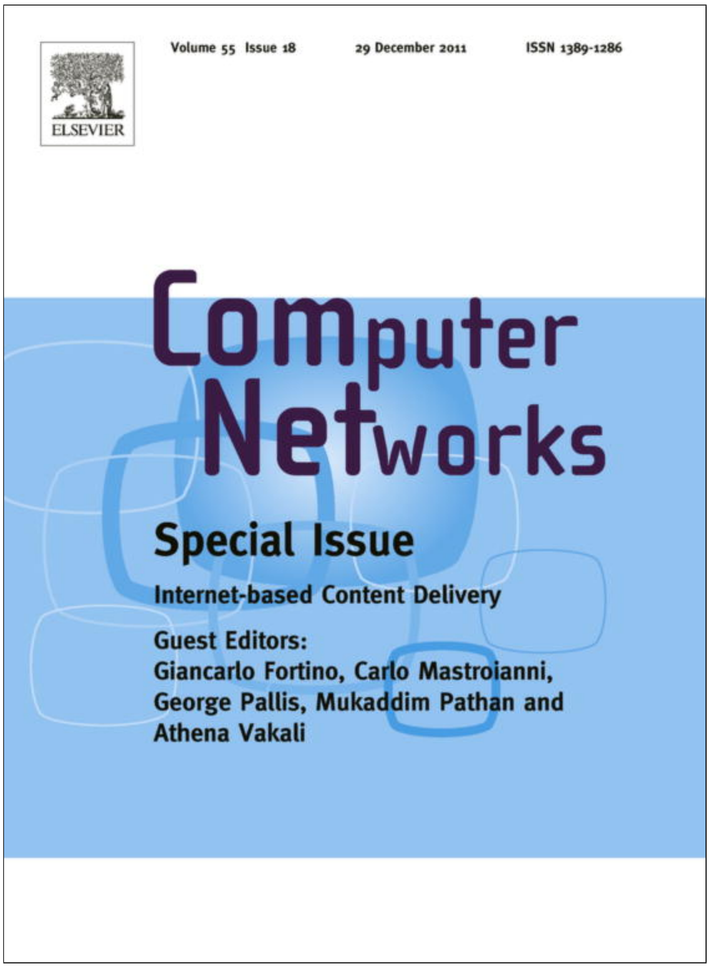


Volume 55 Issue 18 29 December 2011 ISSN 1389-1286

ELSEVIER

Computer Networks

Special Issue

Internet-based Content Delivery

Guest Editors:
Giancarlo Fortino, Carlo Mastroianni,
George Pallis, Mukaddim Pathan and
Athena Vakali

Computer Networks 55 (2011) 4086–4100

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

On predictable large-scale data delivery in prefix-based virtualized content networks

Matthias Wählisch [a,*], Thomas C. Schmidt [b], Georg Wittenburg [c]

[a] Freie Universität Berlin, Institut für Informatik, Takustr. 9, 14195 Berlin, Germany
[b] HAW Hamburg, Dept. Informatik, Berliner Tor 7, 20099 Hamburg, Germany
[c] INRIA / École Polytechnique, Laboratoire d'Informatique de l'École Polytechnique, Route de Saclay, 91128 Palaiseau (CEDEX), France

ARTICLE INFO

Article history:
Received 16 December 2010
Received in revised form 23 June 2011
Accepted 8 July 2011
Available online 5 August 2011

Keywords:
Content delivery and management on the Internet
Dynamic content delivery and adaptive CDNs
P2P networks

ABSTRACT

IPTV, software replication, and other large scale content distribution services raise the need for fast and efficient content delivery mechanisms in underlay as well as overlay networks. Multicast, the natural approach on the network layer, has not been deployed globally, and solutions are pushed to the application layer. For a flexible, sustainable deployment the distribution mechanisms in use should scale up to many thousand group members and provide predictable performance to dynamically adjust to actual performance requirements. In this paper, we present a rigorous analytical model complemented by extensive simulations for content delivery on prefix-based overlay trees. We examine BIDIR-SAM, a generic multicast distribution scheme guided by prefixes that allow for late next-hop binding. Our evaluation quantitatively substantiates all major performance aspects. We derive the distribution functions of hop counts, packet replication loads, as well as all relevant cost measures, which scale logarithmically with network and receiver sizes. Prefix-based content delivery exhibits a churn resistance similar to the underlying key-based routing layer and a multicast efficiency scaling factor close to native group communication protocols. These results make the approach especially suitable for large and very large content groups.

1. Introduction

Large-scale content distribution is a distinguished technical and economical challenge of many facets. IPTV, Video on Demand (VoD), or collaborative social platforms like video chats (e.g., Stickam) place real-time demands of low latency and jitter on the distribution system, while file sharing, Web caching and software replication rely more on efficient, reliable data dissemination and retrieval. Conjointly all content distribution systems need to be highly scalable and should organize data transmission in a provider-friendly way. For any such system, a strong advantage must be seen in a behaviour that is predictable in detail and based on rigorous performance measures. Following this insight, we present a thorough performance analysis with analytically rigorous results and large-scale simulations for the group distribution problem.

Originally, network layer multicast [1] has been designed to deliver data on shortest paths to an unlimited number of receivers. IP multicast would be the most efficient solution for distributing data to content replicas, middlebox caches or end systems. However, providers remain hesitant to globally deploy native multicast, and it thus does not find its role as an inter-domain content delivery service. Overlay distribution schemes, which have been designed and deployed in manifold ways, neither meet the transmission quality nor the routing efficiency of IP multicast. Hence the search continues for mechanisms to spread content faster, more efficiently, reliably, and provider-friendly across a large number of recipients.

[*] Corresponding author. Tel.: +49 30 838 75209.
E-mail addresses: waehlisch@ieee.org (M. Wählisch), t.schmidt@ieee.org (T.C. Schmidt), georg.wittenburg@inria.fr (G. Wittenburg).

1389-1286/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2011.07.020

In recent years, research and development has diverged to address application-centric solutions separate from infrastructure-assisted approaches. In typical deployment scenarios, application layer multicast [2] is optimized for network access constraints, i.e., upload capacity and NAT traversal, whereas content overlay infrastructures follow a server-centric paradigm. Corresponding content delivery of group data within the Internet is traditionally deployed by content delivery networks (CDNs) that replicate data of a well-defined content source towards servers close to the end-users. Content ISPs and end-users attached to eyeball ISPs remain decoupled. To deliver data efficiently between end-users without dedicated support on the network layer, peer-to-peer (P2P) networks have been introduced. Overlays created by P2P networks spawn a virtual content network.

In this work, we argue that content networking should not focus on either content providers or end-users, but consider content delivery as a joint process. Any data replication in the vicinity of receivers increases efficiency. Several novel concepts designed for a joint deployment of network and content providers as well as end-users have been designed, among them nano data centers [3] with replicators at controlled peers (e.g., set top boxes), hybrid Multicast [4–7] based on gateways, or a recent IETF initiative on Content Distribution Network Interconnection (CDNI) [8]. Still keeping in mind the demands of ISPs, which require a valid estimation of resources, and the size of end-users domains, content delivery as well as management in the Internet must be predictable and scalable for the majority of users.

Content distribution on the overlay can proceed via data-driven (mesh-based) approaches that pull content to maximize the utilization of upload capacities at clients at the cost of enhanced control overhead and delay [9,10]. Alternatively, data is pushed down distribution trees, which minimize overhead and delay, but show reduced robustness when inner nodes fail and do not exploit the upload capacities of leaves. Consequently, data-driven solutions face wider deployment at end nodes, while tree-based systems are better adopted in infrastructure-assisted or real-time services. This work examines distribution systems with predictable high-quality performance that, in particular, are suitable for real-time data and focusses on trees.

Trees may reside on structured or unstructured P2P networks. Several debates revolve around DHT performance and about their scalable stability under churn. Current studies reveal that general objections do not hold and structured approaches clearly outperform the unstructured [11,12]. More importantly for its ubiquitous presence in the future, IETF/IRTF activities enforce supplementing Internet services by DHTs, cf. the recently re-chartered P2P RG, as well as the P2PSIP and ALTO working groups. Reload [13], the proposed generic peer-to-peer protocol, will include mandatory support for a DHT. Thus, it is reasonable to assume that DHT substrates will populate the future Internet. With respect to these observations, we focus our discussion on structured overlays.

In this paper, we present an analytical model complemented by extensive simulations for content delivery on prefix-based overlay trees. We examine BIDIR-SAM [14], a generic multicast distribution scheme built as an abstract prefix tree whose vertices are mapped to specific peers only at packet forwarding. In contrast to common (reverse path) routing trees, the use of prefix-directed forwarding allows for a late next-hop binding at runtime and thereby eliminates the stability breach common to trees caused by a departure of inner nodes. Data forwarding on abstract prefix trees attains the following additional advantages: It allows for source-specific bi-directional shared trees, which enable an arbitrary overlay node to distribute data along forward-oriented shortest paths. It does not rely on any kind of rendezvous point, but offers fault-tolerant routing, arbitrary redundancy for packets and paths, and remains mobility-agnostic.

The key contributions of this paper are as follows: We rigorously show that prefix-based content delivery attains strictly predictable performance in forwarding and all relevant cost measures, which scale logarithmically with network and receiver sizes. It exhibits a churn resistance similar to the underlying key-based routing layer and a multicast efficiency scaling factor close to native group communication protocols. These results make the approach especially suitable for large and very large content groups. Detailed comparisons are drawn to Scribe [15], a generic shared tree approach derived from reverse path forwarding on top of the same structured overlay. On the one hand, our analytical insights will help content providers and ISPs to reliably predict future performance demands. On the other hand, they show that structured overlays enable scalable dynamic content delivery and adaptive CDNs.

The remainder of this paper is structured as follows: We present related work on content delivery in the context of structured P2P networks in Section 2. Section 3 describes the prefix-based routing approach, which serves as implementation for our modeling. We model and analyze the content delivery analytically in Section 4 and based on simulations in Section 5, deriving its characteristic performance measures in comparison to the generic shared tree approach Scribe. We discuss the overall insights and conclude in Section 6

2. Related work

Inspired by file sharing demands, many P2P systems have been developed over the past decade with solutions concentrating on maximizing client performance like upload capacities and NAT resilience. Recent research has identified the problems in delivery performance of the resulting pull-based distribution meshes [10], its scheduling [16], and in particular addressed scenarios of real-time video streaming for IPTV-type applications [17,18]. Starting from the interesting first hand principle of stochastic phase space simulations, Carra et al. [19] compared the general performance characteristics of tree- and mesh-based approaches. Leaving control overhead unconsidered, the authors derived that delivery delays of both schemes can approach similar performance values. At the same time, their analysis revealed that – given heterogeneous upload conditions – tree structures need to carefully adapt to deployment conditions and should reflect unbalanced network environments. The analysis of our work focuses on a generic tree construction scheme that admits a fair fan-out distribution at end points. However, using delegation mechanisms described in Section 3.5, trees can include data transfer at a minimal outdegree of 1 and thereby adapt to bottlenecks.

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 [20], Bayeux [21] as derived from Tapestry, and Scribe [15] or SplitStream [22] based on Pastry [23]. These approaches essentially branch in two algorithmic directions: Restricted broadcast uses DHTs to generate a structured sub-overlay network of group members, which thereafter is flooded (e.g., CAN). The second class erects distribution trees from explicit group member management. 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 a structured 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.

DHT-based multicast performance has been thoroughly studied in [24] based on simulations with the comparative focus on tree-based and flooding approaches. The separate construction of mini-overlays per group was shown to incur significant overhead. In addition, flooding was found to be outperformed by forwarding along trees. BIDIR-SAM uses a constrained flooding on prefix-subtrees [25] for group management with exponentially decreasing message load per receiver rank. Its data distribution follows optimal shortest-path trees.

Many unstructured P2P overlay multicast concepts exist [2]. Operating at a lower algorithmic complexity, but significantly higher control signaling efforts, performance characteristics for unstructured schemes differ. While DHT-based schemes are close to optimal with respect to message overhead and forwarding efficiency, they tend to create unbalanced distribution trees as an outcome of structured routing rules. Multicast tree properties comparing structured and unstructured schemes have been explored in [26]. Focusing on Scribe and SplitStream, and in agreement with our results, the authors identified a highly unbalanced forwarding load at inner tree nodes along with large fluctuations in delay. Our subsequent analysis reveals fairly balanced outdegrees in prefix-directed group distribution at very low, predictable jitter.

For the sake of completeness, we mention that probabilistic, gossip-based protocols (e.g., [27]) form an alternate approach to the large-scale content distribution problem. By increasing scalability through reducing coordinative information, those algorithms attempt to optimize the likelihood of a uniformly correct packet delivery. Following a merely probabilistically guided broadcast, these solutions tend to generate duplicates on a large scale.

As part of the fundamental debates on multicast, efficiency has come into focus. Grounded on empirical observations on the IP layer, Chuang and Sirbu [28] proposed a scaling power–law for the total number $L_N(g)$ of links in a multicast shortest path tree with g receivers of the form $L_N(g) \approx < L_U > g^{0.8}$, where $< L_U >$ represents the average number of unicast hops. Subsequent empirical and analytical work in [29–33] has debated the applicability of the Chuang and Sirbu law. Van Mieghem et al. [30] proved that the proposed power law cannot hold in general. Multicast cost efficiency for overlay tree structures have been examined in [34] experimentally and analytically, the latter using *complete* k-ary trees with receivers placed at leaf nodes and inner vertices. In general, no clear indication of a power law could be identified in the paper. The authors conclude that a power exponent of $\alpha = 0.9$ becomes visible for small numbers of receivers. In large scale simulations, this work identifies a power law simular to the Chuang and Sirbu observation for prefix-directed distribution, while Scribe's exponent slightly exceeds the theoretical estimate for overlay multicast.

3. Prefix-based, adaptive large-scale content delivery

Efficient packet forwarding for group data follows a distribution tree, in which branching nodes duplicate packets. The tree is based on a unicast network layer, which in structured overlays is provided by a key-based routing (KBR) implemented in DHTs such as Pastry [23] or Kademlia [35]. In this section, we briefly sketch BIDIR-SAM [14], a group distribution mechanism that follows an abstract prefix tree and allows for late binding. Prefix branching points are mapped to actual nodes only at forwarding. This can be based on any overlay prefix-routing substrate such as Pastry or Kademlia. BIDIR-SAM implements a group communication scheme with the exclusive advantage that *any* source can immediately distribute data according to *source-specific* shortest path trees. It does not require dedicated replication nodes such as rendezvous points. The prefix tree structure gives flexibility for inter-domain content delivery networks, as prefixes may be assigned to specific domains, causing traffic localization. Our subsequent work on an analytical performance model is based on the algorithmic definition of this prefix-based content delivery in structured P2P networks.

3.1. General background

Prefix-directed routing gives rise to a high degree of flexibility, and is used by DHTs (e.g., Pastry). A prefix represents all nodes whose identifier contains it. In general, routing data towards a prefix does not directly address a specific peer, but a set of nodes. Only at the forwarding decision will the prefix be resolved by selecting a destination node. Consequently, any tree structure based on prefixes may adapt to load or mobility, and is shielded from volatile peers as long as the mapping from prefixes to peers is maintained.

3.2. The core protocol

The BIDIR-SAM distribution tree is built from a prefix-based structured overlay. Overlay IDs are created using an alphabet of k digits. A prefix tree covering all DHT members can be immediately derived by identifying leaves as overlay IDs of *all* DHTs members and labeling recursively inner vertices with the longest common prefix (*LCP*) of their children (see Fig. 1). Without group membership management, forwarding along the prefix tree generates a broadcast. For sending a packet from the root to the 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. $\mathcal{C}$ will be updated hopwise with growing length. Downward forwarding is then simply achieved by routing to all neighboring prefixes that share $\mathcal{C}$. This mechanism, called Prefix Flooding [25], can be applied at any level of the tree structure and does not require explicit group management.

3.3. Group membership management

In contrast to broadcast, multicast implements a selective distribution strategy, where group members represent a subset of the peers. Each peer is a potential multicast forwarder, serving as an intermediate destination for a prefix it shares. Consequently, a new multicast listener has to be announced so that all forwarding nodes can store the corresponding neighboring prefix. This prefix represents the root of a subtree which covers multiple multicast listeners. Thus, only the first join and last leave has to be propagated outside this subtree.

To distribute data along a multicast distribution tree, a BIDIR-SAM peer K with overlay ID $\mathcal{K}$ maintains a multicast forwarding table MFT_G for each multicast group G. This list contains all prefixes, which serve as destinations adjacent to K. To join or leave a multicast group, a BIDIR-SAM node injects a state update into the unicast prefix tree. The first and last receiver of the group flood their join and leave message in the complete (unicast) overlay network. For all further group members, the state update is propagated within the smallest subtree including receivers and covering the multicast listener. Using MFT_G tables and Prefix Flooding, the algorithm works as follows:

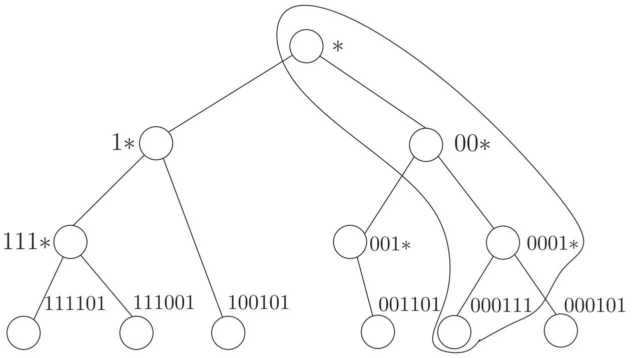


Fig. 1. Prefix tree highlighting all vertices associated with node 000111.

```
▷ Invoking this function at peer $\mathcal{K}$ for group $G$
if $MFT_G = \emptyset$
▷ Flood root (*) of the prefix tree
then PREFIX FLOODING Join/LeaveMessage To *
    ▷ Select longest prefix in forwarding table
else Select $\mathcal{L} \in MFT_G : |\mathcal{L}| \geqslant |\mathcal{L}'|, \forall \mathcal{L}' \in MFT_G$
    ▷ Creates 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 Receive

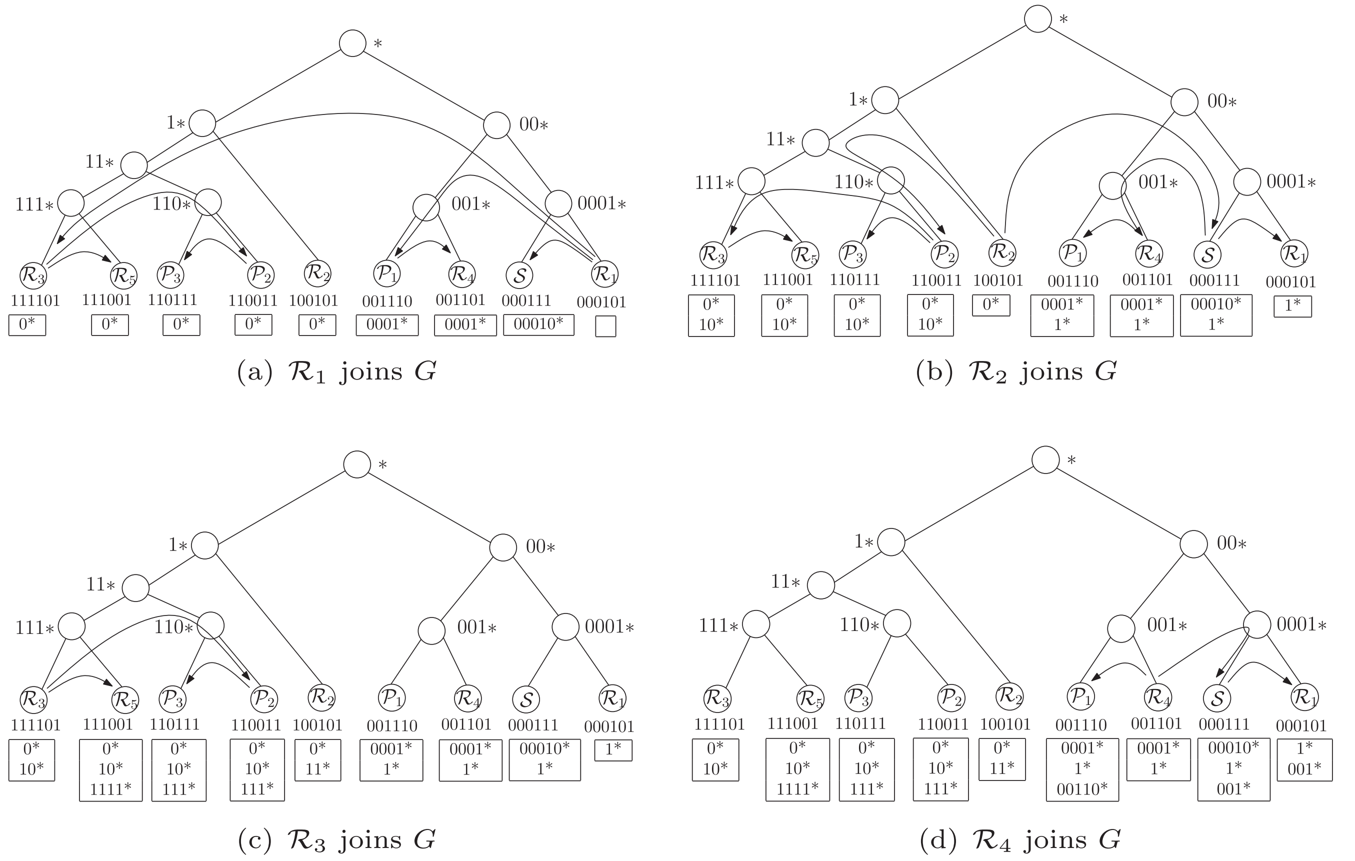
```
BIDIR-SAM JOIN/LEAVE INJECTION
▷ Denote the prefix of length $l$ of a key $\mathcal{A}$ by $prefix(l, \mathcal{A})$
▷ On arrival of message $m$ for group $G$ from peer $\mathcal{P}$ at node $\mathcal{K}$
$\mathcal{L} \leftarrow LCP(\mathcal{P}, \mathcal{K})$
$\mathcal{L}' \leftarrow prefix(|\mathcal{L}| + 1, \mathcal{P})$
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 the stepwise prefix elongation and the tree structure, no loops occur. The BIDIR-SAM join/leave algorithm consequently terminates and sends the group membership messages to all peers of the 'local' subtree. Furthermore, it guarantees a multicast spanning tree. The join procedure and thus the distributed state establishment is illustrated in Fig. 2 for four receivers.

3.4. Data fowarding

Based on its group membership functions, BIDIR-SAM constructs a bi-directional shared tree covering all overlay multicast listeners. Prefix neighbors towards receivers are stored in a decentralized multicast forwarding table MFT_G, which is controlled individually by each overlay node. An arbitrary peer can act as multicast source, while it sends the data to all entries in MFT_G. The packets will then be forwarded to the leaves of the multicast tree:

```
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 $\mathcal{C}_{new}$
```


*
1∗
00∗
11∗
111∗
110∗
001∗
0001∗
$\mathcal{R}_3$ $\mathcal{R}_5$ $\mathcal{P}_3$ $\mathcal{P}_2$ $\mathcal{R}_2$ $\mathcal{P}_1$ $\mathcal{R}_4$ $\mathcal{S}$ $\mathcal{R}_1$
111101 111001 110111 110011 100101 001110 001101 000111 000101
(a): 0* | 0* | 0* | 0* | 0* | 0001* | 0001* | 00010* |
(b): 0* 10* | 0* 10* | 0* 10* | 0* 10* | 0* | 0001* 1* | 0001* 1* | 00010* 1* | 1*
(c): 0* 10* | 0* 10* 1111* | 0* 10* 111* | 0* 10* 111* | 0* 11* | 0001* 1* | 0001* 1* | 00010* 1* | 1*
(d): 0* 10* | 0* 10* 1111* | 0* 10* 111* | 0* 10* 111* | 0* 11* | 0001* 1* 00110* | 0001* 1* | 00010* 1* 001* | 1* 001*

(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.

Data is sent to roots of subtrees as extracted from the multicast forwarding table, and distributed therein with growing destination prefix. Thus, all multicast listeners receive the data exactly once and the algorithm terminates.

3.5. Protocol extensions

The BIDIR-SAM core protocol creates and manages a generic shared family of source trees in prefix space, which allows 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, which are in network proximity for regional moves. 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 [36]. These improvements apply without extra signaling or management overhead.

3.6. Application example: prefix aggregation in P2P TV

The distribution structure of BIDIR-SAM can be naturally applied to next-generation P2P IPTV scenarios, as it allows for inherent proximity between end users and optimized traffic flows. With respect to current viewing practices, the authors in [37] propose a cooperative multicast and P2P approach, which spans P2P distribution among set-top boxes (STBs). Recalling the typical IPTV architecture, STBs are aggregated by Digital Subscriber Line Access Multiplexers (DSLAMs). If all peers behind the same DSLAM carry the same prefix, distribution among them will remain local and backbone crossing is minimized. Group join and leave messaging among receivers will likewise stay local to DSLAM domains, and facilitate fast channel switching. This adaptation to the network architecture occurs automatically and does not require any signaling.

A corresponding prefix assignment can be achieved by splitting the hash function used for node ID creation across the DSLAMs. As overlay IDs do not carry further semantic, a common prefix can be assigned to all nodes behind a DSLAM. Each DSLAM thus corresponds virtually to an inner vertex label of the BIDIR-SAM distribution tree. Based on the BIDIR-SAM distribution structure, locality-aware large-scale content distribution can be deployed without provider interaction.

3.7. Performance metrics

We will analyze the performance of prefix-based content delivery based on the metrics described in this section. Our analytical results (Section 4) are verified by simulations (Section 5). In addition, we compare the results for prefix-directed data distribution with the well-known rendezvous point-based scheme Scribe [15].

Hop count measures the number of overlay routing hops that a packet needs on its way from the source to the destination.

Multicast forwarding entries corresponds to the number of downstream entries required at a peer. This value represents the storage space at a multicast peer. It also characterizes the number of children per overlay node in the distribution tree. Thus, it describes an upper bound for the packet replications in BIDIR-SAM and corresponds to the replication load in Scribe.

Signaling load measures the average number of multicast join messages initiated due to a subscription of a new multicast listener. This value quantifies the cost at peers of incorporating a new receiver in the multicast tree.

Packet replication load quantifies the number of packets a peer has to forward. This metric reflects the number of direct neighbors per node in the distribution tree.

Multicast efficiency defines, similarly to [28], the ratio of the average number of traversed *overlay* hops by distributing the data via multicast and the average *overlay* unicast path length. This normalized measure reflects the economic effect of multicast over repeated unicast.

Delay stretch measures the ratio of the overlay and native multicast path length with respect to [24]. It is worth noting that only the stretch depends on the underlying network topology.

Packet delivery ratio measures the relative number of packets that are correctly transported to the receivers. This metric is applied at the occurrence of churn and estimates the robustness of the overlay scheme in unstable member conditions.

4. Analytical model & performance results

The well defined prefix structure of the content delivery scheme BIDIR-SAM (cf., Section 3) allows for a detailed theoretical analysis, yielding strong analytical results for all major properties. To clarify the underlying model, we first give an overview of the concepts and notations and outline common properties as needed further on.

4.1. The general model and basic properties

For a given key space of alphabet size k and key length h, we consider the corresponding k-ary prefix tree as basic structure. Therein N overlay nodes with prefix set $\{\mathcal{N}\}$ are uniformly placed at leaf nodes of the prefix tree. In particular, for any node K with key $\mathcal{K}$, the probability of attaining a specific digit x reads

$$P(\delta_i(\mathcal{K}) = x) = \frac{1}{k},$$

where $\delta_i(\mathcal{K})$ denotes the i-th digit of $\mathcal{K}$.

Consider an arbitrary prefix $\mathcal{C}$ of length j. The probability for a random overlay node to share this prefix equals $(\frac{1}{k})^j$, any (ordered) sequence of keys, l keys with prefix $\mathcal{C}$ and $N-l$ keys not sharing $\mathcal{C}$, occurs with probability $\left(\frac{1}{k^j}\right)^l \left(1-\frac{1}{k^j}\right)^{N-l}$. Accounting for all possible orderings yields the node distribution in prefix space,

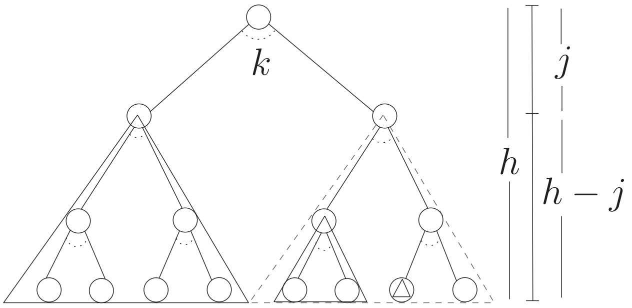


Fig. 3. A Prefix tree with inner vertices defining the root of subtrees with self-similar properties due to the recursive nature of k-ary trees.

$$P(|\{\mathcal{N} \in \{\mathcal{N}\}|\ LCP(\mathcal{C},\mathcal{N}) = \mathcal{C}\}| = l) = \binom{N}{l}\left(\frac{1}{k^j}\right)^l\left(1-\frac{1}{k^j}\right)^{N-l}, \qquad (1)$$

which is Binomial. The prefix $\mathcal{C}$ of length j correspondents to the root of a subtree T_{h-j} of height $h-j$ as visualized in Fig. 3. Hence, Eq. (1) also describes the distribution of nodes populating subtrees.

The keys representing the overlay nodes span a *random recursive k-ary tree* with *inhomogeneous branching rates*. An inner vertex represented by the prefix $\mathcal{C}$ at level $j-1$ on the path to a given node S with key $\mathcal{S}$, will be a branch of the prefix tree, if a node K with key $\mathcal{K}$ exists, such that $LCP(\mathcal{S},\mathcal{K}) = \mathcal{C}$. The latter is equivalent to the existence of a node that attains a dedicated prefix of length $j-1$, and any of $k-1$ from k values at the j-th digit. The probability that none of the $N-1$ remaining nodes carry a dedicated prefix of length j equals $\left(1-\frac{1}{k^j}\right)^{N-1}$. Hence, the branching probability of the overlay prefix structure at level $j-1$ reads

$$P_{Branch}(j-1) = \left(1-\left(1-\frac{1}{k^j}\right)^{N-1}\right)\cdot(k-1) \qquad (2)$$

Any distribution system for a multicast group is organized in accordance with this overlay prefix structure. Consider a group G of g receivers. We assume that receivers are independently chosen among overlay nodes with the uniform probability $r_g = \frac{g}{N}$.[1] The BIDIR–SAM algorithm aggregates multicast receivers according to longest prefixes.

Theorem 1. *For a multicast group G resident in a prefix-structured overlay of k-ary alphabet and N nodes, the probability that a given prefix $\mathcal{C}$ of length j is attained by at least one out of g receivers is given by*

$$P(|\{\mathcal{G} \in G|\ LCP(\mathcal{C},\mathcal{G}) = \mathcal{C}\}| \geqslant 1) = 1-\left(1-\frac{g}{k^jN}\right)^N = 1-e^{-\frac{g}{k^j}} + \mathcal{O}\left(\frac{1}{N}\right) \qquad (3)$$

[1] This assumption is supported in both, theory by [29] and Internet measurements by [31].

Proof. The conditional probability (denoted by $\|$) that none of the nodes is a multicast receiver, while the number $N_{\mathcal{C}}$ of nodes with prefix $\mathcal{C}$ in the overlay equals l, reads

$$P(|\{\mathcal{G} \in G|\ LCP(\mathcal{C},\mathcal{G}) = \mathcal{C}\}| = 0\|\ N_{\mathcal{C}} = l) = (1 - r_g)^l$$

$N_{\mathcal{C}}$ is distributed according to Eq. (1), hence

$$P(|\{\mathcal{G} \in G|\ LCP(\mathcal{C},\mathcal{G}) = \mathcal{C}\}| = 0)$$

$$= \sum_{l=0}^{N} \binom{N}{l} \left(\frac{1 - r_g}{k^j}\right)^l \left(1 - \frac{1}{k^j}\right)^{N-l} = \left(1 - \frac{g}{k^j N}\right)^N,$$

where the last line was obtained by evaluating the binomial expansion series [38]. Taking the complementary weight and observing that $e^x = \left(1 + \frac{x}{N}\right)^N + \mathcal{O}\left(\frac{1}{N}\right)$ proves the theorem. □

It is worth noting that in large overlay networks the prefix distribution of multicast receivers is effectively independent of the overlay size.

4.2. Size of multicast forwarding tables

An upper bound for the state table can be easily derived through the following observation: Any overlay node is situated as a leaf in the prefix tree and has all vertices on the shortest path to the root associated with it. Thus, the number of neighbors equals the sum of the neighbors at each associated vertex. For an alphabet of base k the latter is bound by $k - 1$. The number of vertices towards the tree root is limited by the height of the tree, which is maximal when all branches are binary.

Theorem 2. *For any overlay node in a k-ary prefix tree with g receivers, the number of multicast forwarding table entries is bound by* $\log_2(g)(k - 1)$*.*

We now want to determine the distribution of multicast forwarding states on the prefix tree. At every level j, an overlay node may face 0 to $k - 1$ neighboring vertices connecting different receivers.

Theorem 3. *In the BIDIR-SAM multicast scheme of a group with g receivers, the probability distribution P(j,l) that a given overlay node holds l multicast forwarding entries of prefix length j reads*

$$P(j,l) = \binom{k-1}{l} \left(1 - e^{-\frac{g}{k^{j+1}}}\right)^l \left(e^{-\frac{g}{k^{j+1}}}\right)^{k-1-l} + \mathcal{O}\left(\frac{1}{N^2}\right) \qquad (4)$$

Proof. For a given node consider the possible vertices connecting to the $k - 1$ subtrees at level j. A forwarding state for a particular vertex will be required, if and only if a receiver exists in the corresponding subtree. Being member of a particular subtree with root at level j is equivalent to carrying a prefix of length $j + 1$; its probability was given in Eq. (3). Selecting l forwarders among the $k - 1$ vertices, and adding all possible orderings proves the theorem. □

Mean functions are plotted in Fig. 4(a) for different alphabets. Table entries remain significantly below upper bounds given in Theorem 2, reproducing nicely the logarithmic dependency on g. Increase with k remains sublinear.

4.3. Replication load

The maximal value of the replication load is defined by the number of forwarding table entries and comes into effect with a destination prefix of zero length. Routing from zero prefixes occurs only at the multicast source and leads to the immediate implication of Theorem 2:

Corollary 1. *The multicast replication load for any overlay node remains less or equal to* $\log_2(g)(k - 1)$*.*

In the general case, multicast forwarding occurs in combination with a destination prefix of length j, which rules out all table entries of shorter prefix length.

Corollary 2. *Denote by RPL(j,g) the multicast replication load at a node in an overlay participating in the BIDIR-SAM multicast with prefix length j. Then*

$$\langle RPL(j,g)\rangle = \sum_{i=j}^{h-1} (k-1)\left(1 - e^{-\frac{g}{k^i}}\right)$$

Characteristic distributions of the replication load are drawn in Fig. 4(b), showing a steady decrease in packet replication on the way from source to receivers. Smaller alphabets noticeably smoothen the distributions, which suggests k serving as a tuning parameter of the multicast distribution tree.

4.4. Signaling load

Signaling in schemes like BIDIR-SAM consists of the Join and Leave messages, which are flooded to overlay subtrees selected according to established forwarding states.

Consider an established group G of g receivers in the overlay network. A node newly joining (or leaving) group G will change the group members to $g + 1$ in distributing its Join to the smallest subtree containing its own ID and at least one previous receiver. The probability $P(j,g)$ that a Join injection occurs at level j, or at a subtree of height $h - j$, is equal to the probability that one of the previous g group members shares the prefix of length j with the newly joining node, but none does with the extended prefix of length $j + 1$. Hence, using Eq. (3), we derive the following theorem:

Theorem 4. *The probability P(j,g) for distributing a Join or Leave message within a prefix tree at injection level j reads* $P(j,g) = \left(1 - e^{-\frac{g}{k^j}}\right)e^{-\frac{g}{k^{j+1}}} + \mathcal{O}\left(\frac{1}{N}\right)$*, where g is the number of group members prior to signaling.*

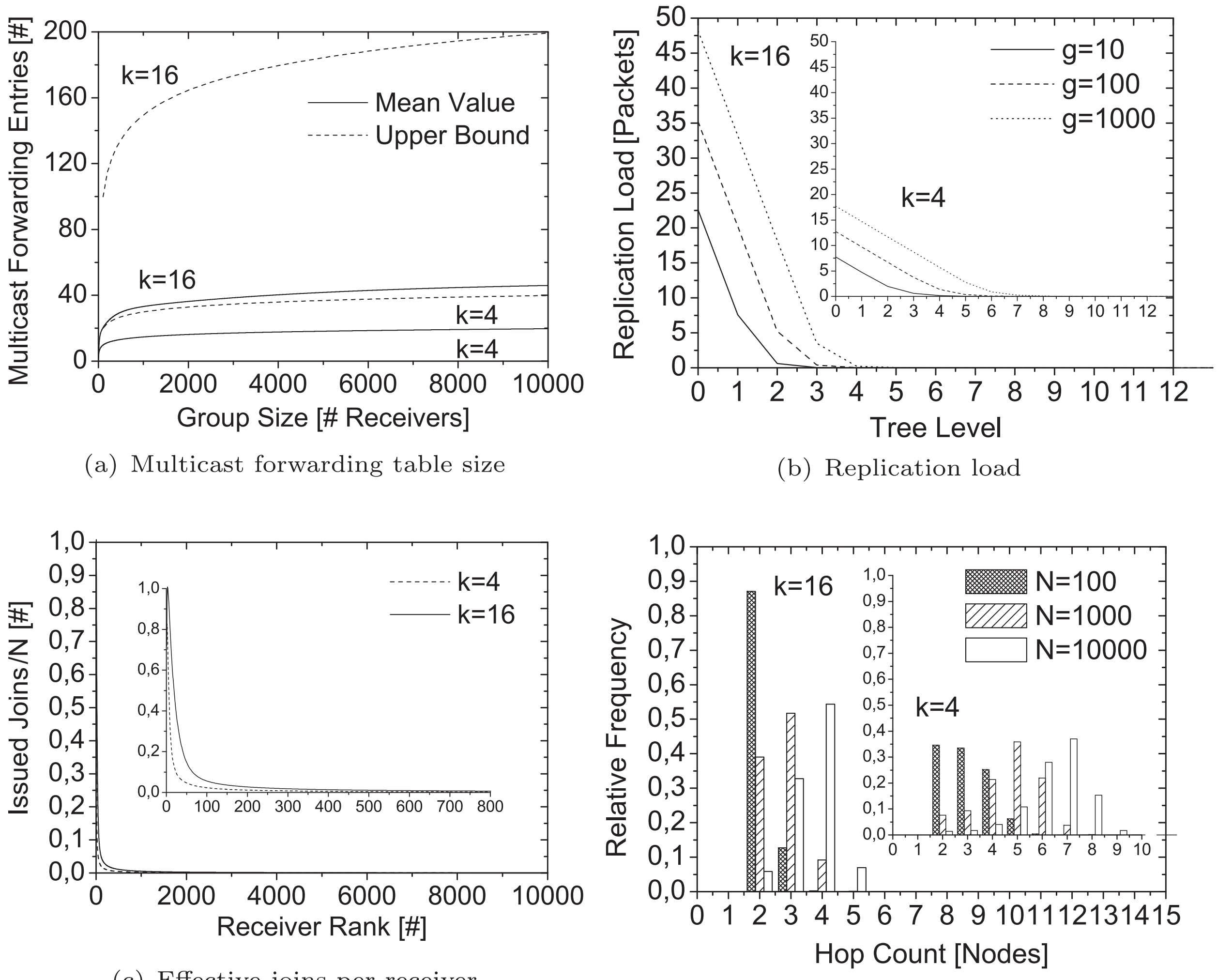
From this distribution the expected number of nodes within the subtrees can be deduced (see [39] for a proof).

Corollary 3. *The expected ratio of flooded nodes is well approximated by*

$$(1 - e^{-g})e^{-\frac{g}{k}} + \frac{k}{g(k+1)\ln k}\left(\left(e^{-\frac{g}{k^{h+1}}} - e^{-\frac{g}{k}}\right)(k+1) + e^{-\frac{g(k+1)}{k}} - e^{-\frac{g}{k^{h+1}}(k+1)}\right)$$

where g is the number of group members prior to signaling.

The results are displayed in Fig. 4(c) as functions of the joining receiver rank. Signaling expenses admit a strong exponential decay in the expected number of flooded nodes. The mean number of messages issued for Join/Leave



(a) Multicast forwarding table size

(b) Replication load

(c) Effective joins per receiver

(d) Normalized hop count distributions

Fig. 4. Analytical results for prefix alphabets of $k = 4$ and $k = 16$.

signaling reduces to below 1 % for group sizes above 500, keeping group management costs marginal in larger distribution scenarios. Signaling expenses decrease almost linearly with k. Again, k serves as a tuning parameter acting in the same direction as for the replication load.

4.5. Hop count

We first recall that prefix routing proceeds down a prefix tree and hence hop numbers are limited by its height.

Theorem 5. *Any multicast receiver in an overlay of N receivers that performs a prefix routing using an alphabet of $k \geqslant 2$ digits will receive a packet after at most $\log_2(N)$ hops. In the presence of Pastry overlay routing, the number of hops attained on average equals $\log_{2^b}(N)$ with $k = 2^b$.*

Compliant with our model, we now want to derive a hop count distribution that represents sparsely scattered receivers in a prefix tree.

On the path from the source to the receivers, a multicast packet traverses an overlay hop, whenever the distribution tree branches at the corresponding prefix $\mathcal{C}$. Taking the branching rate given in Eq. (2), the corresponding recurrence relation of the hop count frequency can be written as

$$f_{h,k,N}(j) = f_{h-1,k,N}(j) + \left(1-(1-k^{-j})^{N-1}\right)\cdot(k-1)\cdot f_{h-1,k,N}(j-1) \qquad (5)$$

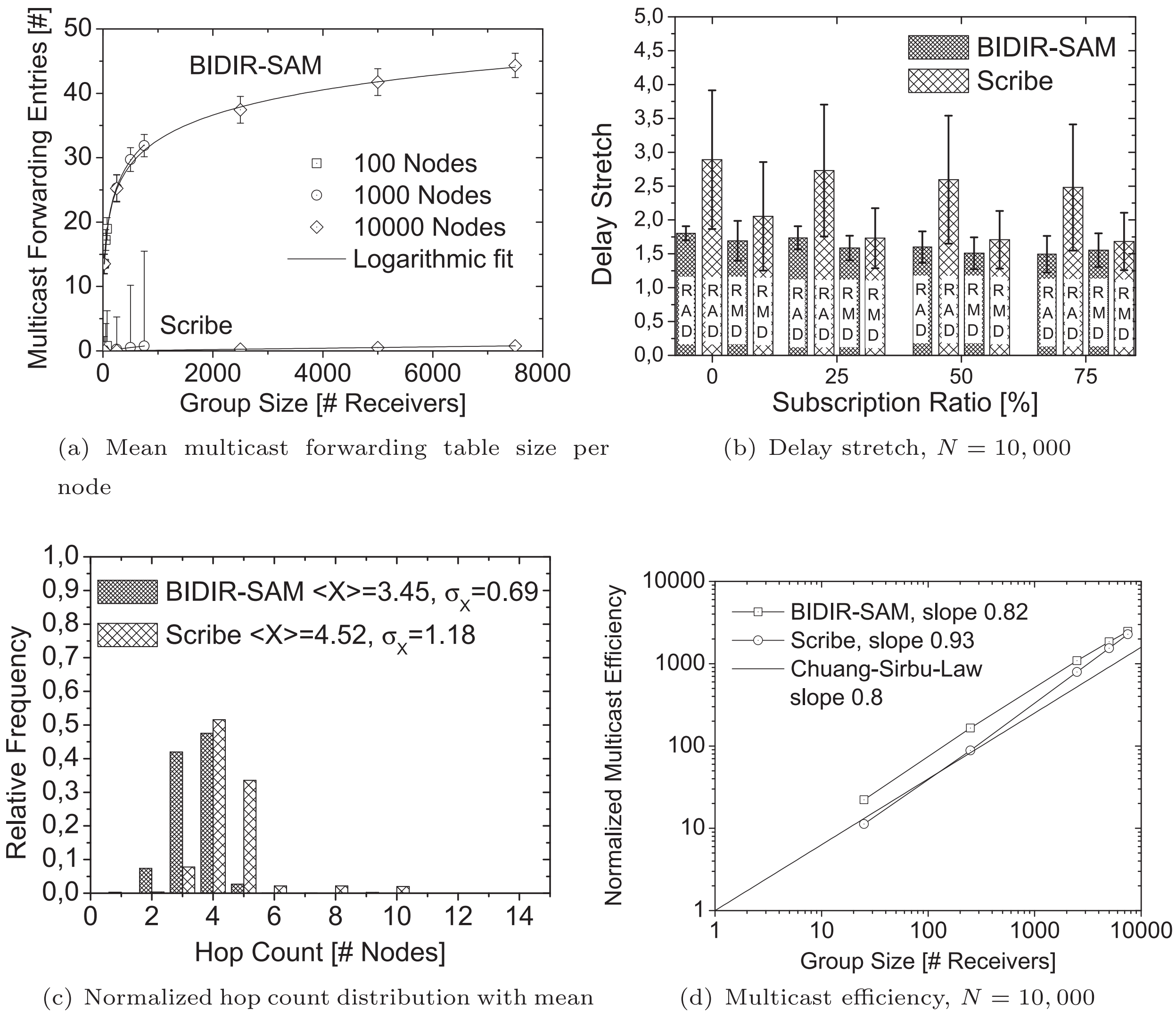
with $f_{1,k,N}(0) = 1$, $f_{1,k,N}(1) = (1-(1-k^{-1})^{N-1})(k-1)$.

Solving the recursion leads to the following theorem:

Theorem 6. *The hop count frequency $f_{h,k,N}$ attained at prefix routing on N overlay nodes with independent uniformly distributed identifiers is given by*

$$f_{h,k,N}(j) = \binom{h}{j}\cdot\prod_{i=0}^{j}\left(1-(1-k^{-i})^{N-1}\right)\cdot(k-1)^{j}. \qquad (6)$$

The hop count frequency $f_{h,k,N}(j)$ is plotted in Fig. 4(d) in normalized form. Mean and width of the distributions grow as k decreases, acting in opposite direction of the branching properties investigated above. A distribution instance optimizing replication and signaling load by using a small prefix alphabet will encounter a moderate increase of routing hops in packet delivery.



(a) Mean multicast forwarding table size per node

(b) Delay stretch, $N = 10,000$

(c) Normalized hop count distribution with mean ($\langle X \rangle$) and standard deviation (σ_X), $N = 10,000$

(d) Multicast efficiency, $N = 10,000$

Fig. 5. Simulation results for prefix alphabet of $k = 16$.

5. Simulation results

In this section, we will verify our analytical results by simulations. We compare the performance of prefix-directed forwarding (BIDIR-SAM) with a rendezvous point approach (i.e., Scribe). Scribe is used in its standard implementation [15]. BIDIR-SAM as well as Scribe are implemented on top of the key-based routing layer Pastry. The simulation starts after a proactive routing maintenance has completely filled Pastry's routing tables. The cost for this additional routing maintenance is low, as incomplete routing tables rarely occur. The simulations are performed on OMNeT++3.3 [40], extended by OverSim [41].

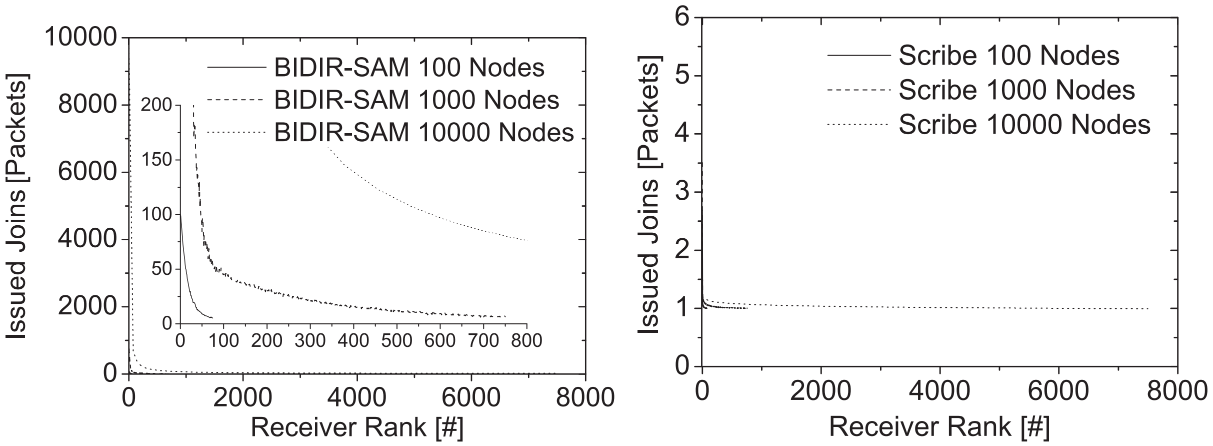
Pastry is configured in its original version with a key length of 128 bits and a varying prefix alphabet size k.[2] In agreement with the multicast routing protocol aspects of investigation (cf., Section 3.7), we either select the simple model of OverSim [41] with a homogeneous link delay of 1 ms, or predicted network distances based on the global network positioning [42] and CAIDA Skitter data.

The simulations are conducted for small, medium and large overlays of 100, 1,000 and 10,000 nodes. Among all peers, one uniformly distributed content source is chosen along with its group address. Receivers are also picked uniformly among nodes, but distinct from the source. We average our results over all samples of identical settings.

5.1. Multicast forwarding table (MFT) size

The average MFT size is visualized in Fig. 5(a) as a function of the number of receivers for different network sizes. Both schemes scale independently of the overlay dimension due to the local view of multicast forwarders. Focusing on mean values, Scribe outperforms BIDIR-SAM as the average number of entries grows only marginally with the group size and remains below 5. However, all BIDIR-SAM tables increase within strict logarithmic bounds as functions of receivers, which complies with the scaling

[2] We present only figures for $k = 16$ as results show qualitative similar behavior to our analytical study.



(a) BIDIR-SAM (b) Scribe

Fig. 6. Effective joins per receiver for prefix alphabet of $k = 16$.

properties of the underlying DHT. It is worth noting that the additional entries in BIDIR-SAM provide inherent redundancy as distributed prefixes cover multiple peers.

Large fluctuations, as indicated by the (first) error bars in Fig. 5(a), are well known for Scribe. Although its average number of tables entries is almost constant, maximal values range up to 5,600 entries for large overlays with a high receiver ratio. The distribution of group states reveals, that almost all peers keep a MFT without entries, but some dedicated overlay nodes maintain single states for up to 80% of the receivers. In contrast, BIDIR-SAM fills its routing tables in a balanced way.

5.2. Signaling load

Fig. 6 displays the average signaling load for join/leave for different group sizes ranked by receiver subscription order. The number of issued joins is higher in BIDIR-SAM than in Scribe, which is also indicated by the different axis scales. This behavior reflects directly the difference of underlying group management algorithms and corresponds to our analysis of the MFT size. In BIDIR-SAM, each new receiver floods a prefix subtree of different height, whereas Scribe submits single subscriptions on unicast paths towards the rendezvous point (RP).

In BIDIR-SAM, the load decays exponentially with the number of receivers. For larger multicast groups with a receiver to overall peer ratio of more than 50%, BIDIR-SAM approximates asymptotically the signaling load of Scribe. Any BIDIR-SAM peer, however, owns at this time a richer, redundant excerpt of the overall prefix tree. While Scribe states represent a single shared tree, which may break into incoherent parts, whenever intermediate states are lost [26], BIDIR-SAM distributes its states among nodes, procuring a redundant source-specific tree infrastructure suitable for resource pooling.

Join signaling may be tuned by adjusting the prefix alphabet size. Decreasing k accelerates BIDIR-SAM convergence (graphs omitted). Scribe shows an opposite effect and increases slightly with the submitted joins as paths to the rendezvous point enlarge.

5.3. Delay stretch

Fig. 5(b) visualizes the delay stretch for the maximum delays (RMD) and average delays (RAD) [24] as function of the receiver population. Clearly, BIDIR-SAM outperforms Scribe in both measures attaining a stretch of 1.5–1.7 with negligible standard variations. Surprisingly, the average gain is considerably higher than the ratio of maximum delays. This discrepancy between RAD and RMD behavior can be explained by regarding rendezvous point effects within a fluctuating underlay topology. Routing via an RP may add a relative delay up to the maximum to the average, but at most a constant to RMD.

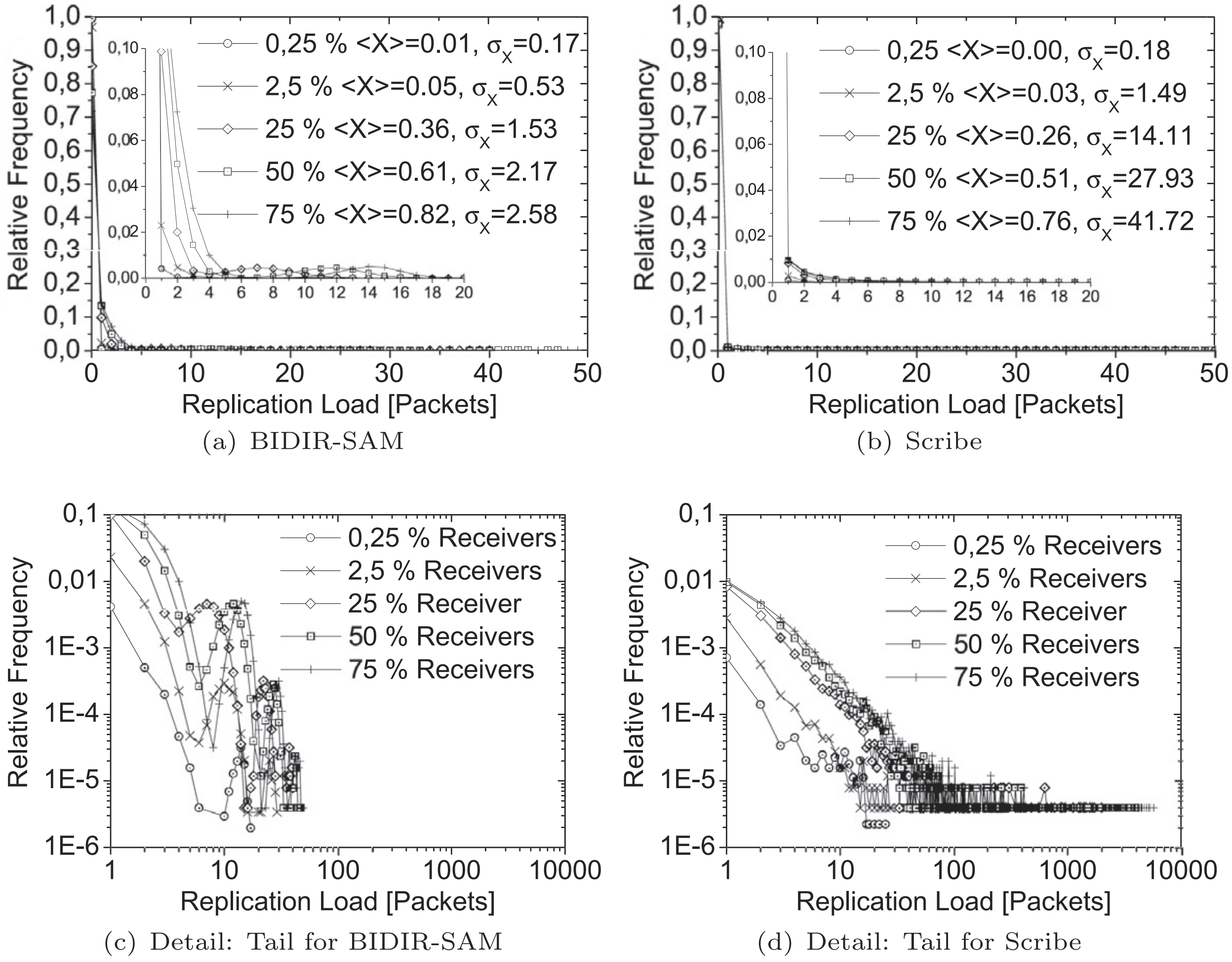
It is worth noting that the delay model does not reflect asymmetric routes, which would positively affect the performance of BIDIR-SAM in contrast to Scribe due to forward oriented path setup.

5.4. Replication load

A detailed view of the packet replication for a varying number of receivers in a fixed size overlay is given in Fig. 7.[3] Both schemes exhibit a sharp peak for low replication values and decay exponentially. However, they differ significantly in detail. The standard variation increases linearly for Scribe and negligibly for BIDIR-SAM with a higher receiver population.

The asymptotical growth of the packet replication in Scribe depends strongly on the group size. Additional listeners of content, thus, increase the maximal replication load. This indicates a tendency that additional receivers construct branches that meet the maximal load replicator, which further implies that a single peer is responsible for forwarding the content to almost all group members. In contrast, BIDIR-SAM balances the load. As visualized in the log–log plots in Fig. 7(c) and (d), the distribution of Scribe is heavy-tailed, decaying like a power law with

[3] We omitted the distribution for different network sizes and receiver populations as the overall shape depends mainly on the number of listeners.



(a) BIDIR-SAM

(b) Scribe

(c) Detail: Tail for BIDIR-SAM

(d) Detail: Tail for Scribe

Fig. 7. Distribution of packet replication in a 10,000 node overlay, a varying receiver subscription ratio and a prefix alphabet of $k = 16$.

significant probabilities for very large replication values up to 7800. In contrast, the distribution of BIDIR-SAM admits a strict exponential decay, with tail weights vanishing at 50. It reduces the replication maxima by distributing the load fairly and evenly over the neighbors.

The branching factor k shifts weights of higher replications, which results in a lower maximal load, but in a slightly increased load per peer and longer paths. The reduced k smoothes the tail of BIDIR-SAM, as branches are populated more densely and replications occur as multiples of $k - 1$ [39].

5.5. Hop count

Fig. 5(c) visualizes the hop count distribution for an overlay of 10,000 nodes. For Scribe, the RP is clearly visible by elongating the paths by at least one hop. Further simulations show that in an overlay with 100 nodes Scribe attains $\langle X \rangle = 2.68$, $\sigma_X = 0.52$ and BIDIR-SAM $\langle X \rangle = 1.91$, $\sigma_X = 0.55$; for 1,000 peers $\langle X \rangle = 3.53$, $\sigma_X = 0.87$ and $\langle X \rangle = 2.68$, $\sigma_X = 0.63$ respectively. Thus, both approaches exhibit a logarithmically increasing path length, which results from the underlying Pastry prefix tree. The standard deviations grow logarithmically in BIDIR-SAM, but linearly in Scribe. Thus, BIDIR-SAM tightly concentrates path lengths around the average and Scribe develops longer branches with higher weights. A smaller prefix alphabet increases the height of the constructed distribution tree, but preserves the general behavior of both approaches.

5.6. Packet delivery ratio

We evaluate the resilience of peers against churn following a lifetime churn model. Real-world measurements revealed that session lengths are similar across different networks and well approximated by a Weibull distribution [43]. According to previous studies [11] and consistent with our deployment scenarios, we analyze the robustness of BIDIR-SAM for moderate lifetimes per peer, varying the scale parameter of the Weibull distribution from 100 to 800 seconds with a constant shape parameter of 1. The model is applied to all peers except the content source, which is not affected by churn. It is worth noting that we use plain Pastry without any special optimizations for churn. This is not a limitation as we do not want to study the robustness properties for specific implementations but want to explore the relative behaviour of the different content distribution concepts.

The packet delivery ratio for BIDIR-SAM, Scribe, and Pastry are shown in Fig. 8. BIDIR-SAM attains performance

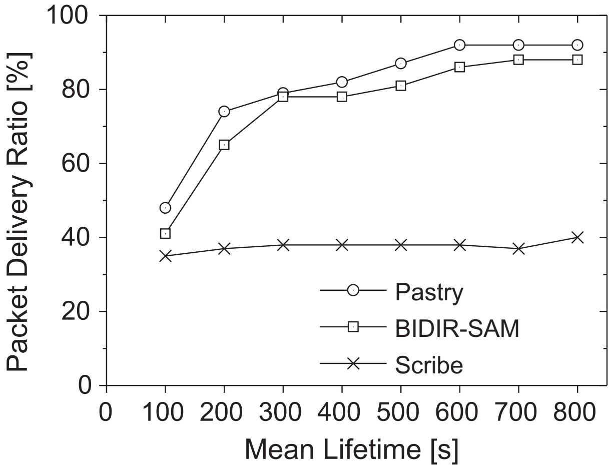


Fig. 8. Successful packet delivery under churn for varying node lifetimes.

values similar to Pastry with a success ratio of up to 90%. For Scribe the packet loss is significantly higher. For long sessions, it delivers only 40% of the data. The results show clearly that – due to the late binding concept – the robustness of BIDIR-SAM depends directly on Pastry. BIDIR-SAM benefits from the virtualized prefix tree concept, which adaptively selects available next hop neighbors, respecting the current state of the underlying key-based routing protocol. In contrast to this, Scribe performs an early binding at group management time and relies on a single rendezvous point. Both aspects remarkably enhance the churn effects inherited from Pastry.

5.7. Multicast efficiency

The normalized multicast efficiency is plotted in Fig. 5(d) for different overlay sizes. For comparison, we also show the normalized multicast efficiency with respect to the analysis by Chuang and Sirbu [28]. Using a standard prefix alphabet with 16 digits, the absolute efficiency values of BIDIR-SAM and Scribe are higher than observed by Chuang and Sirbu for native multicast. However, this metric reflects the scaling behavior of multicast protocols with growing group sizes. The slope, which represents the scaling factor, is calculated based on a linear fit. It is clearly visible that the basic BIDIR-SAM tree grows with a rate similar to native multicast. In contrast, Scribe exhibits a higher scaling factor. This indicates that Scribe paths are constructed less efficiently. The observation coincides with our previous results, which show a high, unicast-like replication load for Scribe around a single node.

6. Discussion & conclusion

This paper presented an analytical model for prefix-directed forwarding in structured P2P networks. The model is based on k-ary trees. It allows for performance predictions and enables stakeholders to forecast capacity requirements. We verified and extended our results by simulations. As a concrete protocol implementation we used BIDIR-SAM, an overlay content delivery approach that enables any peer to distribute data directly into a content group. Using a logical prefix overlay and a bi-directional, shared distribution tree, costs in signaling and forwarding are strictly predictable and scale logarithmically with the network and group size. To the best of our knowledge, this is the first systematic analysis of an overlay content delivery approach that distributes data within a shared tree model on source-specific shortest path trees. Our theoretical analysis quantifies measurements almost in exact agreement with the empirical calculations.

We compared content delivery on bi-directional shared distribution trees with the paradigm of rendezvous point-based schemes. For the latter we used the implementation Scribe. Prefix-directed forwarding admits superior performance in overall data distribution and scaling behavior, in particular when scaling towards very large networks and content groups. Scribe, which erects a single shared tree, outperforms BIDIR-SAM on *average* with respect to the tree construction costs for small size groups, but obligates dedicated overlay nodes with unbound storage and forwarding load. As a result, Scribe performance values fluctuate on a large scale, leading in particular to high jitter values at the receiver nodes.

In the following, we will detail implications for deployments. Under the perspective of practical relevance, we do this for both real protocols, BIDIR-SAM and Scribe.

6.1. Large-scale vs. small groups

Our analysis reveals that BIDIR-SAM packet distribution metrics and overall resource requirements scale evenly as logarithmic functions of the group size, while most performance values of Scribe fluctuate on a scale linearly growing with group members. In contrast to Scribe, BIDIR-SAM has to flood prefix listener subscriptions to prefix (sub-) trees, whose size depend on the receiver population. While the first join message is distributed to all overlay nodes, the cost decays exponentially with further multicast listeners. Groups exceeding 500 receivers in large overlays closely approach the very low costs of Scribe. Thus, for application scenarios with a stable number of receivers, BIDIR-SAM exhibits appropriate cost, even while most of the multicast listeners may change.

Multicast state updates directly influence multicast forwarding table sizes. In general, the average number of entries per peer is higher in BIDIR-SAM than in Scribe. Nevertheless, BIDIR-SAM table sizes exhibit a strict logarithmic bound, while storage in Scribe may grow linearly. Individual Scribe peers frequently store almost all multicast forwarding states. In contrast, BIDIR-SAM instantaneously distributes states fairly among all peers in small and in large groups.

Additionally, BIDIR-SAM outperforms Scribe with respect to the forwarding costs. A BIDIR-SAM multicast sender can control the maximal load it imposes onto the distribution infrastructure, which is a simple but effective QoS instrument. Based on the structural protocol properties it follows that in case of multiple sources, the source-specific distribution model of BIDIR-SAM will balance the load automatically.

BIDIR-SAM always operates on a single virtual distribution tree, which is collectively known at peers. This common forwarding instruction leads to a coherent overlay routing performance among all peers, and is of particular importance for multipath transport. The latter becomes vital in mobility or load sharing scenarios. It further prevents BIDIR-SAM from admitting intolerable jitter values under source variations as have been observed for SplitStream.

6.2. The problem of asymmetric routes

Observing the hop count and packet replication distribution, the question arises about the more fundamental reasons why BIDIR-SAM consistently outperforms Scribe. Leaving aside the RP-issues, the main conceptual difference between data-driven tree approaches and prefix flooding 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. Such a problem does not arise, if the source constructs its tree according to forward routes.

In DHT-based group communication, the direction of tree establishment is even more important. The distribution tree in Scribe is built from receiver subscriptions towards the RP, but the packets flow in the inverse direction. As the association of prefixes to nodes is not unique, two peers may select a different destination for the same prefix. Thus diverse paths will be established, even though packets could uniformly traverse the reverse directions following the RP point of view. In contrast, BIDIR-SAM solely uses forward-oriented directives, extracted directly from unicast DHT routing control.

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.

6.3. Overall performance

The performance of BIDIR-SAM is uniformly and strictly predictable over all peers, whereas Scribe produces an unfair, irregularly fluctuating load at forwarders. BIDIR-SAM constructs shorter paths and creates lower replication loads, which remain quite stable with growing group sizes. Operating on forward-oriented, prefix-defined paths, BIDIR-SAM not only complies with asymmetric links and hop alterations, but takes higher-than-average advantage of proximity selections at the KBR layer.

BIDIR-SAM can nicely be tuned by the prefix alphabet parameter. A smaller prefix alphabet directly smoothes the branching, which in turn reduces signaling and replication load per peer. All deviations from mean values thereby remain small. This overall balancing effect faces the increase in path lengths as its only negative side-effect, while increase logarithmically with the alphabet size. In contrast to this, the branching parameter has only marginal effects on the performance of Scribe.

Prefix-based content delivery based on bi-directional shared tress is an interesting concept with promising performance properties. Our analytical model provides a good foundation to evaluate its strengths and weaknesses. Considering the current debate on ISP-friendly P2P networks, it will be challenging to extend the theoretical analysis to include underlay-overlay performance in follow-up work.

Acknowledgements

The authors thank the anonymous referees for their valuable comments, which helped to improve the paper. This work is partially supported by the German Federal Ministry of Education and Research within the HAMcast project, which is part of G-Lab.

References

[1] S. Deering, Host extensions for IP multicasting, RFC 1112, IETF, August 1989.

[2] M. Hosseini, D.T. Ahmed, S. Shirmohammadi, N.D. Georganas, A survey of application-layer multicast protocols, IEEE Communications Surveys and Tutorials 9 (3) (2007) 58–74.

[3] J. He, A. Chaintreau, C. Diot, A performance evaluation of scalable live video streaming with nano data centers, Computer Networks 53 (2) (2009) 153–167.

[4] B. Zhang, W. Wang, S. Jamin, D. Massey, L. Zhang, Universal IP multicast delivery, Computer Networks 50 (6) (2006) 781–806.

[5] S. Lu, J. Wang, G. Yang, C. Guo, SHM: scalable and backbone topology-aware hybrid multicast, in: 16th International Conference on Computer Communications and Networks (ICCCN'07), 2007, pp. 699–703.

[6] M. Wählisch, T.C. Schmidt, Between underlay and overlay: on deployable, efficient, mobility-agnostic group communication services, Internet Research 17 (5) (2007) 519–534. URL http://www.emeraldinsight.com/10.1108/10662240710830217.

[7] M. Wählisch, T.C. Schmidt, S. Venaas, A common API for transparent hybrid multicast, IRTF Internet Draft – work in progress 03, IRTF, July 2011. URL http://tools.ietf.org/html/draft-irtf-samrg-common-api

[8] B. Niven-Jenkins, F.L. Faucheur, N. Bitar, Content distribution network interconnection (CDNI) problem statement, Internet-Draft – work in progress 02, IETF, March 2011.

[9] J. Liu, S.G. Rao, B. Li, H. Zhang, Opportunities and challenges of peer-to-peer internet video broadcast, Proceedings of the IEEE 96 (1) (2008) 11–24.

[10] M. Zhang, Q. Zhang, L. Sun, S. Yang, Understanding the power of pull-based streaming protocol: can we do better?, IEEE Journal on Selected Areas in Communications 25 (9) (2007) 1678–1694

[11] M. Castro, M. Costa, A. Rowstron, Debunking some myths about structured and unstructured overlays, in: NSDI'05: Proceedings of the 2nd Symposium on Networked Systems Design & Implementation, USENIX Association, Berkeley, CA, USA, 2005, pp. 85–98.

[12] Y. Qiao, F.E. Bustamante, Structured and unstructured overlays under the microscope: a measurement-based view of two P2P systems that people use, in: Annual Tech '06: Proceedings of the ANnual Technical Conference on USENIX, USENIX Association, Berkeley, CA, USA, 2006, pp. 341–355.

[13] C. Jennings, B. Lowekamp, E. Rescorla, S. Baset, H. Schulzrinne, REsource LOcation And Discovery (RELOAD) Base Protocol, Internet-Draft – work in progress 18, IETF, August 2011.

[14] M. Wählisch, T.C. Schmidt, G. Wittenburg, BIDIR-SAM: large-scale content distribution in structured overlay networks, in: M. Younis, C.T. Chou (Eds.), Proceedings of the 34th IEEE Conference on Local

Computer Networks (LCN), IEEE Press, Piscataway, NJ, USA, 2009, pp. 372–375.

[15] M. Castro, P. Druschel, A.-M. Kermarrec, A. Rowstron, SCRIBE: a large-scale and decentralized application-level multicast infrastructure, IEEE Journal on Selected Areas in Communications 20 (8) (2002) 100–110.

[16] L. Abeni, C. Kiraly, R. Lo Cigno, On the optimal scheduling of streaming applications in unstructured meshes, in: Proceedings of the 8th International IFIP Networking Conference, LNCS, vol. 5550, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 117–130.

[17] Y. Liu, On the minimum delay peer-to-peer video streaming: how realtime can it be?, in: Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA '07, ACM, New York, USA, 2007, pp 127–136.

[18] S. Mirshokraie, M. Hefeeda, Live peer-to-peer streaming with scalable video coding and networking coding, in: Proceedings of the first Annual ACM SIGMM Conference on Multimedia Systems, MMSys '10, ACM, New York, USA, 2010, pp. 123–132.

[19] D. Carra, R. Lo Cigno, E.W. Biersack, Stochastic graph processes for performance evaluation of content delivery applications in overlay networks, IEEE Transaction on Parallel Distribution Systems 19 (2) (2008) 247–261.

[20] S. Ratnasamy, M. Handley, R.M. Karp, S. Shenker, Application-level multicast using content-addressable networks, in: J. Crowcroft, M. Hofmann (Eds.), Networked Group Communication, in: Third International COST264 Workshop, NGC 2001, London, UK, November 7–9, 2001, Proceedings, vol. 2233 of LNCS, Springer-Verlag, London, UK, 2001, pp. 14–29.

[21] S.Q. Zhuang, B.Y. Zhao, A.D. Joseph, R.H. Katz, J.D. Kubiatowicz, Bayeux: an architecture for scalable and fault-tolerant wide-area data dissemination, in: J. Nieh, H. Schulzrinne (Eds.), Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV '01), ACM, New York, USA, 2001, pp. 11–20.

[22] M. Castro, P. Druschel, A.-M. Kermarrec, A. Nandi, A.I.T. Rowstron, A. Singh, SplitStream: high-bandwidth content distribution in cooperative environments, in: Peer-to-Peer Systems II, in: M.F. Kaashoek, I. Stoica (Eds.), Second International Workshop, IPTPS 2003 Berkeley, CA, USA, February 21–22, 2003 Revised Papers, vol. 2735 of LNCS, Springer–Verlag, Berlin Heidelberg, 2003, pp. 292–303.

[23] A. Rowstron, P. Druschel, Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems, in: IFIP/ACM International Conference on Distributed Systems Platforms (Middleware), LNCS, vol. 2218, Springer-Verlag, Berlin Heidelberg, 2001, pp. 329–350.

[24] M. Castro, M.B. Jones, A.-M. Kermarrec, A. Rowstron, M. Theimer, H. Wang, A. Wolman, An evaluation of scalable application-level multicast built using peer-to-peer overlays, Proceedings of the Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom 2003), vol. 2, IEEE Computer Society, Washington, DC, USA, 2003, pp. 1510–1520.

[25] M. Wählisch,, T.C. Schmidt, G. Wittenburg, Broadcasting in Prefix Space: P2P Data Dissemination with Predictable Performance, in: M. Perry, H. Sasaki, M. Ehmann, G.O. Bellot, O. Dini (Eds.), Proceedings of the Fourth International Conference on Internet and Web Applications and Services (ICIW'09), IEEE Computer Society Press, Los Alamitos, CA, USA, 2009, pp. 74–83.

[26] S. Birrer, F.E. Bustamante, The feasibility of DHT-based streaming multicast, in: MASCOTS '05: Proceedings of the 13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, IEEE Computer Society, Washington, DC, USA, 2005, pp. 288–298.

[27] S. Voulgaris, M. van Steen, Hybrid dissemination: adding determinism to probabilistic multicasting in large-scale P2P systems, in: R. Cerqueira, R. Campell (Eds.), Middleware 2007, LNCS, vol. 4834, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 389–409.

[28] J.C.I. Chuang, M.A. Sirbu, Pricing multicast communication: a cost-based approach, Telecommunication Systems 17 (3) (2001) 281–297. presented at the INET'98, Geneva, Switzerland, July 1998.

[29] G. Phillips, S. Shenker, H. Tangmunarunkit, Scaling of multicast trees: comments on the Chuang–Sirbu scaling law, in: Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM'99), ACM Press, New York, NY, USA, 1999, pp. 41–51.

[30] P.V. Mieghem, G. Hooghiemstra, R. van der Hofstad, On the Efficiency of Multicast, IEEE Transactions on Networking 9 (6) (2001) 719–732.

[31] R.C. Chalmers, K.C. Almeroth, On the Topology of Multicast Trees, IEEE Transactions on Networking 11 (1) (2003) 153–165.

[32] C. Adjih, L. Georgiadis, P. Jacquet, W. Szpankowski, Multicast tree structure and the power law, IEEE Transactions on Information Theory 52 (4) (2006) 1508–1521.

[33] M. Janic, P. Van Mieghem, On properties of multicast routing trees, International Journal on Communication Systems 19 (1) (2006) 95–114.

[34] S. Fahmy, M. Kwon, Characterizing overlay multicast networks and their costs, IEEE/ACM Transactions on Networking 15 (2) (2007) 373–386.

[35] P. Maymounkov, D. Mazières, Kademlia: A peer-to-peer information system based on the xor metric, in: Proceedings of the 1st International. Workshop on Peer-to Peer Systems (IPTPS '02), Cambridge, MA, USA, 2002, pp. 53–65.

[36] D. Wischik, M. Handley, M.B. Braun, The resource pooling principle, SIGCOMM Comput. Commun. Rev. 38 (5) (2008) 47–52.

[37] M. Cha, P. Rodriguez, S. Moon, J. Crowcroft, On next-generation Telco-Managed P2P TV Architectures, in: Proceedings of the VII. International Workshop on Peer–to–Peer Systems (IPTPS'08), 2008.

[38] M. Abramowitz, I. Stegun, Handbook of Mathematical Functions, Dover Publications, New York, 1964.

[39] M. Wählisch, Scalable Adaptive Group Communication on Bi-directional Shared Prefix Trees, Technical Report, TR-B-08-14, Freie Universität Berlin, Department of Mathematics and Computer Science, Berlin (September 2008). URL http://www.inf.fu-berlin.de/inst/pubs/tr-b-08-14.abstract.html.

[40] A. Varga, OMNeT++. Discrete Event Simulator System, http://www.omnetpp.org (2009).

[41] I. Baumgart, B. Heep, S. Krause, OverSim: A flexible overlay network simulation framework, in: M. Faloutsos et al. (Eds.), Proceedings of the 10th IEEE Global Internet Symposium, IEEE Computer Society, Washington, DC, USA, 2007, pp. 79–84.

[42] T.S.E. Ng, H. Zhang, Predicting internet network distance with coordinates-based approaches, in: 21st IEEE Conference on Computer Communications (INFOCOM'02), IEEE Computer Society, Washington, DC, USA, 2002.

[43] D. Stutzbach, R. Rejaie, Understanding churn in Peer-to-Peer networks, in: Proceedings of the 6th ACM SIGCOMM Internet Measurement Conference (IMC), ACM, New York, NY, USA, 2006, pp. 189–202.

Matthias Wählisch is a PhD candidate and research assistant at the Freie Universität (FU) Berlin. He studied computer science and contemporary German literature at FU Berlin, where he completed his diploma thesis on structured hybrid multicast routing. Matthias continues his research at the Computer Systems & Telematics group there, and is also with the INET research team at HAW Hamburg. He started professional activities at the networking group of the computer centre of FHTW Berlin while at high school. Matthias is the co-founder of link-lab, a start-up company in the field of next generation networking. His major fields of interest lie in efficient and reliable Internet communication. This includes the design and analysis of networking protocols, with a special focus on mobility and group communication in underlay and overlay, as well as Internet topology measurement and analysis.

Thomas C. Schmidt is professor of Computer Networks & Internet Technologies at Hamburg University of Applied Sciences (HAW) and leads the Internet Technologies research group (INET) there. Prior to moving to Hamburg, he headed the computer centre of FHTW Berlin for many years. Thomas studied mathematics and physics at Freie UniversitSt Berlin and University of Maryland. His current interests lie in next generation Internet (IPv6 & beyond), mobile multicast and multimedia networking, as well as XML-based hypermedia information processing. He serves as co-editor and technical expert in many occasions and is actively involved in the work of IETF. Thomas is co-chairing the IRTF Scalable Adaptive Multicast Research Group.

Georg Wittenburg is a postdoctoral researcher at the joint INRIA / École Polytechnique HIPERCOM team located at the Laboratoire d'Informatique de l'+cole Polytechnique (LIX). He received his Ph.D. in Computer Science from from Freie UniversitSt Berlin, Germany, in 2010, and his M.Sc. and B.Sc. from the same university in 2005 and 2003 respectively. Before joining the HIPERCOM team in November 2010, he spent four years working as a research assistant at the Computer Systems & Telematics group at Freie Universität Berlin. His research is focused on wireless ad hoc networking, in particular on the topics of distributed service provisioning, event detection in wireless sensor networks, and accuracy of wireless network simulations.