Dynamic Cross-Domain Group Communication in Hybrid Multicast Networks

Sebastian Wölke, Thomas C. Schmidt, Sebastian Meiling
Hamburg University of Applied Sciences
iNET Research Group, Department Informatik
Berliner Tor 7, 20099 Hamburg, Germany
{t.schmidt, sebastian.woelke, sebastian.meiling}@haw-hamburg.de

Matthias Wählisch
Freie Universität Berlin
CST Research Group, Institut für Informatik
Takustr. 9, 14195 Berlin, Germany
waehlisch@ieee.org

ABSTRACT

Group communication based on multicast enables efficient one-to-many and many-to-many distribution of real-time data. Multicast communication would therefore be beneficial to popular Internet applications such as IPTV, on-line multiplayer games and audio/video conferencing, and is considered as an important network service for future CCN/ICN architectures. However, multicast exists in many flavors and technologies, and on different network layers with incompatible application interfaces and divergent states of deployment. Due to these challenges of multicast plurality, there is no multicast service available on the Internet today.

In this paper, we present a dynamic and technology-transparent group communication scheme by names. We extend H∀Mcast – a hybrid multicast architecture – with a mapping service between technology dependent addressing and an abstract naming. Therefore we show in detail the required components and discuss possible solutions.

I. INTRODUCTION

Multicast enables an efficient group communication with a great variety of application possibilities. It exists in a width range of technologies and flavors, but lacks of an universal available multicast service today. A solution is offered by a hybrid multicast network like H∀Mcast. It connects multicast islands to enable a cross-domain group communication. H∀Mcast based on the RFC *Common Multicast API for Transparent Hybrid Multicast* which presents a technology independent approach. Therefore it provides an abstract programming interface for group communication which is handled by a system-centric middleware. Multicast domains are linked over *Interdomain Multicast Gateways* (IMG) and routing the group data transparent between underlay and overlay. This is possible in a scalable manner by translating indirectly between the multicast technologies over URIs an abstract name scheme. They enable a unique and global identification of multicast groups independent of a specific technology.

In this paper we discuss how a cross-domain group communication in a hybrid multicast network can be implemented. Therefore we have classified the tasks in intra- and inter domain components. In detail we show how a mapping between arbitrary multicast URIs and group address of different technologies can look like and how group data can be scalable distributed between domains.

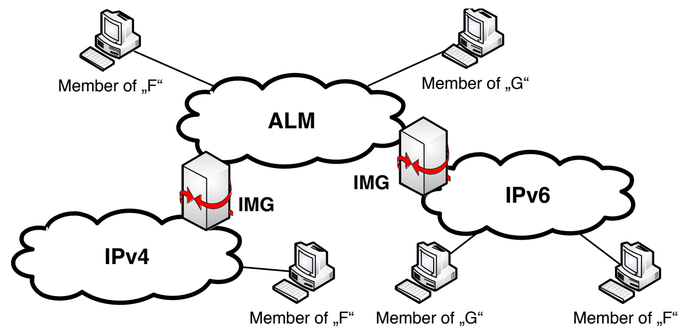


Fig. 1. A hybrid multicast network with members of 2 groups (F, G) distributed across multicast domains of different technologies. IP multicast 'islands' are connected by IMGs via an application layer multicast domain.

The remainder of the paper is organized as follows. In Section II, we give an overview on hybrid multicast networks and discuss related work. A detailed view on the concept of cross-domain hybrid multicast with its intra-domain and inter-domain functionality is presented in Sections III and IV. We conclude in Section V and give an outlook on future work.

II. BACKGROUND AND RELATED WORK

Universal Multicast [1] as well as *Island Multicast* [2] are hybrid multicast approaches that enable a universal multicast service in the Internet. They utilize IP multicast in edge networks if available and connect separated multicast islands by an application layer overlay in the core. These solutions increase efficiency and performance compared to plain overlay multicast approaches, but also improve overall availability (i.e. coverage) compared to IP multicast.

A. Hybrid Adaptive Mobile Multicast

H∀Mcast (Hybrid Adaptive Mobile Multicast) [3] extends the aforementioned solutions by a generic approach. It allows to connect networks of different multicast technologies and enables communication across heterogeneous domains, i.e., H∀Mcast is not bound to IP multicast at edges networks and overlay multicast (ALM) in the core. Multicast domains are inter-connected via IMGs that map group names to domain specific addresses and forward group data accordingly. Fig. 1 shows an example of a hybrid multicast network based on H∀Mcast, where an IPv4 and an IPv6 multicast domain are attached to an Application Layer Multicast (ALM) network.

Groups are globally identifiable, independent of a specific multicast technology based on a uniform and abstract naming scheme using URIs. Furthermore, a generic API [4] is provided to decouple application development from availability of various multicast technologies at runtime.

IMGs route group data according to static or dynamic rules based on the group state of the distribution tree. Each IMG manages the group state of its domains, which arise from the *Join* and *Source-Register* group signaling by its hosts. This group states are communicated between the IMGs to couple sender and receiver. Such a *Rendezvous Process* is required for Multicast, which uses the *Publish/Subscribe* paradigm to decouple sender and receiver for providing a multipoint connection.

In general, the architecture of a hybrid multicast network can be divided into the two main components *Intra-Domain* and *Inter-Domain*. The *Intra-Domain* components comprises of all tasks which arise within a domain and the *Inter-Domain* components handle the interaction between those.

B. Naming and Addressing of Multicast Groups

Multicast group names allow a unique identification of groups across different domains and heterogeneous technologies. They are encoded according to the following URI-based syntax that is also specified in [4]:

```
ham-scheme ":" namespace ":" group
                ["@" instantiation ]
                [":" port ]
                ["/" sec-credentials ]
```

While the *ham-scheme* is fixed to `ham`[1] (ham: hybrid adaptive multicast), the *namespace* specifies a URI prefix for valid multicast group names. *Group* identifies a multicast group uniquely within a given namespace. An (optional) *instantiation* identifies the entity that generates an instance of the group (e.g., a SIP domain or a source in SSM) and the *port* refers to a specific application. *Sec-credentials* can be used to implement optional security features such as authentication and authorization.

C. ID-Locator Mapping

The mapping in a hybrid multicast networks differs from known mapping concepts of IPv4 to IPv6 transition, .e.g, [5] or [6]. Through the separation of Identifier and Locator (*Loc-ID split*), it enables a global and technology-independent identification of groups as well as a *late-binding* of multicast technologies at runtime. The mapping between group names and group addresses is performed locally per-domain from hosts as well as IMGs. For example a host located in an IPv4 domain sends data to the group *ham:opaque:News@BBC.com*. On the technology level this group name has to be translated to an IPv4 multicast address and back to the global identifiable URI by interested hosts and IMGs. The latter may also forward group data to other attached domains based on their routing tables and by taking into account the corresponding technology specific mapping.

[1] registered with the IANA.

III. Intra-Domain Multicast

The Intra-Domain Multicast layer defines the behavior of multicast hosts and IMGs of a distinct multicast technology, which is bounded to a namespace. The Intra-Domain layer can be divided into the tasks *Data-Communication* and *Group-Management*. The *Data-Communication* provides a uniform mapping from group names to group address for the respective multicast technology. The *Group-Management* discovers multicast sources and keeps track of group memberships in a domain. For this, the group signaling of multicast senders and receivers are continually evaluated. Depending on the deployed multicast technology, necessary Join, Leave, or Source-Register messages already exist or otherwise have to be added by additional protocols.

A. Data-Communication

For all participations of a domain, Data-Communication provides a uniform mapping from group name to group address. This mapping is needed by hosts and IMGs to send and receive multicast group data. However, a reverse mapping is required for the reception of group data. On the one side, this is trivial for hosts, as the group was been subscribed before, and the mapping is already known. On the other side, IMGs need to learn the reverse mapping explicitly or derive it from the group name. Mapping concepts for the Data-Communication can be distinguished in stateless and stateful solutions. Furthermore they differ in scalability and in the assistance of the reverse mapping.

Centralized: A central mapping service can be placed in a multicast domain by implementing a single mapping database. It takes care of a coherent mapping of group names to addresses and answers mapping requests of hosts and IMGs - it also provides a reverse mapping. This approach is applicable on every multicast technology and uses a distinct communication channel, that has to implemented by or provided to all domain members. However, this mapping scheme is stateful and scales in the best case linearly with the number of hosts and used multicast group names. Additionally it heightens latency on subscribing groups, as the technology specific address translation has to be requested first.

Distributed: A decentralized approach is base on a voting procedure among all domain members to determine valid mappings of group names to addresses. It utilizes the multicast technology of the respective domain to communicate in a predefined and well-known group. In this group, hosts can inform each other about available mappings, make proposals, or report (and resolve) collisions. Similar to a centralized approach it exhibits higher latency on subscribing groups due to the communication intense mapping process – making it only suitable for small networks. On the other side, the mapping service does not have to be explicitly started and announced. Besides that, it uses the multicast technology itself and thus has a higher reliability than a centralized solution.

Hash Function: The mapping of a group name to an address can be solved implicitly by every group member using a prearranged hash function. This has the advantage that the group address can be determined very quickly. However, a hash function requires an underlying multicast technology with a (fairly) large address space to ensure valid mappings. For

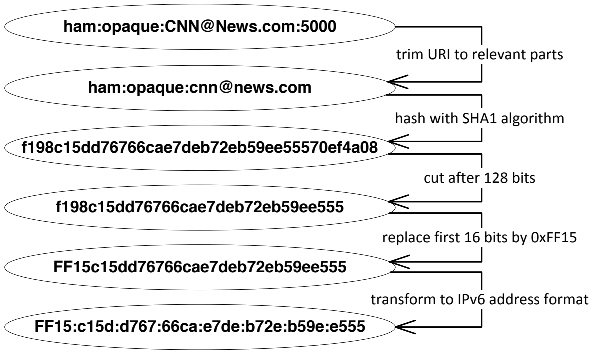


Fig. 2. Transformation of a group name (URI) to an IPv6 multicast address.

example, in an IPv4 domain, the low cardinality of the IPv4 multicast namespace will lead to a high probability of group address collisions. A collision of two groups on the technology level results in receiving unwanted group data and causes unnecessary load at hosts and on the network. Furthermore, a hash function does not provide a reverse mapping, wherefore an additional technique is required to match group address back to group names.

In the following example we map group names (URIs) to valid IPv6 multicast addresses: Therefore, the URI will be hashed as shown in Fig. 2. First, the group URI has to be shorten to the parts, which are relevant for global identification. The *scheme*, the *group*, and if exists the *instantiation* make the URI globally unique. Other parts such as *port* and *sec-credentials* provide only additional information to handle the URI in a proper manner. Second, after the group URI is hashed, the first 16 Bits are replaced by *0xFF15* (see also [7]) to match a dynamically assigned site-local IPv6 multicast address.

Canonical Mapping: It embeds technology specific addresses in URIs. This method is stateless, invertible, and enables a compatibility to native hosts, which do not use the *Common Multicast API*. For example, the IPv4 address 239.99.99.99 with port 5000 can be embedded in a group URI using the `ip` namespace: `ham:ip:239.99.99.99:5000`. In multicast domains that match the namespace, the technology specific information embedded in a URI can be used to derive a mapping. If URIs are routed from one to another IPv4 domain, the technology specific information can be used to create a mapping. Otherwise, the URI is handled like an arbitrary group name using the domain mapping service. Note: the canonical mapping can be used in a parallel to other mapping services, but must not collide with them.

B. Group Management

For each attached domain an IMG has to know, which groups are subscribed and for which groups senders exist. Furthermore, IMGs need the reverse mapping to match received group data to respective group names. These states are stored and managed by the IMGs as a soft state in a group management database. They are acquired through group signaling protocols of a multicast technology (e.g. IGMP [8]/MLD [9]). For example a flat network may have a multicast querier [8] with a group management. If the IMG has access to this information or the IMG is the querier itself all group states are known. In a static multicast tree based on proxies [10] or BIDIR-PIM [11] all control (join, leave) and data messages are forwarded to the root, which allows monitoring and processing at central position.

C. Multicast Protocols

Monitoring group memberships in a multicast domain depends on the underlying multicast routing protocol:

PIM-SM [12] networks are initiated and managed by a *Rendezvous Point* (RP). All PIM routers of a PIM domain are sending the group signaling of all participants to this RP. Thereby all sender and the receiver information are known and can be used by the IMG.

PIM-SSM [13] is a subprotocol of PIM-SM for source specific multicast, which does not need a RP, as of the available source information the group signaling can be send on the shortest path to the source. Thus, group memberships cannot be recorded at a central position. To run an IMG in a PIM-SM domain, group membership states have to be recorded locally by monitoring local PIM networks—,[14] and send these information to the IMG.

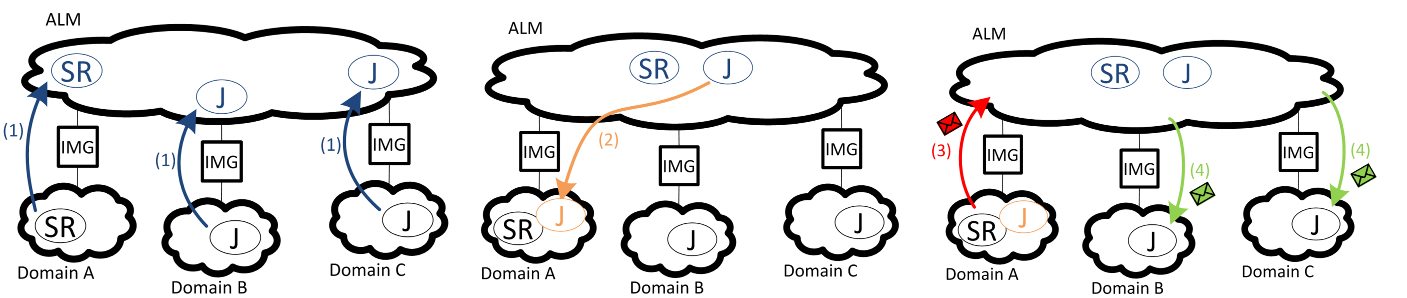
DVMRP [15] networks use flooding to discover and announce group memberships. Similar to PIM-SSM, monitoring group memberships at a central position by an IMG is not possible and will require additional mechanisms. DVMRP is barely used anymore and is mentioned only because of completeness.

Overlay Multicast networks like Scribe [16] or BIDIR-SAM [17] use Distributed Hash Tables (DHT) to abstract from the network topology and pursue a concept of equivalent participants. Such a decentralized system offers no possibility to monitor group states at a central position. Thus, overlay networks must be explicitly adjusted for monitoring group states. For example Scribe builds group specific shared trees, where a host acts as a RP. These RPs can inform the IMGs about their group states and updates.

D. IMG Updating Protocol

The IMG Updating Protocol informs IMGs of a multicast domain about current group states, it provides a reverse mapping and can be implemented in every multicast technology. Every time a host subscribes a group or starts sending data to a group it sends additionally a join or a source register message with the included group name to a predefined and well known group. This group is subscribed by all IMGs of a domain. If an IMG receives a join or a source-register message, it maps the included group name to a group address and saves the tuple in a lookup table. In this way IMGs can match group data to a globally identifiable group name. Furthermore, multicast technologies are mostly non-reliable and packet loss has to be taken into account explicitly.

Robustness and tolerance to packet loss obtains the protocol through measures similar to the MLD [18] protocol. The entries of the lookup table are handled as soft states and are deleted after a period of time to prevent a continuous filling of the table. This requires periodic updates of all entries by



(a) Source and receiver send source-register and join messages, which are aggregated at the upstream.

(b) Join messages are forwarded to a downstream that received source-register messages before.

(c) Group data is routed, when source-register and join messages are received at up- and downstream.

Fig. 3. Example of a cross-domain group communication using hybrid multicast.

(re)sending join and source register messages. To reduce loss probability of these messages, they are duplicated and send delayed – this redundancy is required only during the initial signaling. Furthermore, a number of periodic update messages for the entries of the lookup table are spread over its lifetime. If packet loss occur, then the message is repeated in the next cycle, without deleting the entry. The timing, number of duplicates and number of cycles per lifetime of an entry is based on MLD, which has to solve similar problems. For example, an update cycle of 125 s (*Query Interval*), message duplicate rate of 2 (*Robustness Variable*) in an interval of 1 s (*Unsolicited Report Interval*) is suggested.

IV. Inter-Domain Multicast

The Inter-Domain Multicast layer takes over two tasks: First, it arranges a rendezvous process (RP) between sender and receiver across domain borders, and second, it enables routing of group data across domains and technologies. The rendezvous process is required because of the multicast inherent publish/subscribe paradigm, which decouples sender and receiver spatially and provides a junction on the network level by join- and source-register messages. Routing of group data is done by the IMGs that are connecting different multicast domains and translate between technologies if necessary. Whereby routing decisions are made based on the group management database.

IP Multicast islands are connected through IMGs by a common *Application Layer Multicast* (ALM) network (as shown in fig. 1). ALMs have the advantage that they work independent of the underlying technology, as long as unicast routing is available. This makes them particularly suitable for the integration of multicast islands that are connected over the multicast agnostic Internet. On the one hand ALMs are used to route the group data and on the other hand used as virtual RP, which is described in the following section.

A. Rendezvous Process

The IMG-Updating Protocol was designed to inform all IMGs of a domain by join- and source register messages based on group names, which are send periodically to a predefined group subscribed by all IMGs. If an overlay connects IMGs it mainly consists of them. Thereby the protocol would behave similar to a broadcast technique, which could cause scalability problems in a larger network. Join- and source register messages can be distributed much more targeted by using additional information provided by multicast overlays. This can be shown by multicast overlays such as Scribe [16] and BIDIR-SAM [17].

Scribe implements an multicast overlay using a peer-to-peer (P2P) routing scheme named Pastry [19] to manage the network. Pastry is a *Distributed Hash Table* (DHT), which considers direct neighborhood relations regarding various distance metrics (Proximity Neighbor Selection - PNS), like Hop-Count or RTT. Based on this metric, Scribe creates a group specific shared tree. The first host, which joins a group has to create the group in advance. When creating a group, the responsible host of the related group ID is sought and becomes the RP of the group. All future join-, source-register messages as well as group data are send to this RP and distributed from there. In this way Scribe avoids an unnecessary flooding of group data. BIDIR-SAM however, takes a different approach. It uses a prefix based structure instead of an RP at the root of the shared tree. Therefore, BIDIR-SAM has to flood the group signaling to the subtrees, but builds an efficient source-specific shortest path tree.

An optimization for the rendezvous process in overlays can be implemented in the following way. As ALMs represent a closed system, which controls the network topology and the hosts, hosts can be easy extend by a function to check if a group already exist. By this knowledge, which is for example not easy obtainable in a IP multicast network, it is possible to reduce the rendezvous process from broadcast to relevant IMGs. For that, join- and source register messages are send to a group specific configuration group instead to a global one. This restricts the involved IMGs and has the following impact: If IMGs have local senders or receivers, they subscribe the related configuration groups, which distribute the group states in the overlay. If an IMG sends a join message, it subscribes the transport group. In the case the group does not exits, it has to be created. If an IMGs sends a source register message to the configuration group and the transport group already exists, it can be assumed that subscribers exits without explicitly listen to an join message. In this way, all IMGs receive they relevant group states without latency at the cost of an additional configuration group per group.

B. Routing

The Routing component uses both the information of the Group-Management (group memberships of the locally attached domains) and of the Rendezvous Process (collected group states of foreign domains) to make routing decisions. The following example demonstrate the sequence of forwarding group data between sender and receiver. Figure 3(a) shows three domains of arbitrary multicast technologies connected by IMGs to an overlay (ALM domaind). Domain (A) has one or more sources sending data to group (G). Domain (B) and (C) have receivers that have subscribed group (G). As described by the IMG Updating Protocol, sources announce data with a source-register (SR) messages and receivers are sending join (J) messages to communicate there group state. These messages are aggregated to the upstream and additionally the group of the join messages are subscribed based on the respective multicast technology.

Afterwards the group signaling is distributed by a group specific configuration group to other IMGs. This means, that the IMG, connecting the domain (A) and the overlay, received a SR message from the downstream side and a join message from the upstream. On this setting, the IMG forwards the join message to domain (A) and subscribes the group based on the respective multicast technology (see fig. 3(b)). If the IMG receives group data from group (G) it routes the data to the ALM (fig. 3(c)). The ALM forwards the data to the IMGs of domain (B) and (C) and subsequently data is routed as requested inside domains (B) and (C).

This routing approach has several advantages, as it is simple and based on three routing rules:

1) Join and source-register messages are aggregated to the upstream.
2) Join messages received on the upstream are aggregated to the downstream, if a corresponding source-register message is received on this downstream.
3) IMGs, which received join and source-register messages at the upstream and the downstream side route the group data in the appropriate direction.

This protocol is stateless but needs access to the group management database of the IMG Updating Protocol. Group data is only subscribed and routed if needed and is suitable for single and multi level network hierarchies. However, this protocol does not implement load balancing or non hierarchical connection between domains, e.g, equal peering like in [20].

V. Conclusion and Outlook

In this work, we proposed solutions for a dynamic cross-domain group communication in a hybrid multicast network. We discussed a possible architecture and have given a detailed view on its components. The intra domain component *Data Communication* provides a uniform mapping for all participations from group names to group address. It enables a group communication based on names instead of addresses within a domain. The component *Group Management* creates and manages a database of the domain's group state and the *Rendezvous Process* couples sender and receiver across domains. Therefor, the group states of the domains are distributed in a scalable manner between the IMGs. Based on that we described a simple routing scheme for hierarchical network topologies.

In our ongoing work, we develop technology-independent debugging tools like ping and a trace-route method for real world scenarios.

References

[1] B. Zhang, W. Wang, S. Jamin, D. Massey, and L. Zhang, "Universal IP multicast delivery," *Computer Networks*, vol. 50, no. 6, pp. 781–806, 2006.

[2] X. Jin, K.-L. Cheng, and S.-H. G. Chan, "Island multicast: combining IP multicast with overlay data distribution," *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 1024–1036, 2009.

[3] T. C. Schmidt, M. Wählisch, D. Charousset, and S. Meiling, "On Name-based Group Communication: Challenges, Concepts, and Transparent Deployment," *Computer Communications*, vol. 36, no. 15–16, pp. 1657–1664, Sep–Oct 2013. [Online]. Available: http://dx.doi.org/10.1016/j.comcom.2013.08.001

[4] M. Waehlisch, T. Schmidt, and S. Venaas, "A Common API for Transparent Hybrid Multicast," IETF, RFC 7046, December 2013.

[5] X. Li, C. Bao, M. Chen, H. Zhang, and J. Wu, "The China Education and Research Network (CERNET) IVI Translation Design and Deployment for the IPv4/IPv6 Coexistence and Transition," IETF, RFC 6219, May 2011.

[6] S. Venaas, X. Li, and C. Bao, "Framework for IPv4/IPv6 Multicast Translation," IETF, Internet-Draft – work in progress 03, June 2011.

[7] J.-S. Park, M.-K. Shin, and H.-J. Kim, "A Method for Generating Link-Scoped IPv6 Multicast Addresses," IETF, RFC 4489, April 2006.

[8] B. Cain, S. Deering, I. Kouvelas, B. Fenner, and A. Thyagarajan, "Internet Group Management Protocol, Version 3," IETF, RFC 3376, October 2002.

[9] R. Vida and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6," IETF, RFC 3810, June 2004.

[10] B. Fenner, H. He, B. Haberman, and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ('IGMP/MLD Proxying')," IETF, RFC 4605, August 2006.

[11] M. Handley, I. Kouvelas, T. Speakman, and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)," IETF, RFC 5015, October 2007.

[12] P. Savola, R. Lehtonen, and D. Meyer, "Protocol Independent Multicast - Sparse Mode (PIM-SM) Multicast Routing Security Issues and Enhancements," IETF, RFC 4609, October 2006.

[13] H. Holbrook and B. Cain, "Source-Specific Multicast for IP," IETF, RFC 4607, August 2006.

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

[15] D. Waitzman, C. Partridge, and S. Deering, "Distance Vector Multicast Routing Protocol," IETF, RFC 1075, November 1988.

[16] A. Rowstron, A.-M. Kermarrec, M. Castro, and P. Druschel, "Scribe: The Design of a Large-Scale and Event Notification Infrastructure," in *Networked Group Communication. Third International COST264 Workshop, NGC 2001. Proceedings*, ser. LNCS, J. Crowcroft and M. Hofmann, Eds., vol. 2233. Berlin Heidelberg: Springer–Verlag, 2001, pp. 30–43.

[17] M. Wählisch, T. C. Schmidt, and G. Wittenburg, "On Predictable Large-Scale Data Delivery in Prefix-based Virtualized Content Networks," *Computer Networks*, vol. 55, no. 18, pp. 4086–4100, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2011.07.020

[18] S. Deering, W. Fenner, and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6," IETF, RFC 2710, October 1999.

[19] A. Rowstron and 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)*, ser. LNCS, vol. 2218. Berlin Heidelberg: Springer–Verlag, Nov. 2001, pp. 329–350.

[20] T. C. Schmidt, S. Wölke, and M. Wählisch, "Peer my Proxy - A Performance Study of Peering Extensions for Multicast in Proxy Mobile IP Domains," in *Proc. of 7th IFIP Wireless and Mobile Networking Conference (WMNC 2014)*. Piscataway, NJ, USA: IEEE Press, May 2014, pp. 1–8.