Enabling global multimedia distributed services based on hierarchical DHT overlay networks

Isaias Martinez-Yelmo*

Departamento de Ingeniería Telemática,
Universidad Carlos III de Madrid,
Av. Universidad 30, 28911 Leganés, Madrid, Spain
E-mail: imyelmo@it.uc3m.es
Website: http://www.it.uc3m.es/imyelmo/
*Corresponding author

Alex Bikfalvi

IMDEA Networks,
Av. del Mar Mediterráneo 22, 28918 Leganés, Madrid, Spain
E-mail: alex.bikfalvi@imdea.org

Carmen Guerrero and Rubén Cuevas Rumín

Universidad Carlos III de Madrid,
Av. Universidad 30, 28911 Leganés, Madrid, Spain
E-mail: guerrero@it.uc3m.es
Website: www.it.uc3m.es/carmen
E-mail: rcuevas@it.uc3m.es

Andreas Mauthe

Computing Department at InfoLab 21,
Lancaster University, Lancaster LA1 4WA, UK
E-mail: andreas@comp.lancs.ac.uk

Abstract: Providing innovating multimedia services is a high priority for service providers. Due to the high traffic volume created by multimedia content, the use of decentralised services can lead to better solutions. Starting from the ongoing work of P2PSIP, we define a simple way to interconnect different domains using peer-to-peer networks. We define the needed signalling to provide connectivity between different domains based on P2PSIP. This fact allows an easier deployment of global decentralised multimedia services. We validate the proposed solution through an analytical and experimental study of the routing performance and routing state for two possible scenarios.

Keywords: hierarchical overlay network; peer-to-peer; P2PSIP; multimedia.

Reference to this paper should be made as follows: Martinez-Yelmo, I., Bikfalvi, A., Guerrero, C., Rumín, R.C. and Mauthe, A. (2008) 'Enabling global multimedia distributed services based on hierarchical DHT overlay networks', *Int. J. Internet Protocol Technology*, Vol. 3, No. 4, pp.234–244.

Biographical notes: Isaias Martinez-Yelmo received a MSc on Telecommunication Engineering in 2003 from University Carlos III de Madrid, and a MSc on Telematics in 2007 from University Carlos III de Madrid and University Politecnica de Cataluña, both in Spain. He is researcher and teaching assistant in Telematics Engineering Department and PhD student on Telematics at University Carlos III de Madrid since 2004. His research activities are focused on NGN networks, peer-to-peer overlay networks and content distribution networks. He has been involved in several national and international research projects related with content distribution, overlay networks and broadband networks.

Alex Bikfalvi received a BSc Degree in Telecommunications from Technical University of Cluj-Napoca, Romania in 2006 and a MSc Degree from the University Carlos III of Madrid, Spain in 2008. He is currently a PhD candidate at Carlos III University of Madrid, and works as a research assistant at IMDEA Networks Research Institute in Madrid, Spain. His research interests include peer-to-peer overlays and multimedia content distribution in IMS-based

next generation networks. He has been involved in several research projects including IST CONTENT funded by the European Union and BIOGRIDNET supported by the Madrid regional government.

Carmen Guerrero received the Telecommunication Engineering Degree in 1994 from the Technical University of Madrid and the PhD in Computer Science in 1998 from the Universidade da Coruña . She has been an Assistant (1994–2000) and Assistant Professor (2000–2003) at UDC. She is currently Associate Professor since 2003 at Universidad Carlos III de Madrid (UC3M), teaching computer networks courses. Some of the recent research projects in which she has participated are: CONTENT: Content Home Network and Services for Home Users, MUSE: Multiservice Access Everywhere and E-NEXT: Network of Excellence in Emerging Networking Technologies.

Rubén Cuevas Rumín got his MSc in Telecommunication Engineering and MSc in Telematic Engineering at Universidad Carlos III de Madrid in 2005 and 2007, respectively. Furthermore, he obtained his MSc in Network Planning and Management at Aalborg University in 2006. Currently he is PhD Candidate at the Department of Telematic Engineering at University Carlos III de Madrid. His research interests include overlay and P2P networks and content distribution.

Andreas Mauthe is a Senior Lecturer at the Computing Department, Lancaster University. He has been working in the area of distributed and multimedia systems for more than 15 years. His particularly interests are in the area of content management systems and content networks, large scale distributed systems, peer-to-peer systems, and self-organisation aspects. Prior to joining Lancaster University, he headed a research group at the Multimedia Communications Lab (KOM), at the Technical University of Darmstadt. After completing his PhD in Lancaster in 1997, he worked for more than four years.

1 Introduction

Nowadays the provisioning of multimedia services (VoIP, VoD, IPTV, etc.) is one of the most important objectives of ISPs in delivering new and attractive services. However, these multimedia services have not been as widely deployed as would be expected due to their demanding requirements. The success of applications like Skype[1] (Baset and Schulzrinne, 2006; Rossi et al., 2008) is due to their P2P decentralised design despite their relative complexity. The problem with these applications is that they are closed proprietary solutions and their behaviour is difficult to analyse. Thus, a standardised decentralised scalable solution based on P2P overlay networks is desirable to facilitate a large scale deployment of distributed multimedia services on the Internet.

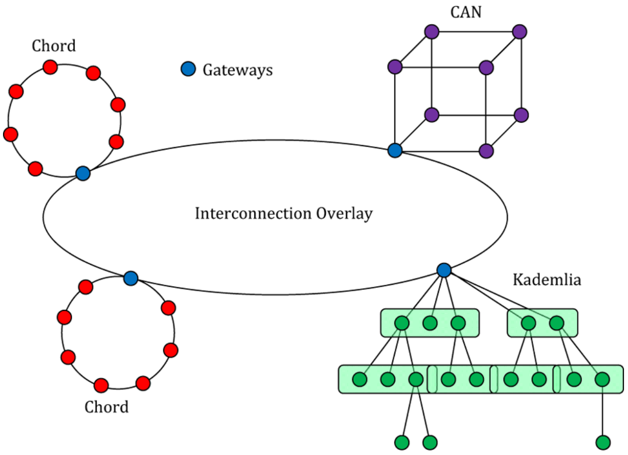
Although a number of solutions have been proposed to support decentralised multimedia services, the new approach of the IETF P2PSIP[2] Working Group is developing into a reference framework. P2PSIP (Bryan et al., 2007) is a peer-to-peer overlay based solution that facilitates a decentralised architecture. The protocol is flexible enough (Jennings et al., 2008) to support most of peer-to-peer networks, allowing the implementation of any DHT overlay network such as Kademlia (Maymounkov and Mazieres, 2002) or Chord (Stoica et al., 2003). However, the design of this protocol does not consider the inter-operation between different domains, a requirement for providing global multimedia services. In this paper we propose a solution to this problem and we assess its Routing Performance (RP) and routing state. Two situations are considered: the first is when the target of the query is independent on the P2PSIP domain, and the second when there is a higher probability of performing a query in the domain where a peer is attached. The latter corresponds to a VoIP service where users belonging to a social network (such as the employees of a company) exchange more calls between them.

Figure 1 illustrates our approach. The idea is that the different domains can deploy their own overlay network and global connectivity between them is established through a dedicated *interconnection overlay*. In this interconnection overlay each domain is represented through at least one *super-peer*. If an item, service or reference, is not in the same domain, a regular peer can asks its super-peer to route the query to the domain of the target peer. To support the routing between the different domains and the interconnection overlay, an extended identifier is used, which is formed by a *prefix ID* for routing in the interconnection and a *suffix ID* for routing in each domain.

Some advantages of this approach are network isolation and the improved scalability, which is intrinsic to the hierarchical architectures. However, issues such as a potential super-peer overload (Beverly Yang and Garcia-Molina, 2003) have to be considered, although this problem, if it occurs, is limited to these peers because the routing state does not increase in legacy peers. For this reason, super peers can be dedicated entities allowing others (such as power-limited handheld device like mobile phones) to be efficiently implemented.

We must highlight that in P2PSIP the overlay networks are used to retrieve the information about users and services, distributed across all peers in the overlay. Usually, this is location information (stored as IP address and port number) and once this information is obtained, the negotiation of the service parameters is done using any suitable signalling protocol. Nevertheless, for compatibility reasons this protocol should be SIP (Rosenberg et al., 2002).

Figure 1 Hierarchical overlay architecture (see online version for colours)



The paper is structured as follows: Section 2 outlines the proposed hierarchical architecture. An analysis of the routing performance and the advantages of this approach are discussed in Section 3. In Section 4 we study the particular case of a hierarchical Kademlia overlay network and validate the theoretical model using the PeerFactSim.Kom simulation framework. Finally, we present the related work in Section 5 and the conclusions and future work in Section 6.

2 Hierarchical DHT overlay networks

2.1 Hierarchical space domain of IDs

In order to provide a hierarchical architecture for interconnecting different domains, and assuming that the main goal of P2PSIP WG is to develop a framework to support any kind of DHT overlay network, we define a hierarchical space of identifiers. The *hierarchical ID* (see Figure 2) is composed by two sub-identifiers: a *prefix ID* and a *suffix ID*. The prefix ID is used for routing in the interconnection overlay between the different domains, whereas the suffix ID is used for routing queries only inside the domain of a peer. This approach can be easily included in the P2PSIP protocol (Jennings et al., 2008) because each header must contain an *overlay ID* which can be used as *prefix ID* and a *node ID* which can be used as *suffix ID*.

Figure 2 Hierarchical ID

Prefix ID (n-bits)	Suffix ID(m-bits)

2.2 Service mapping into the hierarchical ID space domain

One of the main problems in a decentralised architecture is the mapping between the available information and/or services and the peers in the system. For example, if we consider a multimedia environment where services and users are identified by URIs, a resource is identified by `resource@example.com`. In order to map these URIs to the proposed hierarchical ID space domain, the prefix ID is obtained by applying a hash to the domain of the URI: `PrefixID` = `hash(example.com)`. In a similar manner, the suffix ID is generated as the hash of the complete URI: `SuffixID` = `hash(resource@example.com)`.

Once the mapping between the URI and the hierarchical ID has been established, a resource tuple containing the resource hierarchical ID, the URI and the resource itself, is stored at the peer with the closest peer ID. Depending on the DHT protocol this tuple can be also replicated in other peers.

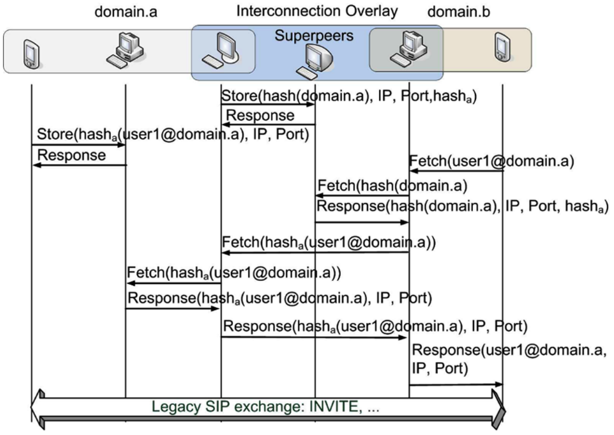
2.3 Hierarchical DHT overlay operation

The behaviour of a hierarchical overlay network can be distinguished into two cases. The first, the search of a resource is limited to the domain where a peer is member. This case is simple because the search for resources is done inside the domain using a flat peer-to-peer overlay and where the routing is based only on the suffix ID. In this situation, the prefix ID of the resource is equal to the hash of the domain name. The second case is the retrieval of a resource in a different domain which is more complex. For instance, this case corresponds to a VoIP call to a user in a different domain. In this circumstance, it is necessary to obtain the contact information published in the domain where the callee is registered. Thus, the initiator sends the query to its own super-peer. Super-peers are selected according to certain characteristics (Min et al., 2006) and the selection mechanism can be integrated in the maintenance protocol of each DHT. Therefore, we can assume that all peers in a domain know their super-peer and they can send a query to the super-peer in one hop. When the super-peer receives the query, it routes the query through the interconnection overlay using the prefix ID. This query will arrive at the super-peer belonging to the domain matching the prefix ID from the query. From this point onwards, the query is forwarded inside the destination domain. If the query reaches a peer that has the desired resource, the peer replies this information following the inverse path. A direct response would be possible if the query includes the {IP address, port} tuple where the requester is waiting the response. This answer must be compliant with the ongoing design of the P2PSIP protocol (Jennings et al., 2008) that it is being defined by the IETF P2PSIP WG. The next section illustrates how to perform this process with P2PSIP.

2.4 Signalling exchange

An example of the signalling from the proposed hierarchical scenario is illustrated in Figure 3, considering the actual status of the P2PSIP protocol (Jennings et al., 2008). In the example, the peer from *domain.b* performs a query for the information related to *user1@domain.a*. For this, it sends a Fetch message to its super-peer that includes the URI of the target, *user1@domain.a*, as plain text. This is because a peer is not required to know the hash functions that are used in the interconnection overlay or in the other domains. The super-peer, which is aware of the interconnection overlay hash function, computes *hash(domain.a)* and forwards the query using the peer-to-peer protocol rules in order to determine the information about *domain.a*. This information includes the address and port of the super-peer and the hash function, $hash_a$, that is used in that domain. Using the query results, the super-peer now sends the Fetch to the super-peer from *domain.a*, which includes $hash_a$*(user1@domain.a)*. In this way we obtain the desired interoperability. Finally, the peers taking care of the desired Resource-ID answer to the super-peer on domain.a, which forwards this information to super-peer in domain.b. Super-peer in domain.b sends the desired Resource-ID to the peer in domain.b. Once this flow finishes, a legacy SIP negotiation is initiated for a VoIP call. We must highlight that Figure 3 represents only a subset of the real flow where the intermediate hops in each overlay are not shown.

Figure 3 Hierarchical P2PSIP signalling (see online version for colours)


domain.a
Interconnection Overlay
domain.b
Superpeers
Store(hash(domain.a), IP, Port,hash$_a$)
Response
Store(hash$_a$(user1@domain.a), IP, Port)
Response
Fetch(user1@domain.a)
Fetch(hash(domain.a)
Response(hash(domain.a), IP, Port, hash$_a$)
Fetch(hash$_a$(user1@domain.a))
Fetch(hash$_a$(user1@domain.a))
Response(hash$_a$(user1@domain.a), IP, Port)
Response(hash$_a$(user1@domain.a), IP, Port)
Response(user1@domain.a, IP, Port)
Legacy SIP exchange: INVITE, ...

2.5 Main characteristics of the hierarchical DHT architecture

Our proposal has several advantages. First, the operations or primitives of the used DHT are not changed. Only some modifications are needed in the maintenance operations to include the selection and update of super-peers (Min et al., 2006). One important advantage is that the routing state in peers does not increase, although much more peers are reachable through the interconnection overlay and the super-peers that are supporting it.

If a global domain would be used to provide global connectivity, this domain would contain all peers in the different domains. If each domain has M peers and K domains want to obtain global connectivity, the number of peers is $N = M \cdot K$. In many DHT-based overlay networks the routing state has a logarithmic dependency with the number of peers. Therefore, we have $O(\log_B N)$ because of the logarithmic property: $O(\log_B M) + O(\log_B K) \sim O(\log_B(M \cdot K))$. This routing state applies only to super-peers, while regular peers only have to maintain the state of their own domain, which is only $O(\log_B M)$.

The drawback is the higher load needed to be supported by the super-peers (Beverly Yang and Garcia-Molina, 2003) although this load is smaller than in other hierarchical DHT proposals (Ganesan et al., 2004; Garces-Erice et al., 2003; Xu et al., 2003; Zoels et al., 2006).

3 Routing performance

This section presents the RP in a system based on hierarchical DHT overlay network with the proposed *hierarchical ID*. This analysis is an extension of the work in Martinez-Yelmo et al. (2008) and a more elaborated model is presented.

3.1 Terminology

In the following lines we list the definition of the parameters for the analytical model:

- K: The number of P2PSIP domains.
- M_k: The number of peers in a P2PSIP domain k.
- N: The number of peers from all the P2PSIP domains. In our case, it is considered that a peer cannot be attached to multiple P2PSIP domains, hence $N = \sum_{i=1}^{K} M_i$.
- S_k: The number of super-peers in a P2PSIP domain k.
- ρ_{ij}: The probability of launching a query from the P2PSIP domain i to the P2PSIP domain j.
- $C(x)$: The number of hops needed to find a super-peer in the interconnection overlay depending on the number of super-peers x. This value depends on the type of overlay used in the interconnection overlay.
- $D_k(x)$: The number of hops needed to find a peer in a flat overlay of type k as function of the number of peers x belonging to the P2PSIP domain.

We assume that all peers from a P2PSIP domain know their super-peers from the interconnection overlay. This assumption implies that only *one hop* is needed to reach the super-peer. The RP inside a P2PSIP domain does not change and is the same as in a flat overlay network.

However, if a query must be routed to other domain, it would be only one hop away to any of its super-peers. The worst case happens when all the super-peers of a domain are attached to the interconnection overlay. Since the number of attached super-peers increases, the number of hops to search a resource in the interconnection overlay also increases. Nevertheless, this increment is small: between one and three hops, depending on the number of super-peers per domain and the overlay used for the interconnection overlay.

Taking into account the above definitions, we obtain the RP of this DHT-based hierarchical overlay networks. First of all, we define the cost of finding a peer in each overlay:

- $D_k(M_k)$: the cost of finding a peer in its own domain
- $C(\sum_{k=1}^{K} S_k)$: the cost of finding a super-peer in the interconnection overlay.

If the probability of obtaining an item in a domain from its super-peer is considered negligible and because the average number of peers in a P2PSIP domain is N/K with $N \gg K$, the average RP experienced by a peer in P2PSIP domain i can be written as follows:

$$\begin{aligned} RP_i &= \rho_{ii} \cdot D_i(M_i) \\ &\quad + \sum_{j=1, j\neq i}^{K} \rho_{ij} \cdot \left[1 + D_j(M_j) + C\left(\sum_{k=1}^{K} S_k\right)\right]. \end{aligned} \tag{1}$$

The first term of the sum is the cost of searching something in the P2PSIP domain of a peer, whereas the second term is the cost for the searches in the other P2PSIP domains.

The average number of hops is given by the next expression:

$$RP = \frac{1}{N} \cdot \sum_{i=1}^{K} M_i \cdot RP_i. \tag{2}$$

If the number of peers is the same in all P2PSIP domains, we have:

$$RP = \frac{1}{K} \cdot \sum_{i=1}^{K} \cdot RP_i. \tag{3}$$

3.2 Random independent queries

If we assume that the number of peers is equal in all P2PSIP domains and each look-up in the overlay is considered randomly independent, we obtain that the probability of looking for a peer attached to other P2PSIP domain is equally distributed among all the foreign P2PSIP domains. This means that $\rho_{ii} = \rho_{ij} = \frac{1}{K}$ and we obtain equation (4) from equation (1):

$$\begin{aligned} RP_i &= \frac{1}{K} \cdot D_i(M) \\ &\quad + \sum_{j=1, j\neq i}^{K} \cdot \frac{1}{K} \cdot \left[1 + D_j(M) + C\left(\sum_{k=1}^{K} S_k\right)\right]. \end{aligned} \tag{4}$$

Finally, if the same overlay is used in all P2PSIP domains the sum can be eliminated from equation (4) and RP_i becomes equal to RP:

$$\begin{aligned} RP_i = RP &= \frac{1}{K} \cdot D(M) \\ &\quad + \frac{K-1}{K} \cdot \left[1 + D(M) + C\left(\sum_{k=1}^{K} S_k\right)\right] \\ &= D(M) + \frac{K-1}{K} \cdot \left[1 + C\left(\sum_{k=1}^{K} S_k\right)\right]. \end{aligned} \tag{5}$$

3.3 Intra-domain queries more likely than inter-domain queries

However, the probability of looking for a peer in the own domain can be different from the one of looking for a peer in other P2PSIP domains. In this situation the inter-domain query probability is $\rho_{ij} = \frac{1-\rho_{ii}}{K-1}$ and we can express equation (1) as follows:

$$\begin{aligned} RP_i &= \rho_{ii} \cdot D_i(M) \\ &\quad + \sum_{j=1, j\neq i}^{K} \cdot \frac{1-\rho_{ii}}{K-1} \cdot \left[1 + D_j(M) + C\left(\sum_{k=1}^{K} S_k\right)\right]. \end{aligned} \tag{6}$$

This expression is useful for some types of scenarios like VoIP calls in community networks where $\rho_{ii} > \rho_{ij}$, meaning that calls between peers that belong to the same company or social network are more likely.

If the same overlay is used on all the P2PSIP domains the sum can be eliminated from equation (6) and RP_i becomes equal to RP:

$$\begin{aligned} RP_i = RP &= \rho_{ii} \cdot D(M) \\ &\quad + (1-\rho_{ii}) \cdot \left[1 + D(M) + C\left(\sum_{k=1}^{K} S_k\right)\right] \\ &= D(M) + (1-\rho_{ii}) \cdot \left[1 + C\left(\sum_{k=1}^{K} S_k\right)\right]. \end{aligned} \tag{7}$$

We define ρ_{ii} as the intra-domain hit probability and it defines the probability of establishing a connection inside the own domain.

4 Case study: hierarchical Kademlia

In this section, we study the RP and the routing state when a Kademlia overlay (Maymounkov and Mazieres, 2002) is used in all domains and in the interconnection overlay. We selected Kademlia because it is one of the most used DHT-based overlays in peer-to-peer applications like e-Mule, Bittorent, etc.

Kademlia is an overlay network that has a RP and a routing state with a logarithmic dependency on the number of peers from the overlay, due to its XOR distance-based routing algorithm.

4.1 Analytical analysis

In order to verify the efficiency of our solution, when the Kademlia protocol is used, we use the following equation: $C(x) = D(x) \sim \log_B x + c$, where B is a configuration parameter that allows to adjust the trade-off between the RP and the routing state in the peers.

In the case of random independent queries, we substitute this expression in equation (5). Therefore:

$$RP = RP_i \sim \log_B(M) + c + \frac{K-1}{K} \cdot [1 + \log_B(K) + c]. \tag{8}$$

If $K \gg 1$ and taking into account the properties of the logarithm, we can write:

$$RP = RP_i \sim 1 + \log_B(M \cdot K) + 2c. \tag{9}$$

On the other hand, for the case of intra-domain queries more likely that interdomain queries we use equation (7):

$$RP = RP_i \sim \log_B(M) + c + (1 - \rho_{ii}) \cdot [1 + \log_B(K) + c]. \tag{10}$$

If $\rho_{ii} \sim 1$, we have:

$$RP = RP_i \sim \log_B(M) + c. \tag{11}$$

For the routing state, the number of entries depends on the number of peers and on the setup parameter B. Actually, the number of overlay routing entries depends on $O(\log_B n)$ where n is the number of peers in the overlay. Super-peers have to support additional entries for the interconnection overlay. The total number of routing entries for a super-peer is approximately to $O(\log_B(K \cdot M))$.

If a flat overlay is used to connect all peers in different domains, peers would need $O(\log_B(K \cdot M))$ routing entries, but using the hierarchical architecture, legacy peers only need $O(\log_B M)$. Therefore, the routing state savings are significant if many domains are interconnected.

4.2 Validation via simulation

In this section, we present several experimental results with the goal to assess the performance of a hierarchical Kademlia overlay network. The results have been obtained with a prototype implementation of the protocol and using the PeerfactSim.KOM[3] P2P network simulator (Darlagiannis et al., 2004), which is a packet-level discrete event-based simulator written in Java. In order to facilitate the simulation of large scale peer-to-peer networks, the simulator uses a simple packet latency model between nodes that is the equivalent of the cumulative propagation, forwarding and queueing delay. However, it does not consider some details such as the processing time and the bandwidth of links (links are over-provisioned).

For the simulation results the corresponding 95% confidence intervals have been calculated. *These confidence intervals give an error less than 10%* which assures the consistency of the data obtained with the simulations.

4.2.1 Simulation setup

To run the experiments, we implemented a prototype of the hierarchical Kademlia protocol and a network scenario generator on top of the simulator engine. The objective was to generate peer-to-peer network models similar to the behaviour of real life Kademlia peers. For this we assumed network scenarios with an average number of peers between 400 and 10,000 and the following number of domains: 1 (i.e., a flat Kademlia overlay), 5, 10 and 20. The peers were uniformly distributed among the domains. In addition, each domain has a super-peer that facilitates the connection of the domains through the interconnection overlay. We limited the number of of super-peers per domain to one ($S_k = 1$) because the RP penalty is negligible (as has been explained in Section 3) and because using more than one super-peer increases greatly the complexity of the simulation. Additionally, the stability of super-peers can be assured as in Skype Rossi et al. (2008) with some mechanism like (Beverly Yang and Garcia-Molina, 2003; Min et al., 2006; Mizrak et al., 2003). The management of the super-peers is not included in the study and it constitutes future work. Thus, we do not consider churn in super-peers and only churn in peers in the manner we explain in the next paragraph.

Each peer executes four types of operations: joining when it attaches itself to the P2P overlay; storing a key-value pair; look-up when searching for a previously stored key in the attempt to find the value and leaving. In order to have scenarios closer to reality, we used an existing study of the KAD implementation of Kademlia (Steiner et al., 2007d) that measures the peer behaviour in terms of churn rate and up-time distributions. Their findings conclude that in a file-sharing KAD network peers arrive and leave with a negative binomial distribution, while the peer session time is similar to a Weibull distribution. Additional details can be found in Steiner et al. (2007a, 2007b, 2007c). This setup can be considered as a medium-high churn rate scenario since the KAD network is used in eMule and BitTorrent applications where the churn is not at all negligible. Thus, our scenario is a worse case study compared to the real situation that occurs in multimedia applications like Skype (Baset and Schulzrinne, 2006; Guha et al., 2006; Rossi et al., 2008).

Due to the simulation constraints (such as simulation duration, required computing resources, etc.) each simulation scenario has two phases. The first is a transitory phase, during which the total number of peers reaches the average targeted in each scenario. This phase does not consider the KAD peers behaviour, since in a real KAD network the arrival and the leaving rate are the same. In the second phase, the peers join and leave the P2P network at the rate given in Steiner et al. (2007d) with a negative binomial distribution (approximately one peer every two seconds). In this phase, the average number of peers in the network is the number of peers at the end of the first phase. Because the results from the KAD study were given for a flat Kademlia network, in the hierarchical case, arriving peers are randomly assigned to any of the existing domains with a uniform distribution.

During a session each peer performs a store that is the equivalent to storing its own URI in the P2P network, and a number of look-up operations that are the equivalent to searching for the URI of other peers. Assuming that the lookups follow the behaviour of the user contacting other peers, we used a Poisson distribution to model them, at an average rate of one call every ten minutes. The transitory first phase was limited to 30 minutes, while the stationary second state spanned up to two hours. As in Kademlia, a maintenance operation was run by each peer every hour after their arrival, in order to refresh their routing tables and republish stored values to neighbour peers. Measurements were taken only during the second phase.

In relation with the setup of the Kademlia overlay, the protocol has been configured with $B = 2^b = 2$, $k = 20$ and $\alpha = 1$. The reason for using $\alpha = 1$ is to facilitate the comparison with other overlays that cannot easily parallelise their operations. Determining the performance for higher values of α is planned as future work. The value of k is used for the size of the buckets and also for the number of replicas of each item inside the overlay.

4.2.2 Routing Performance

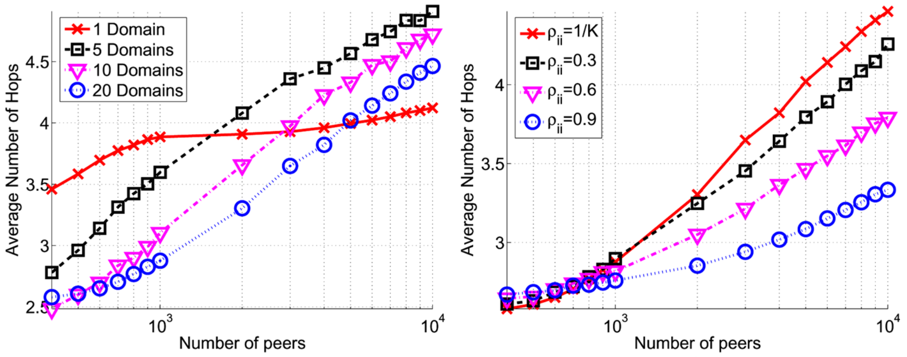
The RP was calculated for both *node look-up* and *value look-up* operations. The former are the result of the maintenance operations (refresh of the routing tables) and are performed solely inside the domain or inside the interconnection overlay between super peers. The latter are modelled based on peer behaviour of searching for stored values and can span two different domains. In addition, since the value look-ups take advantage of key-value replication, we expect the value look-ups to have a better performance. These operations finish as soon as a key is found. According to the analytical model and considering the assumptions on the simulation, the RP is estimated using the equations on Section 4.1.

Figure 4 illustrates the RP for value look-up operations. In Figure 4(a), we have the obtained RP for 1, 5, 10 and 20 domains. The dependency is logarithmic with the number of peers in a domain (linear on a logarithmic scale), when the number of super peers is kept constant. The difference between the values obtained for each number of domains represents the mean of the extra number of hops needed when the number of super-peers in the interconnection overlay increases. Because the increase is almost constant while the number of super peers doubles, the result proves the logarithmic dependency of the RP with the size of the interconnection overlay. The number of hops is bounded by equation (8), which is a constant since it only depends on N. The obtained results are smaller than the theoretical limit due to the replication of the information. Additionally, in Figure 4(b), we have the RP for 20 domains and ρ_{ii} equals to $\frac{1}{K}$, 0.3, 0.6 and 0.9. It can be appreciated how the RP increases as ρ_{ii} increases since the increment of ρ_{ii} makes larger the number of intra-domain look-ups. This difference is especially relevant when the number of peers is large, we have simulated up to ten thousand peers among all the domains and in a real scenario is expected that this number will be much more bigger.

Figure 5 shows the RP for node look-ups for intra-domain operations. Because of the design adopted for the hierarchical architecture, only node look-ups exist on intra-domain operations. They are important because node look-ups look for specific nodes and the performance is worse in comparison with value look-ups since they take advantage of replication. This difference is higher when the number of peers inside the domain is comparable to the replication parameter ($k = 20$) and becomes negligible when the number of domain peers is large enough for the replication to have an important effect. In Figure 5(a), we can see how the RP is smaller than the theoretical ($\log_2 M$ that is for the worst case). As expected, we find in Figure 5(b) that ρ_{ii} does not affect to RP of the node look-ups because is a parameter that defines how many queries are inter-domain and it only affects to value look-ups.

In order to see how good is the analytical analysis on estimating the upper bound of the RP we have obtained the worst case for the simulate value look-ups in Figure 6. We can appreciate in Figure 6(a) how the worst cases are close to the upper bound given in Section 4.1 and how

Figure 4 Routing performance for value look-up operations: (a) random independent queries with several domains and (b) 20 domains and $\rho_{ii} > \rho_{ij}$ (see online version for colours)



(a)

(b)

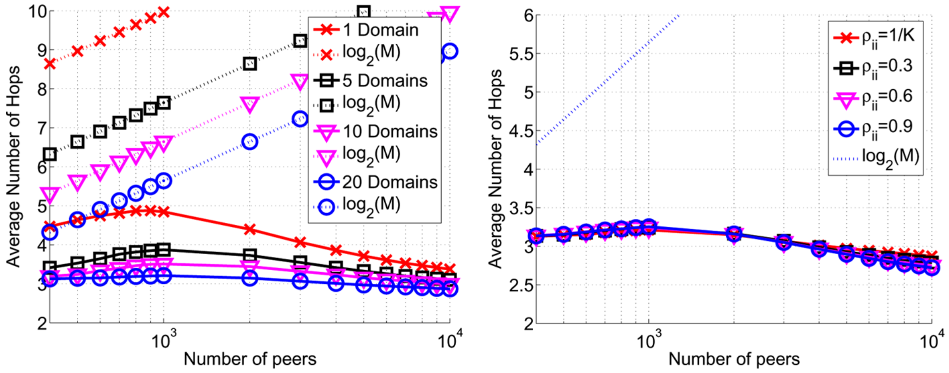
this bound has a logarithmic dependency. Furthermore, in Figure 6(b) we can see how this worse case is independent of ρ_{ii} because the worst case always depends on an inter-domain query.

4.2.3 Routing state

The evaluation of the routing state intends to determine whether the average number of routing entries maintained by the peers lay within the expected ranges and to illustrate the behaviour of the routing state when the number of domains changes. For this, we examine the routing tables used for routing inside domains.

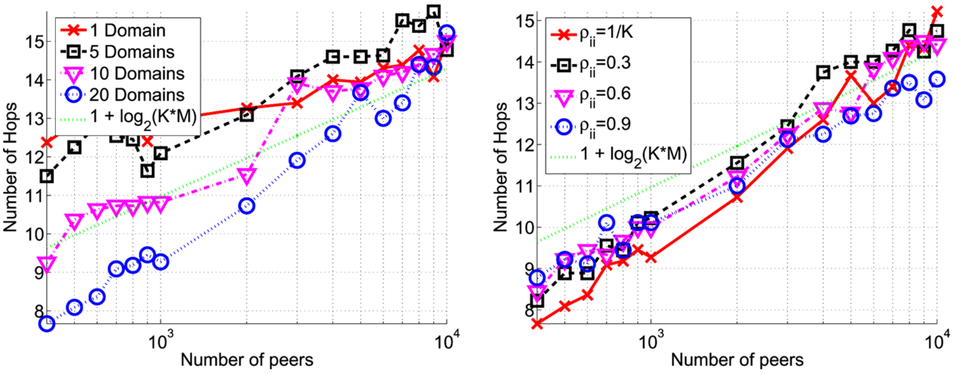
Figure 7 shows the obtained, i.e., $NE \in [\log_2 N, k\log_2 N]$, where NE is the average number of routing entries. In addition, we can observe a slight dependency between the number of domains and the value of the routing state in Figure 7(a). Since the routing state is determined solely by the interaction between peers, the explanation for this dependency is that the

Figure 5 Routing performance for node look-up inter-domain operations: (a) random independent queries with several domains and (b) 20 domains and $\rho_{ii} > \rho_{ij}$ (see online version for colours)



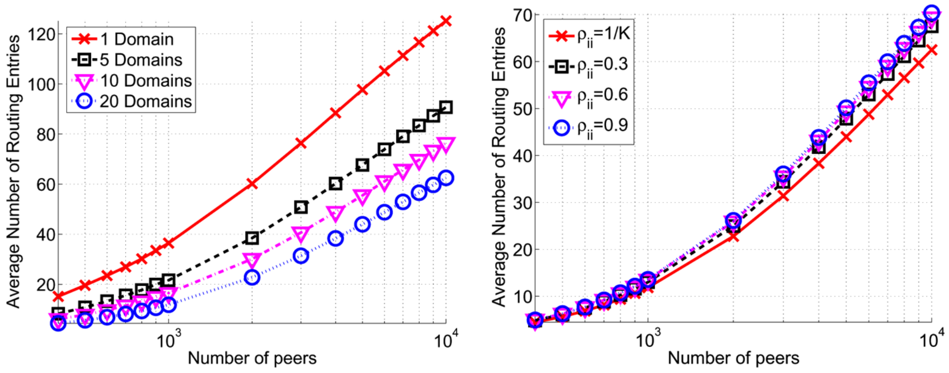
(a) (b)

Figure 6 The worst case of routing performance for value look-ups operations: (a) random independent queries with several domains and (b) 20 domains and $\rho_{ii} > \rho_{ij}$ (see online version for colours)



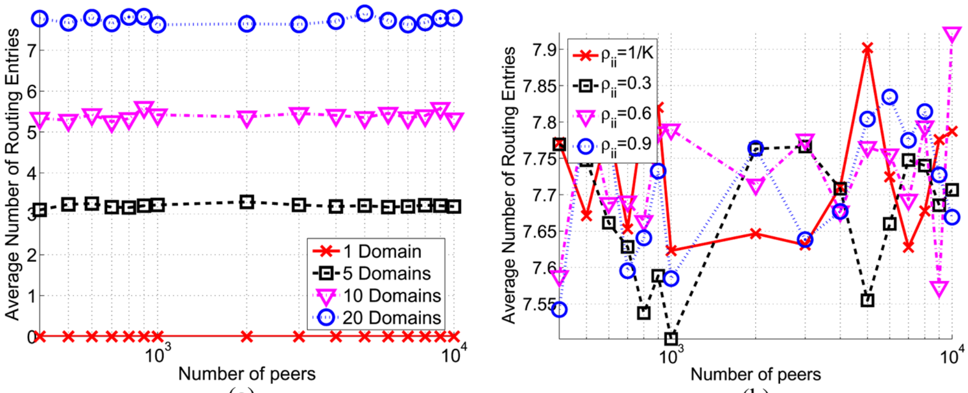
(a) (b)

Figure 7 Routing state for intra-domain routing tables: (a) random independent queries with several domains and (b) 20 domains and $\rho_{ii} > \rho_{ij}$ (see online version for colours)



(a) (b)

Figure 8 Routing state for interconnection overlay routing tables: (a) random independent queries with several domains and (b) 20 domains and $\rho_{ii} > \rho_{ij}$ (see online version for colours)



(a)

(b)

simulation scenarios use the same number of value look-up operations. In general, the value look-ups are originated in one domain and usually terminated in another domain. However, if the number of domains is small, the number of operations that originate and terminate in the same domain increases and consequently the number of routing entries also increases according to the standard Kademlia protocol mechanism to populate the bucket entries. Node look-ups cannot influence the routing state because they take place only inside a domain and have no relationship to the number of domains. We can see also how there is some dependency with ρ_{ii} in Figure 7. If ρ_{ii} is large, the intra-domain queries are more likely and the intra-domain routing tables are slightly more populated.

As expected, the number of hops needed in the interconnection overlay (see Figure 8(a)) is roughly the same for any number of peers, since it only depends on the number of domains, K. In addition, the logarithmic dependency with K can be observed through the large increase in the number of hops from one domain to five domains and the same increase between 5, 10 and 20 domains (the same difference when doubling the number of domains, hence a linear increase on a logarithmic scale). Furthermore, we can see how this value is independent with ρ_{ii} in Figure 8(b).

5 Related work

Overlay networks usually require $O(\log_B N)$ peer hops to reach the desired destination and $O(\log_B N)$ routing entries to maintain the desired structure. This complexity ensures good scalability but it is desirable to have further improvements. Thus, hierarchical overlay networks are being proposed because its benefits are clear (Kwon and Fahmy, 2005).

The first approach is to delegate all the work to super-peers (Garces-Erice et al., 2003; Zoels et al., 2006). They maintain the overlay network and perform all the necessary actions, while legacy peers only have to register their information to their super-peers. Other studies focus on optimising some parameter like the delay. Xu et al. (2003) propose a low delay hierarchical overlay network based on Chord. The drawback is the high routing state needed because *all* the peers in the overlay are attached to *all* the levels in a *n-level* hierarchy. A less aggressive design with the same objective is presented in Ganesan et al. (2004) but the hierarchy is built with the constraint of limiting the maintenance cost to the flat counterpart.

One of the main problems is the selection of super-peers. This selection can be based on the computation capacity of a peer, the available bandwidth to receive and process the queries and the session time to assure a stable set of super-peers (Min et al., 2006). Furthermore some mechanism must be provided to distribute the super-peer related information. One option is to piggyback this information (Joung and Wang, 2007).

6 Conclusions

The objective of the architecture proposed in this paper is to enable the interconnection of different domains in order to support global decentralised multimedia services.

Peers of the same domain are connected via a domain overlay network and by using the $\texttt{SuffixID} = \texttt{hash}(\texttt{user@example.com})$ to route queries. In order to get connectivity with other domains it is necessary to have at least one super-peer in each domain. An interconnection overlay is maintained between the super-peers where the routing is based on $\texttt{PrefixID} = \texttt{hash}(\texttt{example.com})$ values.

We obtain the average number of hops that a peer must perform in our architecture. Furthermore, we also show that the routing state of normal peers does not change and only super-peers are exposed to a higher load. These super-peers might actually be dedicated hosts used in the domain for this specific task. We also study the effect of intra-cluster hit probability ρ_{ii} on the routing performance. This analysis is important because the probability parameter can be used in VoIP scenarios such as a Skype-like service where we have different domains interconnected through the interconnection overlay.

Finally, we perform a simulation of a hierarchical Kademlia overlay network considering the churn for the peers according the values found (Steiner et al., 2007a, 2007b, 2007c, 2007d). The RP is below the value given by the analytical evaluation because of the replication in the information placed in each domain. Particularly, the information related to a specific user or service can be retrieved in 2.5–5 hops for a number of domains between 1 and 20 and with a number of peers per domain between 400 and 10,000. The average number of routing entries is reduced if the number of domains increases. These results illustrate the scalability of the solution.

Acknowledgements

We would like to acknowledge Roberto Gonzalez Sanchez for his collaboration in this work.

This work has been supported by the European Union under the IST Content NoE[4] (FP6-2006-IST-038423), by the Regional Government of Madrid under the BioGridNet[5] project (CAM, S-0505/TIC-0101) and by the Ministry of Science and Innovation under the CONPARTE project (MEC, TEC2007-67966-C03-03/TCM).

References

Baset, S.A. and Schulzrinne, H.G. (2006) 'An analysis of the skype peer-to-peer internet telephony protocol', *INFOCOM: Proceedings of the 25th IEEE International Conference on Computer Communications*, Barcelona, Spain, April, pp.1–11.

Beverly Yang, B. and Garcia-Molina, H. (2003) 'Designing a super-peer network', *Proceedings 19th International Conference on Data Engineering*, Bangalore, India, pp.49–60.

Bryan, D., Matthews, P., Shim, E. and Willis, D. (2007) *Concepts and Terminology for Peer to Peer Sip*, Internet Draft draft-ietf-p2psip-concepts-01.txt

Darlagiannis, V., Mauthe, A., Liebau, N. and Steinmetz, R. (2004) 'An adaptable, role-based simulator for P2P networks', *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods*, Las Vegas, USA, pp.52–59.

Ganesan, P., Gummadi, K. and Garcia-Molina, H. (2004) 'Canon in *g* major: designing dhts with hierarchical structure', *Proceedings of the 24th International Conference on Distributed Computing Systems*, Keio University, Japan, pp.263–272.

Garces-Erice, L., Biersack, E.W., Ross, K.W., Felber, P.A. and Urvoy-Keller, G. (2003) 'Hierarchical P2P systems', *Proceedings of ACM/IFIP International Conference on Parallel and Distributed Computing (Euro-Par)*, Klagenfurt, Austria, pp.1230–1239.

Guha, S., Daswani, N. and Jain, R. (2006) 'An experimental study of the skype peer-to-peer voip system', *The 5th International Workshop on Peer-to-Peer Systems*, Santa Barbara, USA.

Jennings, C., Lowekamp, B., Rescorla, E., Rosenberg, J., Baset, S. and Schulzrinne, H. (2008) *Resource Location and Discovery (Reload)*, Internet Draft draft-bryan-p2psip-reload-04.txt

Joung, Y-J. and Wang, J-C. (2007) 'Chord2: a two-layer chord for reducing maintenance overhead via heterogeneity', *Computer Networks*, Vol. 51, No. 3, pp.712–731.

Kwon, M. and Fahmy, S. (2005) 'Synergy: an overlay internetworking architecture', *ICCCN: Proceedings of the 14th International Conference on Computer Communications and Networks*, pp.401–406.

Martinez-Yelmo, I., Cuevas, R., Guerrero, C. and Mauthe, A. (2008) 'Routing performance in hierarchical DHT-based overlay networks', *Proceedings on 16th Euromicro International Conference on Parallel, Distributed and Network-based Processing*, Toulouse, France, pp.508–515.

Maymounkov, P. and Mazieres, D. (2002) *IPTPS 2002*, Cambridge, MA, USA, 7–8 March, Revised Papers, Vol. 2429/2002 of Lecture Notes in Computer Science, *Chapter Kademlia: A Peer-to-Peer Information System Based on the XOR Metric*, Springer, pp.53–65.

Min, S-H., Holliday, J. and Cho, D-S. (2006) 'Optimal super-peer selection for large-scale P2P system', *ICHIT'06: International Conference on Hybrid Information Technology*, Vol. 2. pp.588–593.

Mizrak, A.T., Cheng, Y., Kumar, V. and Savage, S. (2003) 'Structured superpeers: leveraging heterogeneity to provide constant-time lookup', *WIAPP: Proceedings of the Internet Applications*, pp.104–111.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M. and Schooler, E. (2002) *SIP: Session Initiation Protocol*, RFC 3261 (Proposed Standard), Updated by RFCs 3265, 3853, 4320, 4916.

Rossi, D., Melia, M. and Meo, M. (2008) 'A detailed measurment of skype network traffic', *IPTPS, The 7th International Workshop on Peer-to-Peer Systems*, Tampa Bay (Florida), USA.

Steiner, M., Biersack, E.W. and En Najjary, T. (2007a) 'Actively monitoring peers in KAD', *IPTPS'07: 6th International Workshop on Peer-to-Peer Systems*, 26–27 February, Bellevue, USA.

Steiner, M., En Najjary, T. and Biersack, E.W. (2007b) *Analyzing Peer Behavior in KAD*, Technical Report EURECOM+2358, Institut Eurecom, France.

Steiner, M., En-Najjary, T. and Biersack, E.W. (2007c) 'Exploiting KAD: possible uses and misuses', *SIGCOMM Comput. Commun. Rev.*, Vol. 37, No. 5, pp.65–70.

Steiner, M., En-Najjary, T. and Biersack, E.W. (2007d) 'A global view of KAD', *IMC'07: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, ACM, New York, NY, USA, pp.117–122.

Stoica, I., Morris, R., Liben-Nowell, D., Karger, D., Kaashoek, M., Dabek, F. and Balakrishnan, H. (2003) 'Chord: a scalable peer-to-peer lookup protocol for internet applications', *IEEE/ACM Transactions on Networking*, Vol. 11, No. 1, pp.17–32.

Xu, Z., Min, R. and Hu, Y. (2003) 'Hieras: a DHT based hierarchical P2P routing algorithm', *Proceedings of the International Conference on Parallel Processing*, Kaohsiung, Taiwan, pp.187–194.

Zoels, S., Despotovic, Z. and Kellerer, W. (2006) 'Cost-based analysis of hierarchical DHT design', *P2P: Sixth IEEE International Conference on Peer-to-Peer Computing*, Cambridge, UK, pp.233–239.

Notes

[1] http://www.skype.com

[2] http://www.p2psip.org

[3] http://peerfact.kom.e-technik.tu-darmstadt.de/

[4] http://www.ist-content.eu

[5] http://www.biogridnet.org