Between Underlay and Overlay: On Deployable, Efficient, Mobility-agnostic Group Communication Services *

Matthias Wählisch[1,2] and Thomas C. Schmidt[1]
{waehlisch, t.schmidt}@ieee.org
[1]HAW Hamburg, Department Informatik, Berliner Tor 7, 20099 Hamburg, Germany
[2]link-lab, Hönower Str. 35, 10318 Berlin, Germany

Abstract

Multicast communication services are one of the longest debated issues in the 30 years history of the Internet. Disagreement over innumerous approaches and solutions to the IP host group model has led to a strongly divergent state of deployment. Stimulated by the need of applications alternative multicast mechanisms have been developed. P2P technologies have enabled group distribution on the application or service middleware layer. A significantly simplified routing approach gave rise to the lean, source-specific multicast in IP. Henceforth the debate developed about which approach proves suitable for providing the superior capabilities towards a scalable, efficient *and* deployable group communication service.

This paper discusses problems, requirements and current trends for deploying group communication in real-world scenarios from an integrated perspective. We introduce *Hybrid Shared Tree*, a new architecture and routing approach to combine network- and subnetwork-layer multicast services in end-system domains with transparent, structured overlays on the inter-domain level. This hybrid solution is highly scalable and robust and offers provider-oriented features to stimulate deployment. Furthermore, a straight forward perspective is indicated for a mobility-agnostic routing layer for future use.

Keywords: inter-domain multicast routing, overlay multicast, Hybrid Shared Tree, mobile multicast, multimedia group conferencing, resource discovery, autonomous networks

1 Introduction

When the Internet was still in its early, premature state of development, the idea arose to extend unicast capabilities by a multicast group service (Aguilar 1984). Multicast communication techniques have been under debate since Deering (1989) introduced the host group model to the Internet layer. Until today, the initial approach of Any Source Multicast (ASM) routing remained hesitant to spread beyond limited, controlled environments. Meanwhile, new demands for group communication are arriving with increasing intensity, e.g., multimedia conferencing in mobile environments, service discovery in service-oriented architectures or self-configuring components in autonomous networks.

However, the deployment of IP multicast in general has been slow over the past 15 years, even though all major router vendors and operating systems offer a wide variety of implementations to support multicast (Diot et al. 2000). A fundamental dispute arose on multicast concepts in the end-to-end design principle by Saltzer et al. (1984), questioning the appropriate layer, where group communication service should reside on. For several years, the focus of the research community turned towards application layer multicast, and only recently reconsidered the relevance of IP layer

*This work is supported by the German Bundesministerium für Bildung und Forschung within the project *Moviecast* — http://moviecast.realmv6.org.

multicast. Naturally, the debate on "efficiency versus deployment complexity" overlapped into the mobile multicast domain (Garyfalos et al. 2004). In retrospect, this discourse may well be taken as an expression of ambivalence when considering multicast design concepts for the Internet as a whole.

In the past, vendors and technicians, trying to promote multicast functionality, focused their marketing arguments on network efficiency and unintentionally degraded its paradigm to unidirectional, broadcast-type services. Since then multicast suffers from a reputation of being merely useful for non-interactive, “archaic” mass distribution tasks. Large-scale interactive group applications like massive multiplayer games, conferencing in restricted regimes or complex collaborative environments have only recently drawn attention.

In parallel, mobile multimedia group communication appeared as an emerging applications field. Multicast services in mobile environments may soon become indispensable, when multimedia distribution services such as DVB-H and IP-TV will develop as strong business cases for portables. As IP mobility will unfold dominance and efficiency in costly radio environments will show a larger impact, the evolution of multicast protocols will naturally follow mobility constraints (Biersack 2005).

To all these the Internet uniquely provides the benefit of a globally scalable, dynamic group communication service. Consequently it is not surprising to see a large variety of recent concepts and ideas, but also a range of thorough analysis’ concerning efficiency and deployability of stationary and mobile multicast schemes.

In this paper we summarise and extend these discussions and pursue a discourse on combined efforts of IP layer and overlay multicast solutions. We introduce Hybrid Shared Tree, an architecture and protocol for inter-domain multicast. This scheme inherits major efficiency from the IP layer, while sustaining ease in deployment and infrastructure-transparency from selected group distribution in the overlay. In particular, the core routing of our proposed solution remains mobility-agnostic. Hence it is shown that an easy, deployable, global infrastructure for supporting mobile multicast falls into the realm of possibility.

This paper is organised as follows. We discuss the current state of the art of IP multicast, its potentials, problems and solutions in section 2. A brief review on overlay multicast technologies is given in the subsequent section 3. Section 4 introduces our Hybrid Shared Tree approach to combine underlay and overlay techniques in a symbiotic architecture. Finally, section 5 is dedicated to conclusions and an outlook.

2 IP Multicast – An Elaborate Deployment Challenge

2.1 Intra-Domain Multicast

A large number of today’s enterprise networks provide multicast services within their local domains to facilitate administrative tasks as well as shared group applications. This is indicated by the wide availability of intra-domain multicast protocols s. a. IGMP, MLD, DVMRP, PIM-DM/SM/SSM, Bidir-PIM in routers and end systems and the fairly uniform presence of multicast capabilities in lower layer protocols – i.e., in IEEE 802.3 Ethernet, 802.11 WLAN, 802.16 WIMAX or in 3GPP MBMS and DVB-H.

This deployment success on the one hand can be attributed to the large number of nodes installed in common enterprise domains, which immediately profit from multicast distribution services, on the other hand complex routing services are much easier established, controlled and also restricted within a single administrative domain. Multicast admission and scoping in general and prevention of misuse in DDoS attacks in particular can be managed with reasonable effort at intra-domain level, while these tasks turn into critical challenges in an inter-provider context. Furthermore higher spare capacities of routers and systems at Internet edges allow for concurrent operation of multicast management burdens, while at the same time scaling limitations inherent to

most of the present protocols remain invisible within most enterprise networks.

Nevertheless, intra-domain multicast routing is not considered complete, but remains an active research field. The major reason for discontent results from the handling of data-driven multicast distribution states. They are required at the routing layer, which break the paradigm of stateless forwarders and opens the door for flow-state attacks directed against the routing infrastructure. Recent work on bi-directional PIM by Handley et al. (2007) has advanced this debate by utilising a *group-specific* shared tree within limited domains. States for this bi-directionally operational, but not uniformly optimal distribution tree are established at group creation and thus fully decouple from the data plane.

2.2 Inter-Domain Multicast

In contrast to the success at an internal level, inter-domain multicast deployment largely failed. Inexplicit benefits, complexity and scalability issues with multicast BGP-4 extensions, robustness and security concerns, as well as the threat of intransparently interwoven service models kept ISPs from adding multicast burdens onto their already notoriously overloaded core routers. At present the key issues for inter-domain multicast deployment may be seen as:

Control on groups will allow ISPs to explicitly restrict (or charge for) distribution services, and thus must be considered an important part of a consistent business model;

Controlled load on backbone routers in terms of table spaces, computational and signalling demands will be required for a predictable service quality;

Scalable protocols build the essential foundation for a large-scale deployment;

State aggregation within shared trees will be a technical demand to control the router load;

Forward routing will be of vital importance due to asymmetric backbone routes.[1] Many multicast routing protocols depend on Reverse Path Forwarding and thereby erroneously assume symmetric routes;

Explicit benefits will provide the reasons for ISPs to deploy multicast. Aside from a simple gain analysis, arising applications or new, e.g., mobile services may stimulate appropriate business cases for multicast.

Recent advancements led IP multicast routing in divergent directions. Source Specific Multicast (SSM) (Bhattacharyya 2003, Holbrook & Cain 2006) broke with Deering's open host group model to achieve greatly simplified, domain-transparent routing. In contrast to Any Source Multicast (ASM), optimal (S,G) multicast source trees are constructed immediately from (S,G) subscriptions at the client side, without using network flooding or rendezvous points. Source addresses are to be acquired by out of band channels, limiting its applicability to service–aware parties. By this lack of generality SSM remains unsuitable for self–configuration tasks of distributed systems. Moreover the single source model does not allow for state aggregation in shared trees, while the common PIM-SSM routing by Fenner et al. (2006) uses Reverse Path Forwarding for Internet backbone traversal.

BGMP (Thaler 2004) at the Internet backbone attains a somewhat complementary role of Bidir-PIM by supporting bidirectional shared trees between domain-level rendezvous points, thereby overcoming limitations of scalability. However, BGMP continues to rely on route symmetry throughout the Internet backbone.

Only recently, at SIGCOMM, Ratnasamy et al. (2006) again urged for the adaptation of an any source multicast service on the IP layer. The authors propose BGP extensions to exchange group membership announcements decoupled from multicast route discovery. Routing follows a

[1] In a general attempt Paxson (1997) analyzed 40,000 end–to–end paths and identified half of them as asymmetric.

forward path approach achieved by a tree-based source routing on top of BGP. As BGP routing tables are unaware of global contexts, the authors need to encode the entire distribution tree within forwarded packets. While incorporating original, valuable ideas, this monolithic Free Riding Multicast (FRM) protocol suffers from the drawbacks of not only requiring a complete change of the BGP layer, but also placing the heavy burden of evaluating the distribution tree in the Internet core and performing correspondent source routing. Our proposal presented in section 4 will transform some of these ideas into a solution, which remains transparent with respect to the Internet core.

2.3 Multicast Benefits

Complexity and deployment cost of network layer multicast services need to be compared with benefits gained from the simplicity and efficiency of group applications in network utilisation. Though obvious, network efficiency gained from multicast data distribution has not been quantified for a long time, leaving providers with vague expectations for the outcome of multicast service provisioning.

Only recently, multicast distribution trees have been thoroughly studied with regards to network efficiency. Grounded on empirical observations Chuang & Sirbu (2001) proposed a scaling power–law for the total number $L_M(m)$ of links in a multicast shortest path tree with m receivers of the form

$$L_M(m) \approx < L_U > m^k,$$

where $< L_U >$ represents the average number of unicast hops taken by a message between uniformly chosen nodes in the corresponding network of M nodes. The authors consistently identified the scale factor to attain the independent constant $k = 0.8$. The validity of such universal, heavy–tailed distribution suggests that multicast shortest path trees are of self–similar nature with many nodes of small, but few of higher degrees. Consequently, trees would rather be shaped tall than wide. Providers thus could count on a relative gain in network resource consumption, which uniformly scales in group size as $m^{0.2}$.

Subsequent empirical and analytical work of Phillips et al. (1999), Van Mieghem et al. (2001), Chalmers & Almeroth (2003), Adjih et al. (2006), Janic & Van Mieghem (2006) has debated the applicability of the Chuang and Sirbu scaling law. Schmidt & Wählisch (2006) analyzed its consequences for multicast mobility. Van Mieghem et al. (2001) proved that the proposed power law cannot hold in general, but is indeed a valid approximation for moderate receiver numbers and the current Internet size $N = 10^5$ core nodes.

2.4 Mobile Multicast

Multicast mobility management has to accomplish two distinct tasks: handover operations for mobile listeners and senders. While many solutions exist for roaming receivers (Romdhani et al. 2004, Schmidt & Wählisch 2007), very few schemes have been specified for mobile multicast sources. Following a handover, multicast data reception can be fairly easily regained by a remote subscription approach, cf. MIPv6 (Johnson et al. 2004), possibly expedited by agent–based proxy schemes, cf. Schmidt & Wählisch (2005). In contrast, a multicast sender either defines the root of a source-specific shortest path tree (SPT), distributing data towards a rendezvous point or receivers, or it forwards data directly down a shared tree, e.g., via encapsulated PIM register messages. Aside from tunnelling or shared trees, forwarding along source-specific delivery trees will be bound to a topological network address due to reverse path forwarding (RPF) checks. At the same time a mobile sender must not change source address while re-associating in a different network, since addresses are associated on the application layer, e.g., with RTP media streams.

Within intra-domain multicast routing, the employment of shared trees may considerably relax mobility-related complexity. Relying upon a static rendezvous point, a mobile source may continuously submit data by encapsulating packets with its previous topologically correct or home

source address. Constraints even diminish, when bi-directional PIM is used. Intra-domain mobility is transparently covered by bi-directional shared trees, which are built from a 'virtualised rendezvous point', eliminating the need for tunnelling data to reach the rendezvous point.

However, issues arise in inter-domain multicast scenarios, whenever notification of source addresses is required between distributed instances of shared trees. Problems increase with Source Specific Multicast operated on the IP–layer, as it requires active subscription to contributing sources, thereby relying on topologically correct addresses. On the occurrence of handovers and in the presence of source filters, any mobile SSM routing protocol is required to transform a given (S, G) state into (S', G), while listening applications continue to receive multicast data streams admitting a persistent source address.

Facing multicast deployment problems, it is desirable that any solution to mobile multicast should leave routing protocols unchanged. Mobility management in such deployment-friendly schemes should preferably be handled at the Internet edges, preserving the core routing infrastructure in mobility-agnostic condition. Facing the current state of proposals, the urgent search for such a simple, infrastructure-transparent solution remains, even though there are reasonable doubts about whether this can be achieved for SSM.

In the following section 4 we will demonstrate how a hybrid shared tree scheme may be used to design a mobility-agnostic global multicast routing solution.

3 Overlay Multicast

3.1 The Structured Peer-to-Peer Approach

In recent years the Internet community experienced two significant disruptions. The advent and overwhelming success of Napster and successors from 1999 on demonstrated an imperative desire of Internet users to take advantage of transparent end-to-end application services. The Internet, originally designed as a logical end-to-end overlay on top of heterogeneous physical networks, apparently had failed to serve these needs in its current server-centric and NAT-burdened state of deployment.[2] In the year 2001, when Napster failed legally and early versions of Gnutella broke down technically, proposals for using an abstract name space for combining nodes and content emerged, which organizes within distributed hash tables. The introduced solutions admit the routing geometry of rings as in Chord, trees as in Tapestry and Pastry or a d-dimensional toroidal geometry as in CAN. Their common concept of distributed indexing (Plaxton et al. 1997), which had been initially developed for distributed memory computer architectures, stimulated many of ideas and continues to inspire routing in Manet systems, as well as to heat up the debate on a 'clean slate' reinvention of the Internet.

The DHT substrate Pastry by Rowstron & Druschel (2001), which will be used in this work, combines prefix orientation with topology awareness at the routing layer. Starting from the alphabet of an arbitrary base 2^b, routing proceeds according to a longest prefix match and assures a hopwise increasing prefix coincidence. It terminates after at most $\log_{2^b}(n)$ steps, with $n =$ number of DHT nodes. Under-determined neighbour specification in prefix space is used for a proximity selection of next-hop underlay nodes, which shows enhanced efficiency for higher degrees of freedom, i.e., for shorter prefixes. In combining these two route optimisation mechanisms Pastry arrives at a fairly uniform relative delay penalty factor of about 2, independent of overlay sizes.

3.2 DHT–based Multicast

Derived from structured peer–to–peer routing, a collection of group communication services has been developed, with the aim of seamless depoyability as application layer or overlay multicast.

[2] Characteristically, an ongoing combat arose of P2P suppression on the infrastructure management side and barrier evasion on the application layer.

Among the most popular approaches are multicast on CAN by Ratnasamy et al. (2001), Bayeux by Zhuang et al. (2001), as derived from Tapestry, and Scribe by Castro et al. (2002) or SplitStream (Castro, Druschel, Kermarrec, Nandi, Rowstron & Singh 2003), which inherit their distributed indexing from Pastry. Particularly optimised derivations have been designed for MANETs, cf. Gui & Mohapatra (2003).

Approaches to multicast distribution in the overlay essentially branch in two algorithmic directions. In the first case, DHTs are used to generate a structured sub-overlay of group members, which is then flooded with multicast packets. This mechanism underlies multicast on CAN. In the other, a distribution tree is erected within the full overlay, to be used as a shared or source specific tree. The latter schemes are used in Scribe and SplitStream, where a rendezvous node is chosen from group key ownership, or in Bayeux.

The performance of DHT–based multicast has been thoroughly studied by Castro, Jones, Kermarrec, Rowstron, Theimer, Wang & Wolman (2003) with the comparative focus on tree–based and flooding approaches built onto CAN and Pastry. The separate construction of mini-overlays per group as needed for a selective flooding incurred significant overhead. In addition, flooding was found to be outperformed by forwarding along trees, where a shared group tree combined with proximity-aware routing as in SCRIBE could minimise the overlay delay penalty down to a factor of two. For the sake of completeness we mention that overlay multicast concepts concurrently exist for unstructured peer-to-peer approaches. They operate at lower algorithmic complexity, but show significantly higher efforts in coordinative signalling and thereby admit performance measures too far from native multicast to be of interest in this discourse.

3.3 Discussion

Structured peer-to-peer systems offer multicast services in an infrastructure-agnostic fashion. They are reasonably efficient and scale over a wide range of group sizes. However, they do not allow for layer 2 interactions and thus do not facilitate unrestricted scaling in shared end system domains. Stability issues for tree-based overlay multicast under churn arise as well, as the departure of branching nodes close to the root may have disastrous effects on data distribution. These drawbacks may be mitigated by hybrid approaches, where overlay multicast routing only takes place among selected nodes, which are particularly stable and form a virtual infrastructure. Similar initial propositions have recently been introduced to IRTF (Buford 2007). Such adaptive schemes of cooperative routing in underlay and overlay bear the potential to optimise stability and performance, while sustaining ample flexibility for deployment.

The performance gap between IP and application layer multicast widens, when mobility is introduced. Frequent handoffs and topological re-arrangements degrade the stability of distribution trees and the efficiency of proximity selection. Garyfalos & Almeroth (2005) derived from fairly generic principles efficiency measures for source specific multicast in different metrics. Overlay trees uniformly admitted degradations up to a factor of four over native IP layer multicast in the presence of MIPv6 mobility management. To overcome mobility obstacles, the authors introduce the Intelligent Gateway Multicast (IGM), which assists in reactive handovers at the network access. Although designed from a different perspective, this architectural approach is similar to our proposal in section 4.

4 Hybrid Shared Tree

4.1 Basic Design Principles

In this section we will introduce Hybrid Shared Tree (HST), a hybrid architecture designed to enable global multicast peering at the ISP or enterprise level, while sustaining end system transparency in utilising well-established group distribution services.

The basic concept of HST preserves multicast routing and lower layer packet transmission within domains as discussed in section 2.1, while bridging the inter-domain gap with the help of a structured overlay network to overcome the deployment problems discussed in section 2.2. This approach differentiates the end-to-end design argument Saltzer et al. (1984) with respect to the inhomogeneous nature of the global Internet: While customer-oriented end system networks, which are mainly built on top of multicast enabled subnetwork technologies, do significantly profit of utilising network layer multicast services, the flow-oriented transition networks of the Internet core do not.

In combining a well established DHT with a new overlay multicast routing scheme, we address in particular the following design objectives:

- Provide scalability, robustness and inter-domain transparency for shared distribution trees;
- Detach multicast routing from the Internet core and restrict the backbone infrastructure to plain unicast forward routing;
- Decouple group membership registration from route discovery;
- Decouple multicast state management from the data plane;
- Grant control on group admission to local operators;
- Open a lightweight deployment perspective for mobile multicast services.

This overall design of interconnecting end system domains on the basis of a structured overlay gives full multicast admission control to local operators and may be interpreted as globally distributed service peering. It will enable inter-domain shared trees to multicast group services, which remain invisible to the Internet core, while inheriting the full potential of scalability, self-organisation, redundancy and error resilience from the distributed hash table algorithm in use.

4.2 Architectural Overview

The Hybrid Shared Tree architecture follows the lines of the evolutionary construction scheme of the Internet. Its focus originates from a customer network or an ISP domain, where multicast services are locally deployed. Multicast service exchange is then expected to be implemented like unicast peering, in a dedicated but isolated step. It will operate following the activation of a gateway service, which interconnects the local multicast routing with the distributed peering on the structured overlay. Note that a separation of inter-domain multicast from unicast routing will lead not only to a simplified, more stringently structured approach, but will also segregate malfunctions due to misconfiguration or component overload.[3]

We introduce Inter-domain Multicast Gateway (IMG) as a new architectural entity, which provides a gateway function between the overlay it is a member of, and the multicast routing at the intra-domain underlay that it resides in, cf. figure 1. Those gateways will participate in multicast traffic originating from its residential network, which it will forward into the overlay according to the distributed multicast receiver domains of this group, and will also advertise group membership and receive data according to any subscription from its domain. On the overlay the IMGs will jointly operate a distributed hash table, which is chosen to be Pastry (Rowstron & Druschel 2001) due to its proximity-aware prefix-based routing. Note that our multicast distribution service thereupon will then differ from SCRIBE (Castro et al. 2002), and will follow a new routing scheme, as we will describe below.

The IMG function may be positioned anywhere within the multicast domain, but need to provide a protocol interface to the locally deployed multicast routing. To avoid zigzag transmission,

[3] Caused by experiences with early PIM-SM implementations, there is a common fear of multicast to degrade the unicast forwarding plane.

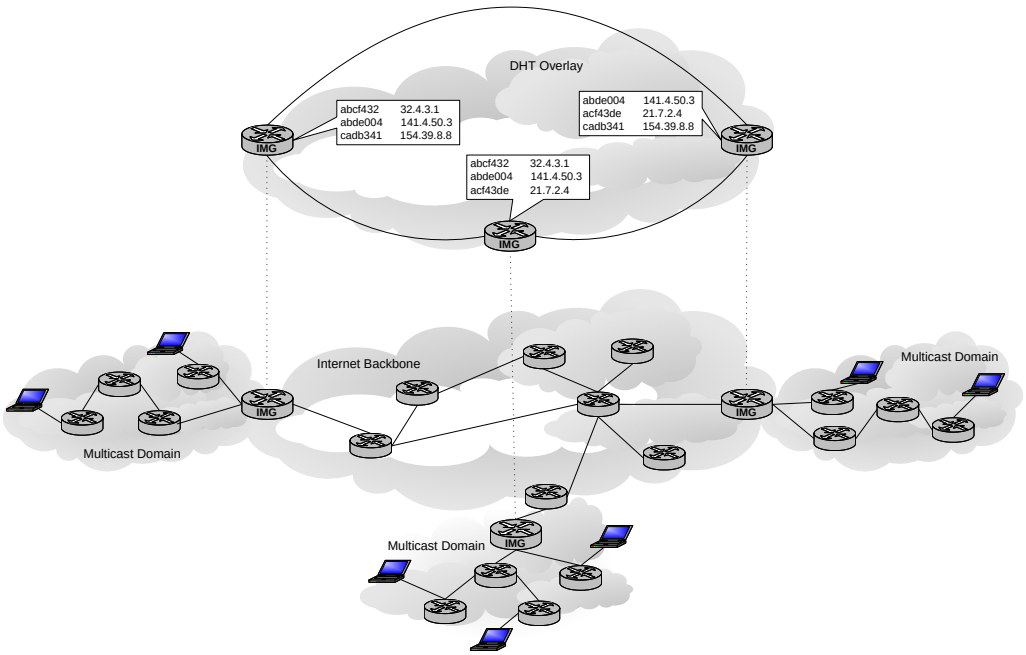

Figure 1: Hybrid Shared Tree Architecture

the IMG may be situated at the domain border router, though, in the example of a PIM-SM or Bidir-PIM domain, the IMG could also be colocated with the rendezvous point or the rendezvous address. Note that the IMG function may be built as a dedicated system entity, but on existing routers may consist of additional intelligence, as well.

Activation of inter-domain multicast gateway services requires only a small amount of selected information for bootstrapping, i.e., an arbitrary contact member of the structured overlay, authentication and authorisation credentials, if applicable. The IMG further on remains under the administrative control of the local network operator, who may restrict admission, scoping and QoS characteristics of the group traffic flowing in and out of the intra-domain. Aside from general multicast peering policies, a service provider is thus enabled to implement firewall-type of packet filters at, or co-located with, these multicast gateways.

This architecture allows for flexibility in several ways. A domain operator is enabled to connect to several multicast overlays in parallel, may choose to replicate IMGs for load balancing or redundancy purposes or may transparently take advantage of the fail-safe unicast peering realised by multi-homed network connectivity. Replication operations will be seamlessly empowered by the self-organisation capabilities of the DHT overlay. Active coordination between gateway peers requires straightforward protocol extensions, whose details are beyond the scope of this paper.

4.3 Inter-domain Group Membership Management

Any Interdomain Multicast Gateway will acquire complete knowledge of group membership requirements from its local intra-domain protocol interface. A PIM-SM group membership registration, for example, will be transmitted to the IMG via the rendezvous point: Bidir-PIM will initiate a corresponding forwarding state at group creation, always leading to per-group accumulated information on active subscriptions. Hence, the DHT could be utilised to store and offer membership states based on per-group keying, which comprises sufficient background to establish shared distribution trees and corresponds to the traditional approach of multicast in structured overlay.

However, to accomplish instantaneous, optimised forwarding throughout the overlay, it is important that group membership information is simultaneously available at each IMG. Therefore,

group registrations and de-registrations learned locally are flooded down the DHT, resulting in complete per domain group membership information at every IMG in the peering overlay. As membership updates are communicated incrementally and aggregated per domain, flooding of state changes is only required in the case of the first arriving or the last leaving group member from a multicast receiver domain.

4.4 Shared Distribution Tree

Efficient multicast packet distribution is realised by the infrastructure based on a distribution tree, where branching nodes duplicate packets. Typically, this tree spans all receivers and is rooted at the source or a rendezvous point. In contrast to traditional approaches, the HST architecture uses a prefix tree, which is built solely of IMG overlay addresses at receiver sites. This tree will serve as a source-specific distribution tree valid for any source.

From the acquired hash IDs of receiver IMGs, each DHT member will be enabled to construct a prefix tree, which connects multicast listening gateways. More detailed, actual receivers – such as all full keys – reside at leaf nodes and inner vertices will be labelled recursively with the longest common prefix of their children (cf. fig. 2). Consequently one-way branches are eliminated and the symbolic path-compressed tree is rooted at "*". Combined with prefix-based routing, we will use this structure as a bi-directional shared distribution tree later on. It is worth noting that virtual prefixes of the branching points do not correspond to existing overlay node IDs, but are dedicated to real node correspondence during routing in the Pastry overlay.

An inner vertex can be mapped to a DHT member if the label represents a prefix of the overlay node address. We will call the prefix to be *'associated with'* the node. As shown in figure 2, any overlay node ID is represented as a leaf and will be associated with all vertices along the shortest path to the tree root due to the recursive labelling. IMGs will derive identical trees in prefix space, whereas routing correspondences are to be extracted from Pastry's routing table and thereby differ from node to node. Furthermore overlay nodes need not memorise the entire group-specific tree, but are only required to persist the prefix neighbour states of all associated vertices. Correspondingly, storage and flooding of receiver IDs is subject to further optimisation on the basis of prefix-controlled forwarding, whose details remain beyond the scope of this paper.

The ab initio construction of such a minimal prefix–spanning tree is enabled solely by the homogeneous key structure of the DHT and cannot be achieved on pure IP or on an AS symbol layer. Based on the structural properties we can dynamically create a sender-specific root which is formed by all associated vertices from the IMG of an arbitrary source. Within the perspective of this virtual root, an IMG can reach all receivers and distribute multicast packets as outlined in the next section.

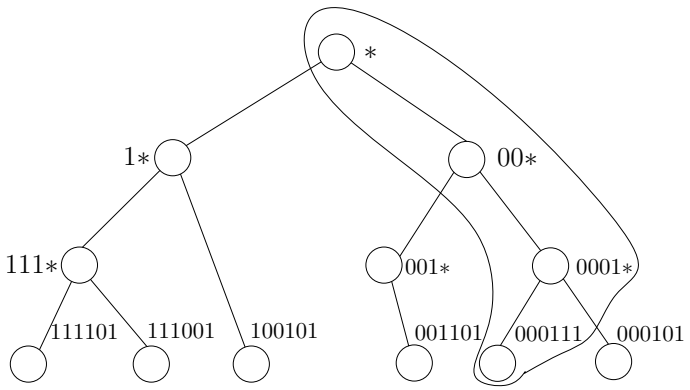

Figure 2: A Path-Compressed Prefix Tree with All Associated Vertices of Node 000111

4.5 Routing

IP layer routing within an ASM domain remains unchanged in a Hybrid Shared Tree architecture, both for sparse as well as dense mode protocols. Subscribed group traffic arriving at an IMG from a remote domain will simply be forwarded into the underlay, the gateway acting as the local source. Within the underlay an IMG operates in the role of a subscribing router being part of a shared or source-specific tree for any (admitted) group with distant receivers. Based on its information of global listeners it will thus participate in multicast distribution within its local domain.

For distributing locally-generated multicast streams to overlay receivers, IMGs will use the receiver-initiated prefix tree as derived in section 4.4 as a bi-directional shared tree. This tree will steer forwarding, substituting the role of destination address in the unicast case. To enable redundancy-free transmission along the tree, a routing node needs to determine

1. its current position within tree vertices
2. the edges currently valid for forwarding.

To account for the first requirement, packets will carry the on-tree prefix they are currently striving for as destination address in the overlay. An overlay node will be a valid receiver or *associated with* a prefix $\mathcal{P}$, whenever $\mathcal{P}$ is a prefix of its hash ID. The second task greatly simplifies by recalling the coherence property of prefixes: If a DHT node is associated with a vertex on the prefix tree, it will likewise be associated with all upper vertices on the path to the tree root. Thus upward routing on the tree degenerates to virtual hops pointing to the node itself. Routing down the tree will proceed according to longest common prefix match in the overlay. Actual forwarding in the IP underlay will be guided by Pastry's proximity-aware routing table.

In detail, to begin multicast forwarding, the initial overlay source $\mathcal{S}$ will identify its position among the tree vertices as the longest matching prefix it is associated with and will replicate data to all adjacent prefix IDs on the tree. As it shares all up-tree prefixes, this forwarding will explore all immediate branch points on upward position and restrict subsequent routing to the downward direction (cf. section 4.4). If a DHT node receives a packet with a destination prefix $\mathcal{P}$ with which it is not associated, it will simply forward it towards $\mathcal{P}$ without consulting the multicast distribution tree. Such an overlay node is just an intermediate forwarder in the DHT, i.e., a node between two destination prefixes. Further on for node IDs $\mathcal{A}$ and $\mathcal{B}$ from the DHT key space let us denote the longest common prefix by $LCP(\mathcal{A}, \mathcal{B}) = \mathcal{P}$ and the length of any given prefix $\mathcal{P}$ by $|\mathcal{P}|$. Then on reception of a multicast datagram with overlay source address $\mathcal{S}$, a DHT node will forward the packet along the bi-directional shared tree according to the following routing algorithm:

PREFIX TREE FORWARDING

```
   ▷ On arrival of packet with destination prefix C
   ▷ at DHT node of ID K
1  if C not associated with K                ▷ K is intermediate forwarder
2     then FORWARD PACKET TO C
3     else
4          for all N_i adjacent IDs on prefix tree
5               do if (LCP(C, N_i) = C)      ▷ N_i is downtree neighbour
6                     then C ← N_i
7                          FORWARD PACKET TO N_i
```

In proceeding along this line, each multicast gateway will be enabled to instantaneously submit group data down a source specific, not always minimal spanning tree of the overlay. Any forwarding step with respect to the prefix tree will transmit the multicast data closer to the receiver IMG by one or more digits. Observe that for every inner vertex label of the prefix tree at least one DHT node exists, since all leaves represent overlay nodes. The prefix tree attains the role

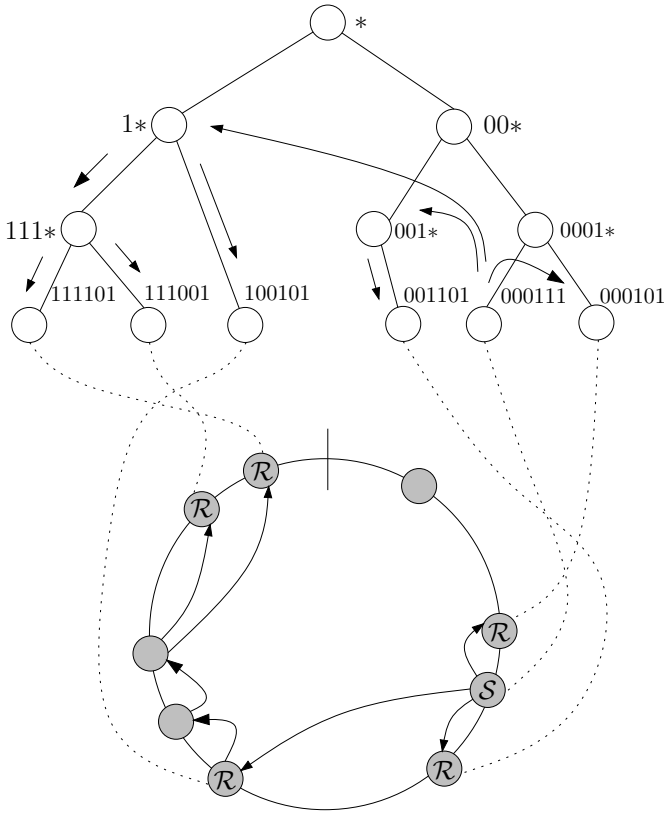

Figure 3: Hybrid Shared Tree Overlay Routing Directed by a Binary Prefix Tree

of an additional overlay directive for routing in prefix space as visualised in figure 3. It will be transferred to actual IP-layer forwarding with the help of Pastry's regular routing procedure: A prefix lookup table will return the IP address of a node, which for nearby prefixes is likely to be the final destination, or otherwise, an indirect hop with longer common prefix to destination. Note that this tree-directed overlay forwarding will lead to forward transmissions in the underlay, which resemble loose source routing along a distribution tree.

4.6 Discussion

The Hybrid Shared Tree multicast architecture attempts to combine scaling efficiency on three levels. First, native network and subnetwork layer multicast services assure optimised traffic distribution towards end systems, shield local group traffic and leave the inter-domain overlay peering unaffected by large receiver groups. Second, the hybrid architecture of super-peer type, e.g., as used in Skype, reduces peering to 'per domain' demands. This significantly decreases costs for group management and route exchange. Finally, the high scalability potential of the structured overlay is inherited from the underlying DHT, which has proved its ability to manage multicast at end-system level.

HST offers a network-transparent shared interconnect between heterogeneous multicast domains, which may operate intra-domain routing protocols of individual choices. As the overlay decouples group and state management from the forwarding plane, multicast transmission will be location-transparent wherever there are intra-domain protocols. Thus in combination with Bidir-PIM at edge domains, HST will lead to a mobility-agnostic routing environment in the sense that listeners and senders may freely move on an inter-domain scale, while a mobility-unaware routing layer will equally enable multicast services. Listeners can benefit from seamless services wherever they encounter previously established group reception.

Routing in the overlay will lead to network layer transparent packet distribution, which will

be less efficient than native IP multicasting. Overlay routing hops will remain bound by $\log_{2^b}(n)$ steps as inherited from the Pastry DHT: the delay penalty is expected to be comparable or below results for SCRIBE (cf. section 3.2), since HST features several improvements. Routing does not proceed via a fixed rendezvous point, thereby avoiding detours, bottlenecks and single points of failure, but will take various, source-dependent ways through the underlay. Additionally, the majority of intermediate hops will be led according to underdetermined prefixes, granting their degree of indetermination to Pastry's proximity selection scheme. Hereof we expect significant route optimisations to take effect with respect to routes chosen on the IP layer. Replication load on overlay forwarders is equivalent to the number of vertices adjacent within the prefix tree and depends on the prefix alphabet parameter b of Pastry. This variable parameter option leads to configurable, strictly predictable per packet processing costs of $\log_2(g)(2^b - 1)$, where g is the number of receiver domains for a given group G. Consequently, the number of neighbouring states required at any overlay member is likewise limited by $\log_2(g)(2^b - 1)$.

It is well known and unavoidable that multicast on the overlay does not scale ad infinitum as IP layer multicast does. However, having at hand logarithmically strong analytical bounds will allow for a very wide range of general deployment and a strict load control by operators or ISPs.

5 Conclusions & Outlook

The Internet uniquely offers the service of distributing data in a multicast host group model. Nevertheless, this fundamental service still suffers from a state of deployment too restrictive to allow for global dissemination of group communication services. In this paper, we discussed potentials, design concepts and pitfalls of multicast solutions, while keeping in mind both the IP layer and overlay technologies. We proposed Hybrid Shared Tree, a new hybrid architecture to interconnect multicast services between local domains, as an attempt to uncouple complexities of inter-domain multicast and unicast backbone routing.

In our hybrid approach, unlike in conventional mono-layer solutions, the well-adopted native multicast in enterprise domains is complemented by scalable, robust and transparent transmission services on structured overlays. Resting upon a newly developed routing scheme, the overlay will allow operators to deploy segregated, individually configurable multicast services with rigorously predictable system load, while leaving the inter-domain Internet unicast backbone untouched. Furthermore shared per group forwarding decouples group state establishment from the data plane, which gives rise to an option of transparent, scalable support for mobile group communication.

Further work will concentrate on a detailed protocol design, evaluation and optimisation, where large-scale experimental testing is foreseen on basis of the PlanetLab platform.

References

Adjih, C., Georgiadis, L., Jacquet, P. & Szpankowski, W. (2006), 'Multicast Tree Structure and the Power Law', *IEEE Transact. on Information Theory* **52**(4), 1508–1521.

Aguilar, L. (1984), Datagram Routing for Internet Multicasting, *in* 'Proceedings of SIGCOMM '84', ACM Press, New York, NY, USA, pp. 58–63.

Bhattacharyya, S. (2003), An Overview of Source-Specific Multicast (SSM), RFC 3569, IETF.

Biersack, E. W. (2005), 'Where is Multicast Today?', *Computer Communication Review* **35**(5), 83–84.

Buford, J. (2007), Hybrid Overlay Multicast Framework, IRTF Internet Draft – work in progress 01, SAM.
URL: *ftp://ftp.rfc-editor.org/in-notes/internet-drafts/draft-irtf-sam-hybrid-overlay-framework-01.txt*

Castro, M., Druschel, P., Kermarrec, A.-M., Nandi, A., Rowstron, A. I. T. & Singh, A. (2003), SplitStream: High-Bandwidth Content Distribution in Cooperative Environments, *in* M. F. Kaashoek & I. Stoica, eds, 'Peer-to-Peer Systems II. Second International Workshop, IPTPS 2003 Berkeley, CA, USA, February 21-22, 2003 Revised Papers', Vol. 2735 of *LNCS*, Springer–Verlag, Berlin Heidelberg, pp. 292–303.

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

Castro, M., Jones, M., Kermarrec, A.-M., Rowstron, A., Theimer, M., Wang, H. & Wolman, A. (2003), An Evaluation of Scalable Application–level Multicast Built Using Peer-to-peer Overlays, *in* 'IEEE Infocom 2003', Vol. 2, IEEE Press, Piscataway, NJ, USA, pp. 1510–1520.

Chalmers, R. C. & Almeroth, K. C. (2003), 'On the topology of multicast trees', *IEEE/ACM Trans. Netw.* **11**(1), 153–165.

Chuang, J. C. I. & Sirbu, M. A. (2001), 'Pricing Multicast Communication: A Cost-Based Approach', *Telecommunication Systems* **17**(3), 281–297. Presented at the INET'98, Geneva, Switzerland, July 1998.

Deering, S. E. (1989), Host Extensions for IP Multicasting, RFC 1112, IETF.

Diot, C., Levine, B. N., Lyles, B., Kassem, H. & Balensiefen, D. (2000), 'Deployment Issues for the IP Multicast Service and Architecture', *IEEE Network Magazine* **14**(1), 78–88.

Fenner, B., Handley, M., Holbrook, H. & Kouvelas, I. (2006), Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised), RFC 4601, IETF.

Garyfalos, A. & Almeroth, K. (2005), 'A Flexible Overlay Architecture for Mobile IPv6 Multicast', *IEEE Journal on Selected Areas in Communications* **23**(11), 2194–2205.

Garyfalos, A., Almeroth, K. C. & Sanzgiri, K. (2004), Deployment Complexity Versus Performance Efficiency in Mobile Multicast, *in* 'Proc. of Intern. Workshop on Broadband Wireless Multimedia: Algorithms, Architectures and Applications (BroadWiM)'.
URL: *http://imj.ucsb.edu/papers/BROADWIM-04.pdf.gz*

Gui, C. & Mohapatra, P. (2003), Efficient Overlay Multicast for Mobile Ad Hoc Networks, *in* 'Proc. of IEEE WCNC'03', Vol. 2, IEEE Press, Piscataway, NJ, USA, pp. 1118–1123.

Handley, M., Kouvelas, I., Speakman, T. & Vicisano, L. (2007), Bi–directional Protocol Independent Multicast (BIDIR-PIM), Internet Draft – work in progress 09, IETF.

Holbrook, H. & Cain, B. (2006), Source-Specific Multicast for IP, RFC 4607, IETF.

Janic, M. & Van Mieghem, P. (2006), 'On properties of multicast routing trees', *Int. J. Commun. Syst.* **19**(1), 95–114.

Johnson, D. B., Perkins, C. & Arkko, J. (2004), Mobility Support in IPv6, RFC 3775, IETF.

Paxson, V. (1997), 'End-to-End Routing Behavior in the Internet', *IEEE/ACM Trans. Netw.* **5**(5), 601–615.

Phillips, G., Shenker, S. & Tangmunarunkit, H. (1999), Scaling of multicast trees: comments on the chuang-sirbu scaling law, *in* 'SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication', ACM Press, New York, NY, USA, pp. 41–51.

Plaxton, C., Rajaraman, R. & Richa, A. (1997), Accessing Nearby Copies of Replicated Objects in a Distributed Environment, *in* 'Proc. of 9th ACM Sympos. on parallel Algor. and Arch. (SPAA)', ACM Press, New York, NY, USA, pp. 311–330.

Ratnasamy, S., Ermolinskiy, A. & Shenker, S. (2006), Revisiting IP Multicast, *in* 'Proceedings of SIGCOMM '06', ACM Press, New York, NY, USA, pp. 15–26.

Ratnasamy, S., Handley, M., Karp, R. M. & Shenker, S. (2001), Application-Level Multicast Using Content-Addressable Networks, *in* J. Crowcroft & M. Hofmann, eds, 'Networked Group Communication, Third International COST264 Workshop, NGC 2001, London, UK, November 7-9, 2001, Proceedings', Vol. 2233 of *LNCS*, Springer–Verlag, London, UK, pp. 14–29.

Romdhani, I., Kellil, M., Lach, H.-Y., Bouabdallah, A. & Bettahar, H. (2004), 'IP Mobile Multicast: Challenges and Solutions', *IEEE Comm. Surveys & Tutorials* **6**(1), 18–41.

Rowstron, A. & Druschel, P. (2001), Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems, *in* 'IFIP/ACM International Conference on Distributed Systems Platforms (Middleware)', pp. 329–350.

Saltzer, J. H., Reed, D. P. & Clark, D. D. (1984), 'End-to-End Arguments in System Design', *ACM Trans. Comput. Syst.* **2**(4), 277–288.

Schmidt, T. C. & Wählisch, M. (2005), 'Predictive versus Reactive – Analysis of Handover Performance and Its Implications on IPv6 and Multicast Mobility', *Telecommunication Systems* **30**(1–3), 123–142.

Schmidt, T. C. & Wählisch, M. (2006), 'Morphing Distribution Trees – On the Evolution of Multicast States under Mobility and an Adaptive Routing Scheme for Mobile SSM Sources', *Telecommunication Systems* **33**(1–3), 131–154.
URL: *http://dx.doi.org/10.1007/s11235-006-9010-4*

Schmidt, T. C. & Wählisch, M. (2007), Multicast Mobility in MIPv6: Problem Statement and Brief Survey, IRTF Internet Draft – work in progress 01, MobOpts.
URL: *http://tools.ietf.org/html/draft-irtf-mobopts-mmcastv6-ps*

Thaler, D. (2004), Border Gateway Multicast Protocol (BGMP): Protocol Specification, RFC 3913, IETF.

Van Mieghem, P., Hooghiemstra, G. & van der Hofstad, R. (2001), 'On the Efficiency of Multicast', *IEEE/ACM Trans. Netw.* **9**(6), 719–732.

Zhuang, S. Q., Zhao, B. Y., Joseph, A. D., Katz, R. H. & Kubiatowicz, J. D. (2001), Bayeux: An Architecture for Scalable and Fault-tolerant Wide-Area Data Dissemination, *in* 'Proc. of the 11th International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV 2001)', pp. 11–20.