O(\log n)$ node state DHTs, these parameters characterize the overlay independent of network size.*

Definition 3 (Logarithmically transformed view)**.** Let $(\mathcal{I}, d)$ be the metric space embedding node identifiers of a DHT where the distance between a node x_0 and another node x_i is defined as $d(x_0, x_i)$. Then the distance of x_0 and x_i in the logarithmically transformed view of x_0 is defined as:

$$d'(x_0, x_i) = -\ln\left[d(x_0, x_i)\right]. \tag{1}$$

The transformed view of a base node x_0 can be used to characterize distances between x_0 and a set of other nodes in a DHT. This transformed view can be represented along a half-line $[0, \infty]$ as follows: the base node x_0 itself is in $+\infty$ while other DHT nodes x_i (e.g., peers of the base node, or the target node of a lookup process) are represented along this half-line at distance $d'(x_0, x_i)$ from the 0 point.

To demonstrate linearity of the sequence of long-range connections in the transformed view of a node, Figure 2 represents long-range connections of nodes in various DHT implementations: Chord [5], Pastry [6] Kademlia [8][4] and probabilistic power-law routing overlays (e.g. Symphony [9] or Accordion [10]). For each of these DHTs, the upper line shows long-range peers of the node in the real metric space of the DHT (to ease graphical representation, the ring geometry of the metric space have been straightened) while the lower line shows these peers in the logarithmically transformed view of the node. In the real metric space, the represented node is in point 0. In the transformed view, this point corresponds to $+\infty$. Finally, long-range connections in the transformed view span within the range $[0, -\ln d_S)$, where d_S is the distance from the farthest short-range peer of the node (represented by grey circle in Figure 2).

Although long-range connection distribution differs for each of the above DHT implementations, Figure 2 shows that it is possible to partition the long-range connection domain in the transformed view into equally sized partitions of length Δx so that the number of long-range connections $N_L(\Delta x)$ be the same inside each of these partitions (either deterministically or in expected value). Based on this observation, I have defined a $\lambda_{\Delta x}$ long-range connection density parameter as:

$$\lambda_{\Delta x} = \frac{E\left[N_L(\Delta x)\right]}{\Delta x}. \tag{2}$$

For $O(\log n)$ node state DHTs, this $\lambda_{\Delta x}$ parameter characterizes long-range connection distribution of the DHT overlay independent of network size. In general, the choice of Δx is not arbitrary. In order to obtain constant long-range connection density in the entire long-range connection domain of the transformed view, Δx might need to be set to a DHT specific value (see again Figure 2).

I have identified an important family of DHTs where the sequence of long-range connections in the transformed view of the node is even more "regular":

[4]For Pastry, the parameter b is the bit length of numbers in the routing table, while the parameter k for Kademlia is the maximum size of buckets

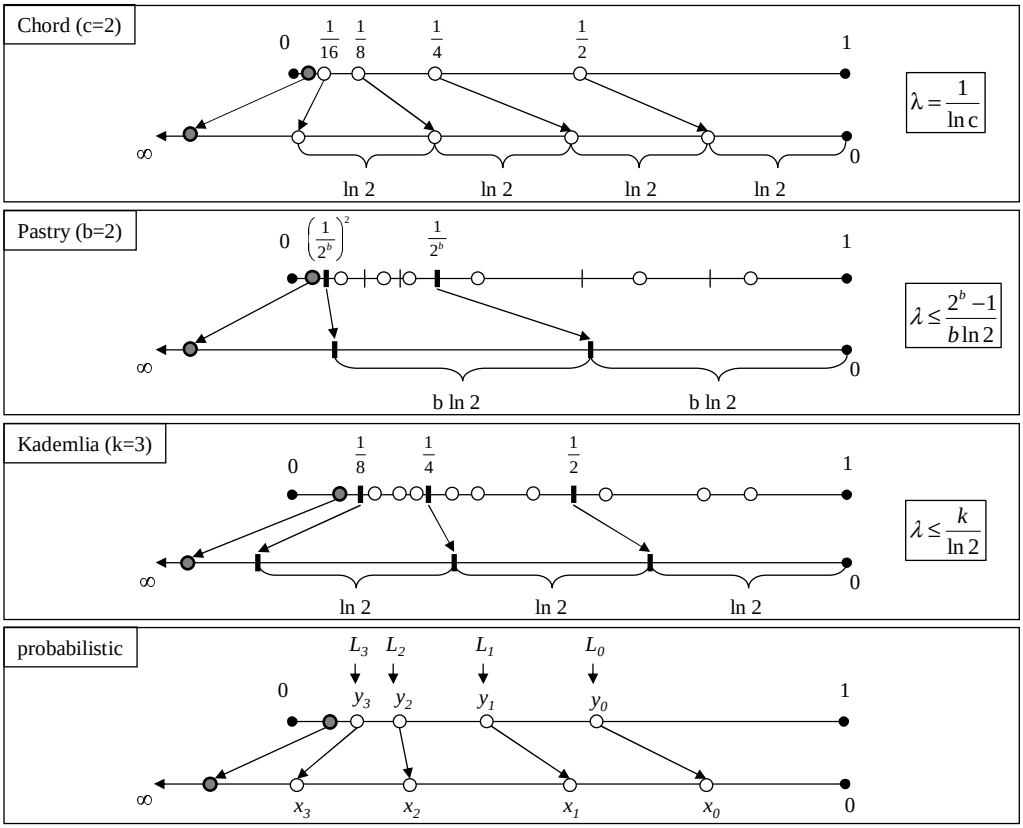

Figure 2: Comparison of the routing table of some well known DHTs

Definition 4 (Regular power-law routing overlays (RPLRO))**.** A power-law routing overlay is called regular if the sequence of long-range connections of a node in its transformed view corresponds to a randomly chosen sample of length $(-\ln d_S)$ from an infinite renewal process and these random samples are chosen independently for each node.

Hence distances between subsequent long-range connections of a node in its transformed view are *i.i.d* and the distribution function $F(x)$ of these *i.i.d* random variables identifies unambiguously a regular power-law routing overlay. For regular power-law routing overlays, long-range connection density can be defined independent of Δx as

$$\lambda = \lim_{\Delta x \to 0} \frac{E\left[N_L(\Delta x)\right]}{\Delta x} = \frac{1}{\mu} \tag{3}$$

and $\lambda = \frac{1}{\mu}$ uniformly for any distances within the long-range connection domain, where μ is the mean distance between subsequent long-range connections of a node in its transformed view.

Probabilistic power-law routing overlays are regular (see Thesis 1.2). Pastry [6]

and Kademlia [8] are not regular but are close to being regular with only small distortions. Finally, Chord [5] and deterministic power-law routing overlays in general are not regular, but, they can be made regular: Considering the transformed view of a node in a DPLRO, its first long-range peer is always located at $\ln c$. Substituting the constant q in Definition 2 by a random variable so that this first long-range peer be evenly distributed in the range $[0, \ln c]$, the overlay becomes regular.

Probabilistic power-law routing overlays form a very special subset of regular power-law routing overlays:

THESIS 1.2 [J2, J3, B1]: *I have proved that the sequence of long-range connections of a node in a probabilistic power-law routing overlay can be described as a random realization of a truncated Poisson process in the transformed view of this node and long-range connection density of this overlay is equal to the λ intensity of this Poisson process.*

Analysis of routing

Routing in most DHTs is based on the combined use of short-range and long-range connections[5]. The first routing hops usually take place via long-range connections while the last hops usually take place via short-range connections. Except for completely deterministic routing overlays, the probability of routing via a short-range connection increases monotonously approaching to the target and it is not possible to clearly separate routing process into distinct long-range and short-range routing phases.

As a first approximation, in order to allow analytical study of this complex dual routing process, I have restricted analysis to routing via long-range connections and replaced short-range connections by additional imaginary long-range connections. The sequence of long-range connections in the transformed view of a node spans until the farthest short-range connection of this node. When considering long-range only forwarding, this sequence continues until infinity in this transformed view as imaginary long-range connections. Hence, for regular power-law routing overlays, the concatenation of the sequence of real and imaginary long-range connections corresponds to an infinite sample from an infinite renewal process. In the real routing overlay, forwarding takes place via a short-range peer only when the target node is closer than the closest long-range peer of the forwarding node. When the real routing overlay is forwarding via a short-range peer, the long-range only model is forwarding through an imaginary long-range peer at a smaller distance.

Figure 3 demonstrates this difference between real forwarding via a short-range connection and long-range only forwarding via an imaginary long-range connection. The upper line in the figure represents the real metric space while the lower line shows the transformed view of the forwarding node. For a real overlay, forwarding occurs

[5]There are a few exceptions: e.g., Kademlia [8] nodes only have long-range connections while CAN [7] nodes only have short-range connections.

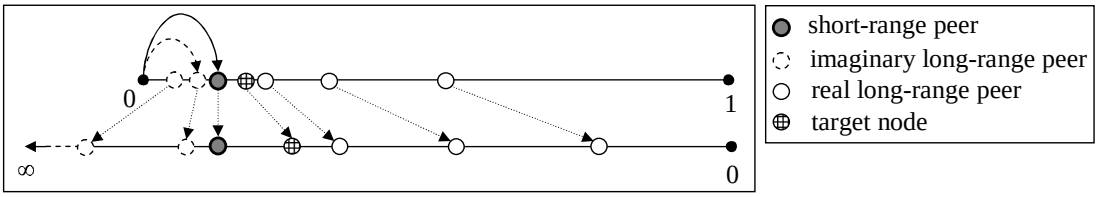

Figure 3: Forwarding through imaginary long-range connections

via a short-range connection. However, using the long-range only model, forwarding takes place via the imaginary long-range connection being the closest to the target node.

I have proven that the expected value of the per-hop routing progress is always smaller for the long-range only model than for the real routing overlay, hence results obtained from the long-range only model can be used as a lower bound on per-hop routing progress. In Theses 1.3, 1.4, 1.5 and 1.6, I have analyzed long-range only routing using the renewal process model of Thesis 1.1. Then, in Thesis 1.7 and 1.8, I have used the obtained results to derive closed-form upper bounds on the expected number of routing hops as a function of network size and overlay parameters (for real routing via both short-range and long-range connections).

THESIS 1.3: ***[B1]*** *Considering long-range only forwarding in a regular power-law routing overlay, let F_k and F_{k+1} be the forwarding nodes in hops k and $k+1$ of a lookup process. Furthermore, let u_k be the random variable corresponding to the per hop routing progress, defined as the distance between the images of forwarding nodes F_k and F_{k+1} in the transformed view of the target node. I have proved that (i) the sequence of the random variables u_k for subsequent hops of a lookup process are* i.i.d *and (ii) the expected value of u_k can be lower bounded by:*

$$E[u_k] \geq -\ln\left[1 - e^{-\frac{1+c_v^2}{2\lambda}}\right], \tag{4}$$

where λ is the long-range connection density and c_v is the long-range connection density coefficient of variation in the overlay.

An important consequence of Thesis 1.3 is that the progress of routing via long-range connections can be also modeled using renewal theory in the transformed view of the target (note that the renewal process describing long-range connections of a node and the renewal process describing the progress of routing are different).

As expected, the lower bound on the expected per-hop progress increases monotonously with increasing λ and decreases monotonously with increasing c_v. For two specific regular power-law routing overlay families (probabilistic and regularized deterministic overlays) I have also derived the closed form distribution of the random variable u as a function of the λ long-range connection density:

THESIS 1.4 [J2, B1]: *For probabilistic power-law routing overlays, I have derived the distribution of the per-hop routing progress u via long-range connections in the transformed view of the target node:*

$$f_{prob}(u) = \lambda(1 - e^{-u})^{(\lambda-1)}e^{-u} \tag{5}$$

and

$$F_{prob}(u) = (1 - e^{-u})^{\lambda}, \tag{6}$$

where λ is the long-range connection density of the overlay.

As a consequence, the expected value of the per hop progress towards the target in its transformed view is:

$$E_{prob}[u] = \int_0^{\infty} f_{prob}(u)u\, du = H_{\lambda}, \tag{7}$$

where H_x is the harmonic number of x (generalized for real numbers).

Substituting $\lambda = 1$ into Equation 5 gives the *pdf* of the exponential distribution with parameter 1. Hence, $\lambda = 1$ is a very special long-range connection density value for probabilistic power-law routing overlays where routing progress can be described by a Poisson process of rate 1 in the transformed view of the target node.

Results of Thesis 1.4 can be transformed back from the transformed view of the target node to the real metric space of the DHT as follows:

THESIS 1.5: ***[J3, B1]*** *I have derived the distribution and the expected value of the ratio of distances from the target after and before a routing hop via a long-range connection for a probabilistic power-law routing overlay:*

Consider the routing process in a probabilistic power-law routing overlay of long-range connection density λ. Then the series of random variables $w_k = \frac{d_{k+1}}{d_k}$ describing the ratio of distances from the target after and before a routing hop via a long-range connection are i.i.d *and the* pdf *and expected value of w_k are:*

$$f^w_{prob}(w) = \lambda(1 - w)^{(\lambda-1)} \quad \textit{if} \quad 0 < w < 1 \quad \textit{and 0 otherwise} \tag{8}$$

and

$$E[w] = \frac{1}{1 + \lambda}. \tag{9}$$

THESIS 1.6 [B1]: *For "regularized" deterministic power-law routing overlays, I have derived the distribution of the per-hop routing progress via long-range connections in the transformed view of the target node:*

$$f_{det}(u) = \begin{cases} \lambda\dfrac{e^{-u}}{1 - e^{-u}} & \textit{if} \quad u > -\ln\left[1 - e^{\frac{1}{\lambda}}\right] \\ 0 & \textit{otherwise} \end{cases} \tag{10}$$

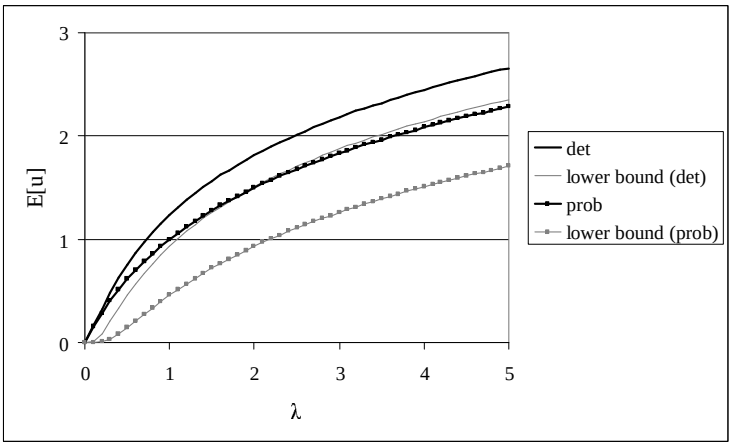

Figure 4: Expected value of u and its estimated lower bound as a function of λ

The relationship between the λ long-range connection density and the parameter c used in the Definition 2 is the following:

$$\lambda = \frac{1}{\ln c} \quad \Leftrightarrow \quad c = e^{\frac{1}{\lambda}}. \tag{11}$$

Figure 4 compares $E[u]$ (the expected per hop routing progress in the transformed view of the target) and its lower bounds obtained from the general formula of Thesis 1.3 for deterministic and probabilistic power-law routing overlays as a function of long-range connection density. As expected, for the same long-range connection densities, per-hop routing progress is higher for DPLROs ($c_v = 0$) than for PPLROs ($c_v = 1$).

Using Thesis 1.3 and the Lorden bound [20] from renewal theory, it is possible to derive a closed form upper bound on the expected number of routing hops using long-range only forwarding. This upper bound can in turn be used to derive an upper bound for real routing (via both short-range and long-range connections) on the expected number of routing hops as a function of network size:

THESIS 1.7: ***[B1]*** *I have derived an upper bound on the expected number of routing hops U in a regular power-law routing overlay:*

$$U(n, \lambda, c_v, N_S) \quad \leq \quad \frac{\ln n - H_{N_S-1} - 0.42}{-\ln\left[1 - e^{-\frac{1+c_v^2}{2\lambda}}\right]} + \frac{2.41\lambda}{\ln^2\left[1 - e^{-\frac{1+c_v^2}{2\lambda}}\right]} + \epsilon, \tag{12}$$

where n is the number of nodes in the overlay, λ is the long-range connection density, c_v is the long-range connection density coefficient of variation, N_S is the number of short-range connections per node and ϵ is a small positive error term

$$\epsilon \in o\left(\frac{n^{N_S}}{e^n}\right)$$

negligible except for very small network sizes.

For regular power-law routing overlays, where $E[u]$ and $E[u^2]$ can be expressed in closed form, the upper bound of Thesis 1.7 can be further improved:

THESIS 1.8: ***[B1]*** *I have provided a tighter upper bound on the expected number of routing hops U in probabilistic power-law routing overlays:*

$$U(n, \lambda, N_S) \quad \leq \quad \frac{\ln n - H_{N_S-1} - 0.42}{H_\lambda} + \frac{1.645 - \psi'(1+\lambda)}{H_\lambda^2} + 1 + \epsilon, \qquad (13)$$

where n is the number of nodes in the overlay, λ is the long-range connection density, and N_S is the number of short-range connections per node, $\psi'(x)$ is the first derivative of the digamma function and ϵ is a small positive error term

$$\epsilon \in o\left(\frac{n^{N_S}}{e^n}\right)$$

negligible except for very small network sizes.

4.2 Asymptotically optimal stochastic overlay maintenance in DHTs

THESES 2 [**J3**]: *I have proposed a novel architectural component to further improve DHT performance under churn. This architectural component applies stochastic long-range connection maintenance to a probabilistic power-law routing overlay with bidirectional connections. I have shown – both analytically and by extensive simulations – that the proposed architectural component guarantees asymptotically minimal maintenance overhead.*

The role of short-range and long-range connections is complementary: while short-range connections guarantee the success of routing, long-range connections expedite the routing process to $O(\log n)$ hops. These different roles imply different requirements and optimization opportunities in the maintenance process. To guarantee successful routing, short-range connection maintenance has to be strict, self-stabilizing and proactive. In contrast, to guarantee $O(\log n)$ routing performance, long-range connections only have to meet probabilistic requirements. In these theses, I have proposed a maintenance solution which meets these requirements while reducing long-range connection maintenance overhead to the theoretical lower bound.

THESIS 2.1 [**J3**]: *Using the Poisson process model of Thesis 1.2, I have shown a self-healing property of long-range connections in probabilistic power-law routing overlays with bidirectional connections under churn. Assuming a steady state when node arrival and departure rates are equal, distance distribution of long-range connections remains unchanged without any explicit maintenance solely as a side-effect of new connections established by joining nodes.*

Although the self-healing property of probabilistic power-law routing overlays with bidirectional connections provides solid foundations to minimize maintenance overhead, it does not entirely replace maintenance mechanisms. When the departure rate of nodes is higher than the arrival rate of new nodes, lost long-range connections need to be created by maintenance process. The contrary is required for growing networks where arrival rate of new nodes is higher than departure rate: without removing some connections, long-range connection density of nodes would grow continuously resulting in extreme node degrees for older nodes.

THESIS 2.2 [J3]: *I have proposed a stochastic long-range connection maintenance algorithm for probabilistic power-law routing overlays with bidirectional connections, which ignores individual connections and only considers the distance distribution of long-range connections. I have shown analytically that in steady state (where departure/failure and arrival rate of new nodes are equal) the per node average rate at which new long-range connections are created or deleted by the maintenance algorithm is upper bounded by a constant independent of network size. Additionally, the algorithm provides balanced node degree distribution and guarantees an average routing performance which is a function of a λ_{opt} system parameter and scales logarithmically with network size.*

The proposed algorithm maintains distance distribution of long-range connections by keeping the $\widehat{\lambda}$ estimated long-range connection density of each node within a range $[\lambda_{min}, \lambda_{max}]$, where $\lambda_{min} = \lambda_{opt} - \Delta\lambda$ and $\lambda_{max} = \lambda_{opt} + \Delta\lambda$. Applying Thesis 1.2, $\widehat{\lambda}$ is calculated for each node by maximum likelihood estimation from the N_L number of its long-range connections and the d_S distance from its farthest short-range peer. Given the bidirectional nature of the overlay, the algorithm uses separate $\widehat{\lambda}_r$ and $\widehat{\lambda}_l$ estimations for right and left side long-range connection densities:

$$\widehat{\lambda}_r = -\frac{N_L^r}{\ln d_S^r} \quad \text{and} \quad \widehat{\lambda}_l = -\frac{N_L^l}{\ln d_S^l}, \tag{14}$$

where d_S^l and d_S^r denotes the distance from the farthest short-range peer at the left and right sides respectively while N_L^l and N_L^r denotes the number of long-range connections at the left and right sides respectively.

The proposed maintenance algorithm is reactive and is triggered either when another node creates a new (bidirectional) connection to this node or when a connection failure is detected (via timeouts). Upon these events, maintenance algorithm recalculates the value of the estimated connection density for both sides ($\widehat{\lambda}_r$ and $\widehat{\lambda}_l$). If $\widehat{\lambda}$ is within the range $[\lambda_{min}, \lambda_{max}]$, then no maintenance actions are taken. Otherwise, if $\widehat{\lambda}$ falls below λ_{min} at either the left or right side, then new connections are created at this side (using the same distance distribution than for joining nodes) until $\widehat{\lambda}$ reaches λ_{opt}. Similarly, if the estimated connection density exceeds λ_{max} then randomly selected connections are deleted until $\widehat{\lambda}$ reaches λ_{opt}.

Let $r_{in} = r_{out} = r$ be the arrival and departure (failure) rate of nodes in the

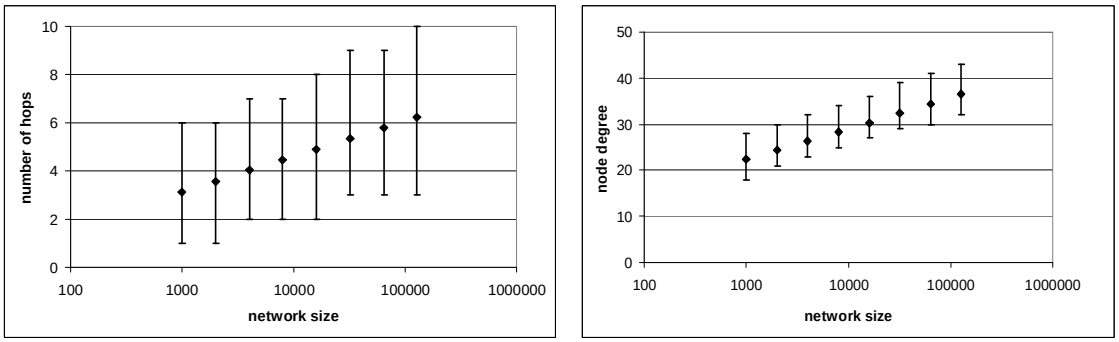

Figure 5: Routing hop and node degree distribution as a function of network size

systems. Furthermore let R_{cm} and R_{dm} be the rate at which long-range connections are created and deleted by maintenance, respectively. All of these rates are normalized by the number of nodes in the systems (n).

Using a linear equation system and a Markov-chain model, I have shown that the following upper bounds hold for long-range connection creation and deletion rates:

$$R_{dm} \leq \frac{\lambda_{opt}}{\Delta\lambda} r \tag{15}$$

and

$$R_{cm} \leq \frac{\lambda_{opt}}{\Delta\lambda} r. \tag{16}$$

As a consequence of Thesis 1.8, the average number of routing hops in the overlay maintained by this algorithm is upper bounded by a function of λ_{min} for a given network size and for a given λ_{opt} setting, it scales logarithmically with network size.

I have also verified these statements using extensive simulations. The left graph in Figure 5 shows the number of routing hops (average, 5^{th} and 95^{th} percentiles) as a function of network size for networks with $\lambda_{opt} = 1/\ln 2$ and $\Delta\lambda = 0.2$ long-range connection density parameters and $N_S = 3$ short-range connections at both sides of nodes. The right graph demonstrates balanced node degree representing its average as well as the 5^{th} and 95^{th} percentiles as a function of network size for the same overlay parameters.

THESIS 2.3 [J3]: *I have proposed a range-based long-range connection establishment algorithm where the expected communication overhead of one connection establishment is upper bounded by a constant independent of network size.*

Instead of creating a connection to the node being the closest to the point drawn at random according to the required distance distribution, I define a small range around the selected point, and connection is created to the first node found in this range during the lookup process. Hence, the first (fastest) lookup hit (node ID) matching the determined range will be used. The selection of this range is illustrated in Figure 6,

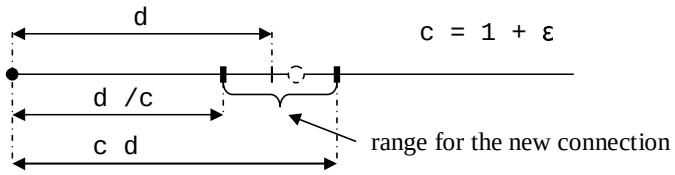

Figure 6: Defining ID range to create new long-range connections

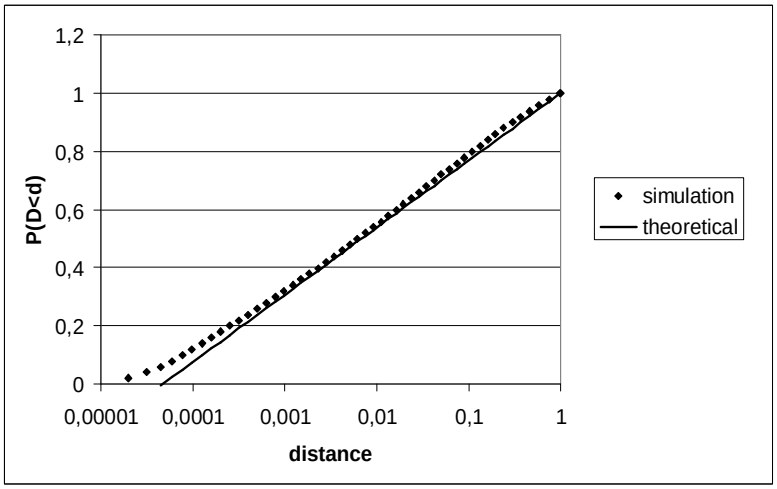

Figure 7: Distance distribution of long-range connections

d is the distance drawn at random to create a new long-range connection. I define a range $[\frac{d}{c}, dc]$ where $c = 1 + \epsilon$, $0 < \epsilon \ll 1$ and ϵ is a constant system parameter. If there are no nodes within this range (the probability of this event increases with smaller d and ϵ values), than the lookup process will terminate at the closest node and connection will be created to this node (outside the range).

I have shown analytically that this method reduces average communication overhead of one long-range connection establishment from $O(\log n)$ to $O(1)$. Using extensive simulations, I have also shown that the small random offsets introduced by the use of the above ranges do not significantly distort the distance distribution of long-range connections in the probabilistic power-law routing overlay. Figure 7 compares the theoretical distance distribution of long-range connections for a probabilistic power-law routing overlay of 128k nodes with $\lambda = 1/\ln 2$ and the empirical distribution function for the same overlay obtained by simulations.

Definition 5 (Network half life)**.** Network half life is defined as the time during which half of the nodes are replaced by new arrivals in the network. [21]. Network half-life characterizes dynamism of a network independent of the number of nodes.

THESIS 2.4 [J3]: *I have shown that using the stochastic long-range connection mechanism proposed in Thesis 2.2 and the sequential range-based connection establishment mechanism proposed in Thesis 2.3, the overall maintenance traffic per node and per network half-life is $O(\log n)$, where n is the number of overlay nodes. This is asymptotically optimal, since $o(\log n)$ had been proven to be a lower bound on maintenance traffic in order to prevent disconnection of any overlay at high probability [21].*

Using Thesis 2.2 and 2.3, Table 1 summarizes various components of long-range connection maintenance overhead per node and per network half-life. Adding these components, I have obtained the $O(\log n)$ bound on overall long-range connection maintenance overhead. Given the constant number of short-range connections, per node short-range connection maintenance overhead does not increase with network size, hence the overall maintenance overhead is also $O(\log n)$.

	conn. cre./del. rate	per conn. overhead	overall overhead
New nodes joining	$O(\log n)$	$O(1)$	$O(\log n)$
Create by maint.	$O(1)$	$O(1)$	$O(1)$
Delete by maint.	$O(1)$	$O(1)$	$O(1)$

Table 1: Overview of the components of long-range connection maintenance overhead

5 Applicability of New Results

New results in Theses 1 are mainly theoretical, helping to understand better the common foundations of routing in distributed hash tables. In addition, the proposed generic stochastic model can also be used to compare analytically the routing performance of different DHT implementations. Routing performance depends on different overlay parameters which are usually incompatible across different DHT families. In the proposed stochastic model, these incompatible parameters can be translated into a common set of overlay parameters $\{\lambda, c_v, N_S\}$. Using this common parameter set, static routing performance of different DHTs can be compared analytically applying the proposed upper bounds on the expected number of routing hops. Finally, these closed form upper bounds can also be used as an input to derive bounds on other DHT performance metrics; e.g., lookup latencies (in both static networks and under a given level of churn) applying the analytical framework presented in [15].

New results in Theses 2 are more application oriented. Since dynamism and churn are inherent properties of most peer-to-peer system, efficient maintenance mechanisms are critical to provide good lookup performance in DHTs. Combining the proposed novel stochastic maintenance mechanism with other churn-tolerant architectural components (e.g., churn tolerant lookup strategies [15], fine tuning of timeout handling [17]), it is possible to further decrease overlay maintenance overhead in DHTs while

preserving high availability and good lookup performance. To demonstrate the applicability of stochastic maintenance, I have implemented a DHT (in a simulation environment) which uses the proposed stochastic maintenance component. Based on extensive simulations, I have also proposed protocol parameter settings (summarized in Table 2) for this DHT implementation in order to achieve the best tradeoff between maintenance overhead, lookup performance and availability (M denotes the number of short-range connection maintenance cycles per network half-life).

parameter	proposed range
M	$[10, 50]$
λ	$[1.4, 3.0]$
$\frac{\Delta\lambda}{\lambda}$	$[0.2, 0.3]$
N_S	$[3, 5]$

Table 2: Proposed protocol parameter ranges

References

[1] S. Guha, N. Daswani, and R. Jain, "An experimental study of the Skype peer-to-peer VoIP system," in *Proceedings of the 5th International Workshop on Peer-to-Peer Systems (IPTPS'06)*, (Santa Barbara, CA, USA), 2006.

[2] A. Sentinelli, G. Marfia, M. Gerla, S. Tewari, and L. Kleinrock, "Will IPTV ride the peer-to-peer stream?," *IEEE Communications Magazine*, vol. 45, no. 6, pp. 86–93, 2007.

[3] H. Schulze and K. Mochalski, "Internet study 2007," tech. rep., Ipoque, 2007.

[4] K. Wehrle and R. Steinmetz, *P2P Systems and Applications*. LNCS 3485, Springer, 2005.

[5] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: Scalable peer-to-peer lookup service for internet applications," in *Proceedings of ACM SIGCOMM*, (San Diego, CA, USA), pp. 149–160, 2001.

[6] A. Rowstron and P. Druschel, "Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems," in *Proceedings of the 18th IFIP/ACM International Conference on Distributed Systems Platforms*, LNCS 2218, (Heidelberg, Germany), pp. 329 – 350, Springer, 2001.

[7] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A scalable content-addressable network," in *Proceedings of ACM SIGCOMM*, (San Diego, CA, USA), pp. 161–172, 2001.

[8] P. Maymounkov and D. Mazieres, "Kademlia: A peer-to-peer information system based on the xor metric," in *Proceedings of the 1st International Workshop on Peer-to-Peer Systems*, LNCS 2429, (Cambridge, MA, USA), pp. 53–65, 2002.

[9] G. S. Manku, M. Bawa, and P. Raghavan, "Symphony: Distributed hashing in a small world," in *Proceedings of the 4th USENIX Symposium on Internet Technologies and Systems*, (Seattle, WA, USA), pp. 127–140, 2003.

[10] J. Li, J. Stribling, R. Morris, and M. F. Kaashoek, "Bandwidth-efficient management of DHT routing tables," in *Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation*, (Boston, MA, USA), pp. 99–114, 2005.

[11] T. Locher, S. Schmid, and R. Watternhofer, "eQuus: A provably robust and locality-aware peer-to-peer system," in *Proceedings of the 6th IEEE International Conference on Peer-to-Peer Computing*, (Cambridge, UK), pp. 3–11, 2006.

[12] K. Gummadi, S. Gribble, S. Ratnasamy, S. Shenker, and I. Stoica, "The impact of DHT routing geometry on resilience and proximity," in *Proceedings of ACM Sigcomm*, (Karlsruhe, Germany), pp. 381–394, 2003.

[13] 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 *Proceedings of INFOCOM '05*, (Cambridge, MA, USA), pp. 225–236, IEEE, Mar. 2005.

[14] J. S. Kong, J. S. A. Bridgewater, and V. P. Roychowdhury, "A general framework for scalability and performance analysis of DHT routing systems," in *Proceedings of the International Conference on Dependable Systems and Networks (DSN'06)*, (Philadelphia, PA, USA), pp. 343–354, 2006.

[15] D. Wu, Y. Tian, and K.-W. Ng, "Analytical study on improving DHT lookup performance under churn," in *Proceedings of the 6th IEEE International Conference on Peer-to-Peer Computing*, (Cambridge, UK), pp. 249–258, IEEE, 2006.

[16] D. Stutzbach and R. Rejaie, "Understanding churn in peer-to-peer networks," in *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement (IMC '06)*, (Rio de Janeiro, Brazil), pp. 189–202, 2006.

[17] S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz, "Handling churn in a DHT," Tech. Rep. UCB/CSD-3-1299, UC Berkeley, Computer Science Division, UC Berkeley, USA, Dec. 2003.

[18] K. Aberer, L. O. Alima, A. Ghodsi, S. Girdzijauskas, S. Haridi, and M. Hauswirth, "The essence of P2P: A reference architecture for overlay networks," in *Proceedings of the 5th IEEE International Conference on Peer-to-Peer Computing*, (Konstanz, Germany), pp. 11–20, 2005.

[19] J. M. Kleinberg, "The small-world phenomenon: an algorithmic perspective," in *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, (Portland, OR, USA), pp. 163–170, 2000.

[20] F. Beichelt and L. P. Fatti, *Stochastic Processes and Their Applications*. CRC Press, 2001.

[21] D. Liben-Nowell, H. Balakrishnan, and D. Karger, "Analysis of the evolution of peer-to-peer networks," in *Proceedings of the 21st Annual Symposium on Principles of Distributed Computing (PODC '02)*, (Monterey, CA, USA), pp. 233–242, 2002.

Publications

[B] BOOK CHAPTERS

[**B1**] **Péter Kersch**, Róbert Szabó. Analysis of DHT routing. Invited chapter in the *Handbook of Peer-to-Peer Networking*, Springer, 2008. (submitted)

[J] JOURNALS

[**J1**] **Péter Kersch**, Róbert Szabó. A graph theoretical lower bound on maintenance overhead of structured P2P overlays. *PIK, Vol. 31. No. 1, pp 24–28, special issue on Modeling of Self-Organizing Systems*, March. 2008. DOI: 10.1515/piko.2008.005

[**J2**] **Péter Kersch**, Róbert Szabó. DHT routing analysis in a logarithmically transformed space. *Peer-to-Peer Networking and Applications, Vol 1. No. 1, pp 64–74*, Springer, March 2008. DOI: 10.1007/s12083-007-0002-2

[**J3**] **Péter Kersch**, Róbert Szabó, Lawrence Cheng, Kerry Jean, Alex Galis. Stochastic maintenance of overlays in structured P2P systems. *Elsevier Computer Communications, Vol 31/3 pp 603–619, special issue on Disruptive Networking with Peer-to-peer Systems*, Febr. 2008. DOI: 10.1016/j.comcom.2007.08.017

[**J4**] Kis Zoltán Lajos, Kovácsházi Zsolt, **Kersch Péter**, Simon Csaba. Mobil többesadás protokollok vizsgálata IPv6 hálózatokban. *Híradástechnika*, Vol. LIX. pp. 20–25, March 2004.

[**J5**] **Kersch Péter**, Vajda Lóránt, Török Attila. IP mikromobilitási protokollok ad hoc kiterjesztése. *Híradástechnika*, Vol. LVIII. pp. 14–19, Apr. 2003.

[**J6**] **Kersch Péter**, Kürthy Lóránt, Simon Csaba, Vajda Lóránt. IP mikromobilitási protokollok ad hoc kiterjesztésének tesztelése. *Híradástechnika*, Vol. LVIII. pp. 20–28, Apr. 2003.

[C] CONFERENCES

[**C1**] Lawrence Cheng, Roel Ocampo, Kerry Jean, Alex Galis, Zhaohong Lai, Csaba Simon, Robert Szabo, **Peter Kersch**, Raffaele Giaffreda. Distributed Hash Tables Composition in Ambient Networks. In *Proceedings of IEEE DSOM'06,* 17^{th} *IFIP/IEEE International Workshop on Distributed Systems*, pp.258–268, Dublin, Ireland, October 23–25, 2006 DOI: 10.1007/11907466

[**C2**] **Péter Kersch**, Zoltán Lajos Kis, Róbert Szabó. Self Organizing Ambient Control Space - An Ambient Network Architecture for Dynamic Network Interconnection. In *Proceedings of the* 1^{st} *International ACM Workshop on Dynamic Interconnection of Networks*, pp.17–21, Cologne, Germany, September 2^{nd} 2005. DOI: 10.1145/1080776.1080782

[**C3**] Róbert Szabó, **Péter Kersch**, Balázs Kovács, Csaba Simon, Márk Erdei, Ambrus Wagner. Dynamic Network Composition for Ambient Networks: a Management View. In *Proceedings of Eurescom Summit 2005*, pp.35–42, Heidelberg, Germany, April 27–29 2005.

[**C4**] Csaba Simon, Rolland Vida, **Péter Kersch**, Christophe Janneteau, Gösta Leijonhufvud. Seamless IP Multicast Handovers in OverDRiVE. In *Proceedings of IST Mobile and Wireless Communications Summit*, Lyon, France, June 27–30 2004.

[**C5**] Zoltán Lajos Kis, Zsolt Kovácsházi, **Péter Kersch**, Csaba Simon. Adaptation of IPv6 multicast protocols to heterogeneous mobile networks. In *Proceedings of 10th Eunice Summer School and IFIP WG 6.3 Workshop on Advances in fixed and mobile networks*, pp.99-103, Tampere, Finland, June 14–16 2004.

[**C6**] **Péter Kersch**, Csaba Simon, Lóránt Vajda. Ad Hoc Extension for IP Radio Access Networks. In *Proceedings of TRANSCOM 2003 (*5^{th} *European Conference Of Young Research and Science Workers in Transport and Telecommunications*, pp.191–196, Zilina, Slovakia, June 23-25 2003.

Citations

[CI1] I. Scholtes, S. Kolos, P.F. Zema. The ATLAS Event Monitoring Service — Peer-to-Peer Data Distribution in High-Energy Physics. *IEEE Transactions on Nuclear Science, Vol. 55., No. 3, pp 1610–1620*, Jun. 2008

refers to

[18] P. Kersch, R. Szabo, L. Cheng, K. Jean, and A. Galis, "Stochastic maintenance of overlays in structured P2P systems," *Comp. Commun.*, vol. 31, no. 3, pp. 603–619, 2008.

as follows:

"The efficient handling of high exit rates – commonly called churn – involves self-organization and self-stabilization issues and is still a subject of ongoing research [18]."

[CI2] N. Wakamiya, M. Murata. Bio-inspired Analysis of Symbiotic Networks. in *Managing Traffic Performance in Converged Networks, LNCS Vol. 4516, pp 204–213*, Springer, 2007

refers to

5. Kersch, P., Szabo, R., Kis, Z.L.: Self organizing ambient control space – an ambient network architecture for dynamic network interconnection. In: Proceedings of the 1st ACM workshop on Dynamic Interconnection of Networks (DIN'05), pp. 17–21 (2005)

as follows:

"In [5], they considered a hierarchical overlay model, in which overlay networks were interconnected by an upper level network of representative peers, called super peers. They mentioned two types of composition of overlay networks, they were, absorption and gatewaying. By absorption, two overlay networks which accept each other are merged into one and represented by one super peer. On the other hand, if two overlay networks cannot agree to be merged for some reasons, e.g., incompatibility, they build a new upper level overlay which interconnects them by gatewaying."