

Multicast in $\mathcal{DKS}(N, k, f)$ Overlay Networks *

Luc Onana Alima¹, Ali Ghodsi¹, Sameh El-Ansary², Per Brand² and Seif Haridi¹

¹MIT-Royal Institute of Technology, Kista, Sweden

²Swedish Institute of Computer Science, Kista, Sweden

{onana, ali, seif}@imit.kth.se, {sameh,perbrand}@sics.se

Abstract

In [1] a family of DHT-based infrastructures, termed $\mathcal{DKS}(N, k, f)$, with a number of desirable properties is presented. In the current paper, we show how multicast is achieved in $\mathcal{DKS}(N, k, f)$ overlay networks. Each multicast group is represented by an instance of $\mathcal{DKS}(N, k, f)$, which is created and maintained exactly as the underlying overlay network. Multicast messages are efficiently disseminated thanks to a correcting broadcast algorithm that allow each multicast message to be delivered exactly once to all application layer processes despite the presence of erroneous routing information.

1. Introduction

Recent developments in P2P computing show that DHT-based overlay networks scale well and can serve as infrastructures for P2P applications [1, 7, 4, 6]. To ease the development of large scale applications, such as media distribution, that require one-to-many communication, these infrastructures should provide multicast service [2, 5].

Multicast has been addressed according to two main approaches in the context of DHT-based infrastructures. In the first approach (see e.g. [2]), a *multicast distribution rooted tree* is maintained for each group, by some nodes of the underlying infrastructure. An advantage of this approach lies on the fact that the root of a multicast tree can be used for access control. The main drawback of this approach is the possibility of bottlenecks at the root node through which all multicast messages must pass. In the second approach (see e.g. [5]), group members self-organize into an overlay network similar to the underlying overlay network. This approach has the advantage that only group members (or bootstrap nodes like in [5]) maintain information about the group. Bottlenecks can only occur at bootstrap nodes and not because of multicasting itself. However, access control might be difficult in this case.

In this paper, we briefly present how multicast is achieved in $\mathcal{DKS}(N, k, f)$ overlay networks. Our design follows the second approach. The main advantage of our design over that of [5] is that each group is tailored to meet specific requirements regarding maximum group size, degree of fault-tolerance and maintenance

*This work was funded by the European project IST-2001-32234, PEPITO and Vinnova PPC project in Sweden.

cost. Members of a multicast group self-organize in an instance of $\mathcal{DKS}(N, k, f)$, which is created and maintained exactly as the underlying overlay network. Multicast messages are disseminated efficiently thanks to a correcting broadcast algorithm, which ensures that each application layer process receives each multicast message exactly once, despite the presence of erroneous routing information at the infrastructure level.

The rest of the paper is structured as follows. Section 2 discusses how groups are created and managed. Section 3 briefly describes how multicast messages are disseminated. Finally, Section 4 summarizes the paper.

2. Group creation and management

Let O be an instance of $\mathcal{DKS}(N, k, f)$, which we call the underlying overlay network. Let n be a node of O and assume that node n wants to create a new group identified by g . Then, node n first determines the characteristics of the group g . These include, (i) k_g : the arity parameter within the group. This parameter serves for the construction of the topology of the overlay network representing the group g . (ii) N_g : a power of k_g that denotes the maximum number of members that the group g can have. This number serves to determine the logical ring onto which group members are mapped. (iii) f_g : the fault-tolerance parameter within the group g . (iv) H_g : the hash function used to map nodes in O onto the logical ring of maximum size N_g . Then, once these parameters are set up, node n inserts itself as the first node of the $\mathcal{DKS}(N_g, k_g, f_g)$ representing group g . The characteristics of the group, g , and the address of node n are made available such that other nodes of O interested in g can join the group g .

The join, leave or failure of a group member is handled exactly as in the underlying overlay. Note however that due to the fact that the size of a group can be relatively small, a policy has to be decided to handle collisions. In our current design, we adopt the first-come-first-in policy.

The principle of the multicast group in $\mathcal{DKS}(N, k, f)$ overlay networks is illustrated in Figure 1. In this figure, *a*) shows a $\mathcal{DKS}(N = 32, k = 2, f = 3)$ instance, serving as the underlying overlay network, in which only nodes with identifiers 4, 11, 15, 19, 25 and 28 are present. *b*) shows a $\mathcal{DKS}(N = 16, k = 4, f = 2)$ representing a multicast group with identifier $g = 4$, in which nodes 28, 25, 19 and 15 of O are members. Note that because of the

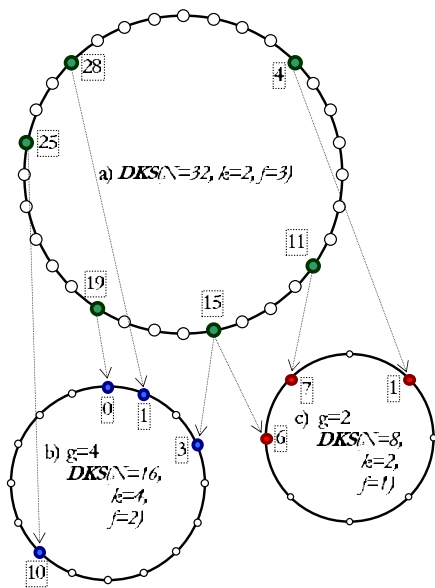


Figure 1: Two groups based on a $DKS(32, 2, 3)$.

hashing function for the multicast group $g = 4$, denoted H_4 , each of the nodes 28, 25, 19 and 15 receive a new identifier relative to this multicast group. For example, node 28 receives 1 as its identifier in the multicast group $g = 4$. c) shows a $DKS(N = 8, k = 2, f = 1)$ representing a multicast group with identifier $g = 2$ in which nodes 4, 11 and 15 of O are members.

3. Correcting broadcast

The main challenge in doing multicast in $DKS(N, k, f)$ overlay network is to achieve optimal forwarding of multicast messages. Due to the space limitation we only sketch the principle of the correcting broadcast. A full description can be found in [3].

In a $DKS(N, k, f)$ network, from a node n 's perspective, the identifier space (ring) is “observable” from $\log_k(N)$ levels. At each level, node n has a restricted view of the identifier space, which consists of k disjoint parts. Node n is itself responsible for one part and node n maintains information about the responsables of the other parts. This constitutes the routing information at node n .

Assuming that every node has correct routing information, the basic idea of the broadcast algorithm is as follows. When a node n wants to broadcast a message msg , node n sends msg to each responsible node, n' , piggybacking the interval, $I_{n'}$, for which node n considers n' responsible. Upon receipt of this message, node n' sends msg to each responsible it knows within the interval $I_{n'}$, piggybacking a sub-interval of $I_{n'}$. Because of the disjointness of the intervals, there is no redundant messages. This process is repeated until all nodes are covered.

The above sketched algorithm works well when every node has correct routing information. However, due to the dynamism of the overlay network, routing informa-

tion becomes outdated. In $DKS(N, k, f)$, this is tackled by the correction-on-use technique (see [1, 3] for details).

4. Concluding remarks and future work

We have briefly shown how multicast is achieved in $DKS(N, k, f)$ overlay networks. The main idea lies on the fact that for each group, a specific instance of $DKS(N, k, f)$ is created and maintained exactly as the underlying overlay network. Within a given group, each multicast message is delivered to all application processes exactly once despite erroneous routing information thanks to the correcting broadcast algorithm.

Acknowledgments

We would like to thank Mr. Thomas Sjöland for correcting some typos in the preliminary version.

References

- [1] Luc Onana Alima, Sameh El-Ansary, Per Brand, and Seif Haridi. $DKS(N, k, f)$: A Family of Low Communication, Scalable and Fault-Tolerant Infrastructures for P2P Applications. In *The 3rd International workshop on Global and Peer-To-Peer Computing on large scale distributed systems - CCGRID2003*, Tokyo, Japan, May 2003.
- [2] M. Castro, P. Druschel, A-M. Kermarrec, and A. Rowstron. SCRIBE: A large-scale and decentralised application-level multicast infrastructure. *IEEE Journal on Selected Areas in Communications (JSAC) (Special issue on Network Support for Multicast Communications)*, 2002.
- [3] Ali Ghodsi, Luc Onana Alima, Sameh El-Ansary, Per Brand, and Seif Haridi. Self-Correcting Broadcast in Structured P2P Networks. Technical Report TRITA-IMIT-LECS R 03:03, ISSN 1651-4661, ISRN KTH/IMIT/LECS/R-03/03-SE, KTH, March 2003.
- [4] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A Scalable Content Addressable Network. Technical Report TR-00-010, Berkeley, CA, 2000.
- [5] Sylvia Ratnasamy, Mark Handley, Richard Karp, and Scott Shenker. Application-level Multicast using Content-Addressable Networks. In *Third International Workshop on Networked Group Communication (NGC '01)*, 2001.
- [6] Antony Rowstron and Peter Druschel. Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. *Lecture Notes in Computer Science*, 2218, 2001.
- [7] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In *ACM SIGCOMM 2001*, pages 149–160, San Deigo, CA, August 2001.