2009 IEEE 34th Conference on Local Computer Networks (LCN 2009)
Zürich, Switzerland; 20-23 October 2009

BIDIR-SAM: Large-Scale Content Distribution in Structured Overlay Networks

Matthias Wählisch
Freie Universität Berlin
Institut für Informatik
Email: waehlisch@ieee.org

Thomas C. Schmidt
HAW Hamburg
Dept. Informatik
Email: t.schmidt@ieee.org

Georg Wittenburg
Freie Universität Berlin
Institut für Informatik
Email: wittenbu@inf.fu-berlin.de

***Abstract*—IPTV, software replication and other large-scale distribution tasks urge the need for efficient multicast mechanisms in overlay networks. Current multicast solutions on the application layer are either efficient, structured, but inflexible, or flexible, unstructured, but of lesser efficiency. This paper introduces Scalable Adaptive Multicast on BI-DIRectional shared trees, a new structured but flexible approach to content distribution. BIDIR-SAM is the first DHT-based overlay multicast that distributes any source multicast data according to source-specific shortest path trees. Built upon bi-directional shared prefix trees, the approach distributes packets uniquely via fully redundant paths, and allows for highly flexible network adaptivity. Guided by an overlay abstraction, it operates directly on top of a prefix routing and does not rely on any kind of rendezvous point or bootstrapping.**

I. Introduction

Large-scale content distribution is one of the most fascinating technical and economical challenges. Recently, IPTV and Video on Demand (VoD), software replication, and collaborative social platforms, e.g., video chats like Stickam experience rapid deployment. Nevertheless, there is a continuous search for mechanisms to spread content more efficiently, reliably and provider-friendly across a large number of recipients.

Originally, network layer multicast [1] has been designed to deliver data to multiple receivers. As providers remain hesitant to globally deploy native multicast, many ideas arose to disseminate data by replicating streams at end user devices or hybrid gateways [2], with overlay distribution facilitated by P2P protocols. Content replication on the overlay can reside on unstructured or structured P2P networks, i.e., DHTs.

Several debates rank around the performance of distributed hash tables (DHT) and their stability under churn. Current studies reveal that general objections do not hold and structured approaches clearly outperform the unstructured [3], [4]. More importantly for a future deployment, IETF/IRTF activities enforce to supplement Internet services by DHTs – cf., the P2PSIP and ALTO working groups, as well as P2P and SAM research groups. A generic peer-to-peer protocol supported by a mandatory DHT is currently designed [5]. Thus, it may be reasonable to assume that DHT substrates will populate the future Internet. With respect to these observations, we focus on and limit our discussion to structured overlays.

Current structured approaches to group distribution either implement a source-specific network flooding, a source-specific or a shared tree. Flooding was found to be outperformed by forwarding along trees [6]. Tree-based schemes uniquely deliver packets on paths that often are close to optimal. However, the tendency arose to discard trees from P2P multicast [7], mainly due to two reasons: a lack of robustness, as inner vertices may fail, and an unfair load distribution, as leaf nodes never replicate traffic.

Additionally, current shared tree approaches introduce dedicated rendezvous points, which concentrate traffic and may add jitter boosts [8]. A new content distribution approach should fulfill the following requirements: optimal, but variable traffic paths from the source to the receivers, robustness, balancing of fan-out among peers, and a predictable performance.

In this paper, we present Scalable Adaptive Content Distribution on Bi-directional Shared Trees (BIDIR-SAM) in large-scale structured overlay networks. BIDIR-SAM constructs one abstract source-specific bi-directional shared tree per group, with vertices mapped to groups of nodes. It enables an *arbitrary* overlay node to distribute data along *forward-oriented* shortest paths, without utilizing a rendezvous point or flooding. Replication load is balanced inherently fair among all receivers, making the approach suitable for media type broadcasting, as well. BIDIR-SAM operates directly on top of a DHT that uses a prefix-based routing scheme and may be proximity-aware. It thus inherits performance and error resilience directly from the structured overlay. Our scheme offers arbitrary redundancy for packets and paths based on dynamic multipath transport.

The remainder of this paper is organized as follows: In section II, we review related work. We present the core mechanism of our approach in section III. A discussion of the achievements in section IV concludes this paper.

II. Related Work

Derived from structured P2P routing, several group communication services have been developed with the aim of seamless deployability as application layer or overlay multicast. Among the most popular approaches are multicast on CAN [9], Bayeux [10] as derived from Tapestry and Scribe [11] or SplitStream [12] based on Pastry. These approaches essentially branch in two algorithmic directions. The first uses DHTs to generate a structured sub-overlay network of group members, which thereafter is flooded (CAN). The second class erects distribution trees. Identifying rendezvous points from group ID hashes, Scribe and SplitStream generate shared trees from reverse path forwarding, while Bayeux constructs shortest paths trees from source-specific client subscriptions. Performing receiver tracking at a source-centric group control, Bayeux exhibits linear growth in listener-state information. To the best of our knowledge, neither an any source multicast scheme is known that distributes data along shortest path trees, nor a structured overlay multicast that strictly adheres to logarithmically scalable costs.

Multicast tree properties comparing structured and unstructured schemes [13] have been explored in [8]. Focusing on Scribe and SplitStream, the authors identified a highly unbalanced forwarding load at inner tree nodes along with large fluctuations in delay. The latter were found to accumulate in SplitStream to mainly intolerable jitter values. In addition, conventional trees composed of fixed DHT nodes are highly vulnerable to failures along the paths. A single drop-out in the center of a spanning tree amplifies damage in the system.

In summary, tree-based approaches bear the advantage of unique and efficient packet transmission. Their main drawback lies in lack of redundant paths, as an inner vertex equals exactly one overlay peer, and largely unbalancing replication load among nodes. In the following, we present BIDIR-SAM, which solves these issues by constructing a *single, abstract* tree, on with each vertex is represented by a set of receivers. This multi-path structure does not rely on a group controller, but exhibits scalable, balanced load at individual peers.

III. BIDIR-SAM

BIDIR-SAM is full multicast solution including group membership management. It operates on the prefix structure in the key space of an overlay. Packet forwarding in BIDIR-SAM follows a virtual distribution tree in prefix space, whose vertices are dynamically mapped to overlay nodes by the underlying DHT. Branching nodes duplicate packets as in topological trees, but the variable, late binding of vertices to physical nodes may lead to different mappings according to DHT node selection. The structured overlay is provided by a key-based routing (KBR) as implemented in DHTs like Pastry. At first, we draw on the concept of prefix directed routing that will be used further on to resolve the virtual tree.

A. Prefix-Directed Forwarding

Prefix-directed routing is used by several DHTs such as Pastry and gives rise to a high degree of flexibility. Each peer maintains a set of prefixes that includes a prefix of each available peer (for example, the routing table in Pastry). A prefix represents all nodes with an ID sharing it. In general, routing towards a prefix does not directly address a specific peer, but a set of potential nodes. Only at the forwarding decision, the prefix will be resolved to a specific destination. Consequently, any tree structure based on prefixes may adapt to load, and is shielded from volatile peers, as long as the binding of prefixes to peers will remain at the late step of forwarding.

B. The Core Protocol

The BIDIR-SAM protocol constructs its virtual distribution tree on the prefix structure of multicast members in the overlay. Overlay IDs are created using an alphabet of k digits. Data then will flow to group members by forwarding along this prefix tree, which proceeds by mapping prefixes to nodes as visualized in figure 1. Before discussing the general multicast, we first want to recall the simpler case of broadcasting.

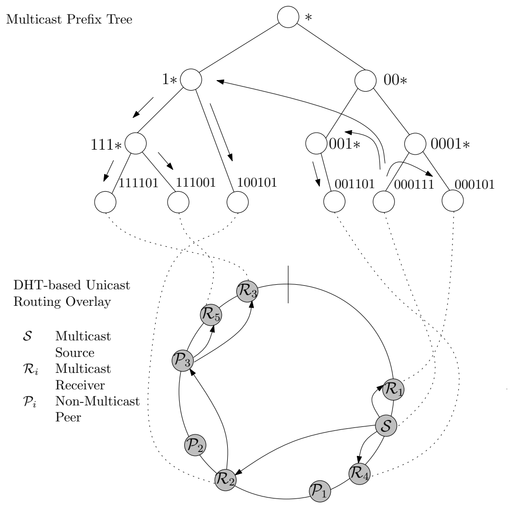


Fig. 1. BIDIR-SAM: A source sends data to prefixes that cover receivers. Prefix-directed routing forwards to nodes that represent the selected prefixes.

a) Broadcast on a Prefix Tree: A prefix tree covering all DHT members can be immediately derived by identifying leaves as overlay IDs of *all* DHTs members and labeling inner vertices recursively with the longest common prefix (LCP) of their children.

For sending a packet from the root to all leaves of the broadcast tree, each peer needs to decide on packet replication according to its current branching position on the tree. This context awareness can be gained from adding a destination prefix $\mathcal{C}$ to the packets, which will be hopwise updated with growing length. Downward forwarding is then simply achieved by routing to all neighboring prefixes that share $\mathcal{C}$. This mechanism, called PREFIX FLOODING [14], [15], can be applied at any level of the tree structure and does not require explicit group management.

b) Multicast Group Management: The BIDIR-SAM distribution tree is constructed as follows: Identify leaves as IDs of group listeners instead of all overlay nodes as shown in Fig. 3. This member-based tree must be generated by a specific group signaling that allows for proper prefix selection.

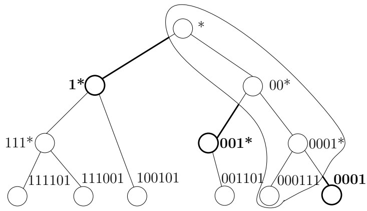
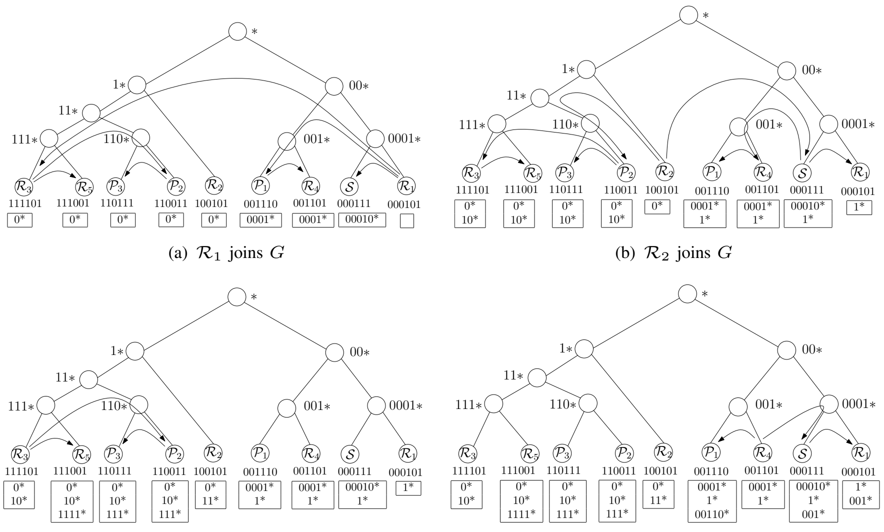


Fig. 3. Node Embedding in a Prefix Tree

To learn about a multicast group structure, a peer does not need to memorize the entire group-specific multicast tree, but will only be required to store the neighboring prefix labels (highlighted bold in Fig. 3) of all inner vertices that share its label. For each multicast group G, a BIDIR-SAM peer K individually maintains a (decentralized) multicast forwarding table MFT_G. This list contains all prefixes from the distribution tree, which serve destinations adjacent to K. The first and last receiver of the group flood their join/leave message in the complete (unicast) overlay network. For all further group members, the state update is propagated according



(a) $\mathcal{R}_1$ joins G

(b) $\mathcal{R}_2$ joins G

(c) $\mathcal{R}_3$ joins G

(d) $\mathcal{R}_4$ joins G

Fig. 2. Consecutive receivers join group G. BIDIR-SAM join procedure shows signaling flow (arrows) and evolving multicast forwarding table per peer

to MFT_G-entries at a peer K. Signaling occurs only within the smallest subtree covering the new multicast listener. Its root equals the LCP of K and the longest prefix in MFT_G:

```
BIDIR-SAM JOIN/LEAVE INJECTION
    ▷ Invoking this function at peer with ID $\mathcal{K}$ for group $G$
if $MFT_G = \emptyset$
    then PREFIX FLOODING Join/LeaveMessage To *
    else Select $\mathcal{L} \in MFT_G : |\mathcal{L}| \geq |\mathcal{L}'|, \forall \mathcal{L}' \in MFT_G$
            ▷ Create root of subtree to flood
         $\mathcal{C} \leftarrow LCP(\mathcal{L}, \mathcal{K})$
         PREFIX FLOODING Join/LeaveMessage To $\mathcal{C}$
```

On the reception of a state update, the following function will be called to include or delete multicast forwarding entries and to route the message down the unicast prefix tree:

```
BIDIR-SAM JOIN/LEAVE PROCESSING
    ▷ Denote the prefix of length $l$ of a key $\mathcal{A}$ by $prefix(l, \mathcal{A})$
    ▷ On arrival of message $m$ for group $G$ from peer $\mathcal{P}$ at node $\mathcal{K}$
$\mathcal{L} \leftarrow LCP(\mathcal{P}, \mathcal{K})$
$\mathcal{L}' \leftarrow prefix(|\mathcal{L}| + 1, \mathcal{P})$
if type($m$) = LEAVE
    then $MFT_G \leftarrow MFT_G \setminus \mathcal{L}'$
elseif type($m$) = JOIN
    then $MFT_G \leftarrow MFT_G \cup \mathcal{L}'$
PREFIX FLOODING $m$ TO $\mathcal{L}$
```

Based on this group membership, a BIDIR-SAM peer constructs a bi-directional shared tree covering all overlay multicast listeners (see [14] for a formal proof). The join procedure is illustrated in Fig. 2 for four receivers. The visualized prefix tree represents the (unicast) routing overlay, which is used to distribute multicast signaling.

c) Data Distribution: Each node controls an individual multicast forwarding table MFT_G that holds all neighboring prefixes that cover receivers. Thereby an arbitrary peer can act as multicast source by issuing data to all entries in MFT_G. The packets will then be forwarded to the leaves of the multicast tree as follows:

```
BIDIR-SAM FORWARDING
    ▷ On arrival of packet with destination prefix $\mathcal{C}$
    ▷ for group $G$ at DHT node of ID $\mathcal{K}$
for all $\mathcal{N}_i$ IDs in $MFT_G$
    do if $LCP(\mathcal{C}, \mathcal{N}_i) = \mathcal{C}$
          ▷ $\mathcal{N}_i$ is downtree neighbor
        then $\mathcal{C}_{new} \leftarrow \mathcal{N}_i$
             FORWARD PACKET TO $\mathcal{C}_{new}$
```

Data is sent to all roots of subtrees as extracted of the multicast forwarding table, and distributed therein with growing destination prefix. Thus, all multicast listeners receive the data exactly once and the algorithm terminates (for a proof cf. [14]).

It is worth noting that error resilience directly follows from the prefix structure. In contrast to other scalable approaches around, inner vertices of the tree are represented by prefixes, *not* individual nodes. Thus a volatile peer does not change the tree structure, but only changes the underlying DHT (unicast) routing table. Consequently, BIDIR-SAM is just as stable as the DHT. An arbitrary redundancy can be achieved by sending data in parallel to different prefixes.

C. Protocol Extensions

The core protocol maintains a generic shared family of source trees in prefix space, which allow for unique multicast data transmission from any node in a prefix-optimized fashion. It is open to additional features as desired by the application or network scenario.

As all peers in a BIDIR-SAM overlay multicast are equally suited to serve as a content root for a given group, neighboring peers may serve as relays. Thereby it offers *fault-tolerant routing, arbitrary redundancy for packets and paths* and remains *mobility agnostic* in the sense that mobile senders can seamlessly transmit multicast data from any location, while listeners may need to activate prefix branches for distribution. Furthermore, it facilitates *dynamic multipath transport* without effort and may give rise to end-to-end resource pooling in multicast, thereby filling the gap left in [16]. These improvements apply without additional signaling or management overhead.

IV. Discussion & Conclusion

In this paper, we have presented BIDIR-SAM, a novel overlay multicast approach, which enables any peer to distribute multicast data directly into a multicast group. Using a logical prefix overlay, BIDIR-SAM peers autonomously construct a bi-directional, shared distribution tree, which disseminates data according to source-specific shortest paths. There is no need for dedicated, infrastructure entities such as rendezvous point among the overlay nodes.

The protocol has been thoroughly analyzed based on its analytically available properties, as well as by extensive simulations. For evaluation results, which are omitted here due to space limitations, we refer to [14]. Our analysis revealed that the protocol costs in signaling and forwarding are strictly predictable and scale logarithmically with the network and group size. Thereby, and to the best of our knowledge, BIDIR-SAM is the only structured multicast scheme, which distributes data on source-specific shortest path trees at logarithmic costs, and as well the only solution, which utilizes shortest path trees within a shared tree model.

Our analysis, which draws detailed comparison with Scribe, further reveals that BIDIR-SAM packet distribution metrics and overall resource requirements evolve evenly on logarithmic scales, while most performance values of Scribe fluctuate on a scale linearly growing with group members. In particular, the following discussions grant stimulating insights.

1) Redundancy and Reliability: The common major weakness of tree-based, structured multicast approaches debated in literature and practice lies in a limited reliability and an increased vulnerability with respect to node or link failures. It is one of the major strength of BIDIR-SAM to overcome this deficit. On the price of an enhanced initial signaling load, a distribution of relevant multicast state information is achieved among all peers. In particular, each peer is enabled to initiate or perpetuate the distribution of any packet it receives. This inherent redundancy allows for a drop out of nodes, without affecting multicast forwarding. On the contrary, only a few selected Scribe peers hold forwarding entries. If they disappear or suffer from disturbances, the multicast distribution tree collapses. The reliability of BIDIR-SAM is enforced by storing only prefixes instead of node keys, where each prefix covers a group of interchangeable overlay nodes.

2) The Problem of Asymmetric Routes: Overall measures on routing performance raise the question about more fundamental reasons why BIDIR-SAM consistently outperforms Scribe. Leaving aside the rendezvous point issues, the main conceptual difference between data-driven tree approaches and the BIDIR-SAM follows from the method of tree establishment. In general, data-driven trees will be constructed from reverse path forwarding. The tree is optimal, as long as the routing table entries are invertible. But if links between nodes admit asymmetrical weights, a source may deliver data along suboptimal paths. This is particularly important in DHTs, which construct forwarding paths according to the specific, limited views at intermediate nodes. Routes commonly are highly asymmetric, which leads to suboptimal, early branching as Scribe exhibits close to its rendezvous point. Such problems do not arise, if the source constructs its tree according to forward routes as in BIDIR-SAM.

Optimized tree construction and data transmission throughout the underlay are key controls for efficient group communication in DHTs. This work has identified that reverse path selection in overlay and underlay turns into a severe problem in the presence of asymmetric routing. Asymmetric routing paths are also a problem for native group communication, because common multicast routing is based on data-driven trees. Establishing forward paths in the Internet is not as easy as it is in DHTs due to scaling issues. BIDIR-SAM changes the paradigm of data-driven trees to source-driven distribution: Each source represents the root of an implicitly defined distribution tree under appropriate performance values.

References

[1] S. Deering, “Host extensions for IP multicasting,” IETF, RFC 1112.

[2] M. Wählisch and T. C. Schmidt, “Between Underlay and Overlay: On Deployable, Efficient, Mobility-agnostic Group Communication Services,” *Internet Research*, vol. 17, no. 5, pp. 519–534, Nov. 2007.

[3] M. Castro, M. Costa, and A. Rowstron, “Debunking some myths about structured and unstructured overlays,” in *Proc. of NSDI’05*. Berkeley, CA, USA: USENIX Association, 2005, pp. 85–98.

[4] Y. Qiao and F. E. Bustamante, “Structured and unstructured overlays under the microscope: A measurement-based view of two P2P systems that people use,” in *Proc. of USENIX’06*. Berkeley, 2006, pp. 341–355.

[5] C. Jennings, B. Lowekamp, E. Rescorla, S. Baset, and H. Schulzrinne, “REsource LOcation And Discovery (RELOAD) Base Protocol,” IETF, Internet Draft – work in progress 01, Dec. 2008.

[6] M. Castro, M. B. Jones, A.-M. Kermarrec, A. Rowstron, M. Theimer, H. Wang, and A. Wolman, “An Evaluation of Scalable Application-level Multicast Built Using Peer-to-peer Overlays,” in *Proc. of the 22nd IEEE Infocom 2003*, vol. 2. IEEE ComSoc, 2003, pp. 1510–1520.

[7] V. Pai, K. Kumar, K. Tamilmani, V. Sambamurthy, and A. E. Mohr, “Chainsaw: Eliminating Trees from Overlay Multicast,” in *Proc. of IPTPS’05*, ser. LNCS, M. Castro and R. van Renesse, Eds. Berlin Heidelberg: Springer–Verlag, 2005, vol. 3640, pp. 127–140.

[8] S. Birrer and F. E. Bustamante, “The Feasibility of DHT-based Streaming Multicast,” in *Proc. of MASCOTS ’05*.IEEE ComSoc, 2005.

[9] S. Ratnasamy, M. Handley, R. M. Karp, and S. Shenker, “Application-Level Multicast Using Content-Addressable Networks,” in *Proc. of NGC 2001*, ser. LNCS, vol. 2233. London: Springer–Verlag, 2001, pp. 14–29.

[10] S. Q. Zhuang, B. Y. Zhao, A. D. Joseph, R. H. Katz, and J. D. Kubiatowicz, “Bayeux: An Architecture for Scalable and Fault-tolerant Wide-area Data Dissemination,” in *Proc. of NOSSDAV’01*, J. Nieh and H. Schulzrinne, Eds. New York, NY, USA: ACM, 2001, pp. 11–20.

[11] M. Castro, P. Druschel, A.-M. Kermarrec, and A. Rowstron, “SCRIBE: A large-scale and decentralized application-level multicast infrastructure,” *IEEE JSAC*, vol. 20, no. 8, pp. 100–110, 2002.

[12] M. Castro, P. Druschel, A.-M. Kermarrec, A. Nandi, A. I. T. Rowstron, and A. Singh, “SplitStream: High-Bandwidth Content Distribution in Cooperative Environments,” in *IPTPS 2003 Berkeley, CA, USA*, ser. LNCS, vol. 2735. Springer–Verlag, 2003, pp. 292–303.

[13] M. Hosseini, D. T. Ahmed, S. Shirmohammadi, and N. D. Georganas, “A Survey of Application-Layer Multicast Protocols,” *IEEE Communications Surveys & Tutorials*, vol. 9, no. 3, pp. 58–74, 2007.

[14] M. Wählisch, “Scalable Adaptive Group Communication on Bi-directional Shared Prefix Trees,” Freie Universität Berlin, Dept. of Math. and CS., Berlin, Tech. Rep. TR-B-08-14, September 2008. [Online]. Available: http://www.inf.fu-berlin.de/inst/pubs/tr-b-08-14.abstract.html

[15] M. Wählisch, T. C. Schmidt, and G. Wittenburg, “Broadcasting in Prefix Space: P2P Data Dissemination with Predictable Performance,” in *Proc. of ICIW’09*, IEEE ComSoc, May 2009, pp. 74–83.

[16] D. Wischik, M. Handley, and M. B. Braun, “The Resource Pooling Principle,” *SIGCOMM CCR*, vol. 38, no. 5, pp. 47–52, Oct. 2008.