Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Theodor Nolte

Optimale Platzierung von Gateways in hybriden Multicast Routing-Architekturen

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Theodor Nolte

Optimale Platzierung von Gateways in hybriden Multicast Routing-Architekturen

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung
im Studiengang Technische Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer : Prof. Dr. Thomas C. Schmidt
Zweitgutachter : Prof. Dr. rer. nat. Bernd Kahlbrandt

Abgegeben am 8. Januar 2010

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einführung

Die Hybrid-Shared Tree Architektur (4) verfolgt das Ziel einen Internet-weiten Multicast zu ermöglichen, indem IP-Layer Multicast Inseln über eine Overlay-Multicast Infrastruktur miteinander verbunden werden.

Das Herz dieser Architektur ist das Inter-domain Multicast Gateway (IMG). Es ist der Vermittler zwischen dem Overlay und dem Underlay.

Diese Arbeit befaßt sich damit, wo das IMG innerhalb einer IP-Layer Multicast Domain optimal platziert werden kann.

Zunächst befassen wir uns mit dem Problem der Multicast-Kommunikation. Nachdem wir die Funktionsweise des IP-Layer und Applikation-Layer Multicast betrachtet haben, widmen wir uns dem hybriden Multicast. Anschließend überlegen wir, wo das IMG in einer IP-Layer Multicast Domain platziert werden kann.

2 Multicast

Kommunikation bedeutet, daß zwischen den Kommunikationsteilnehmern Nachrichten ausgetauscht werden. Man unterscheidet zwischen bidirektionaler und unidirektionaler Kommunikation. Bei der bidirektionalen Kommunikation ist jeder Teilnehmer Sender und Empfänger, d.h. jeder Teilnehmer hat die Möglichkeit, sowohl Nachrichten-Pakete zu versenden als auch zu empfangen. Bei der unidirektionalen Kommunikation wird hingegen zwischen Sender und Empfänger von Nachrichten unterschieden.

Findet Kommunikation per Unicast statt, werden Nachrichten zwischen genau zwei Teilnehmern ausgetauscht. Dies kann sowohl bidirektional als auch unidirektional geschehen – entweder wechseln sich beide Teilnehmer im Senden und Empfangen der Nachrichten ab, oder einer ist der Sender und der andere der Empfänger.

Stellen wir uns ein Szenario vor, bei dem es genau einen Sender aber viele Empfänger gibt, n an der Zahl (unidirektionale Kommunikation mit mehreren Empfängern). Alle Empfänger sollen eine bestimmte Nachricht vom Sender erhalten. Wäre nur eine Unicast-Kommunikation möglich, so müßte der Sender diese eine Nachricht an jeden Empfänger einzeln verschicken, also n-mal. Dies bedeutet, daß der Sender den Vorgang des Sendens ein und derselben Nachricht n-mal wiederholen müßte (siehe Abbildung 2.1). Wird nun ein Strom von Nachrichtenpaketen an die Empfänger verschickt, so wird der Link des Senders n-mal so stark

Abbildung 2.1: Unicast-Kommunikation – Der Sender S verschickt das gleiche Paket bzw. die gleichen Pakete an jeden Empfänger R (Receiver) separat

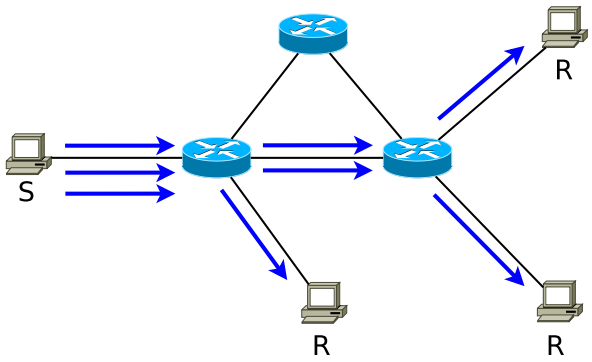
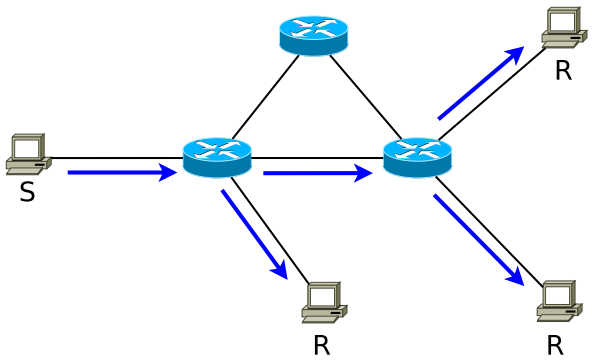


Abbildung 2.2: Multicast-Kommunikation – Der Sender S verschickt ein Paket in einem Vorgang an alle Empfänger R



ausgelastet als bei genau einem Empfänger. Ebenso ist der Aufwand des Senders um den Faktor n erhöht.

Genau an diesem Problem setzt die Multicast-Kommunikation an. Anstatt eine bestimmte Nachricht n-mal zu versenden, findet nur ein Vorgang des Versendens statt. Im Netzwerk selber, in dem die Nachrichten verschickt werden, gibt es einen Mechanismus, der dafür sorgt, daß die Nachricht dann bei den Empfängern ankommt. Mit anderen Worten: Beim Routing der Nachrichtenpakete werden diese an Knotenpunkten geeignet vervielfacht, so daß jeder Empfänger schließlich eine Kopie der Nachricht erhält (siehe Abbildung 2.2). Daraus ergibt sich über das Netzwerk ein Baum, der sogenannte Verteilbaum, mit der Wurzel beim Sender, Verzweigungen an den Routern und den Blättern bei den Empfängern.

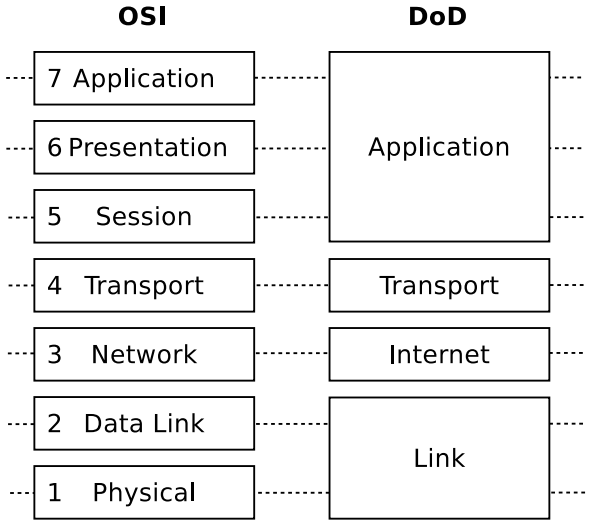
Die Zustände über die Gruppenmitgliedschaften sind über die Knotenpunkte, die Router, verteilt. Jeder Router führt eine Routingtabelle, anhand derer er entscheidet, über welche ausgehenden Links Multicast-Pakete weitergeleitet werden.

Die n Empfänger bilden eine Multicast-Gruppe. Ihr ist eine Adresse aus dem Namensraum des Netzwerks zugeordnet. Ein Multicast-Paket, das an die Mitglieder einer Gruppe verschickt wird, ist an diese Adresse adressiert. Die Multicast-Gruppe wird über die Adresse identifiziert.

Multicast-Kommunikation ist also eine Gruppenkommunikation. In einer Multicast-Gruppenkommunikation gibt es drei Aufgabenbereiche, um das Routing von Multicast-Paketen zu bewerkstelligen:

Gruppenmanagement Hinzufügen neuer Teilnehmer in eine Multicast-Gruppe und das Entfernen von Mitgliedschaften, wenn die Empfänger nicht mehr existieren bzw. ihre Mitgliedschaft aufkündigen.

Abbildung 2.3: Schichtenmodelle paketorientierter Netzwerkkommunikation



Bildung und Pflege des Verteilbaums Den Baum für die Verteilung der Multicast-Pakete aufbauen und warten, so daß bei Änderungen in den Gruppenmitgliedschaften weiterhin alle Mitglieder erreicht werden und ausgeschiedene Teilnehmer nicht mehr mit Multicast-Paketen bedient werden.

Weiterleitung von Multicast-Paketen Multicast-Pakete entlang des Verteilbaums an die Empfänger zustellen.

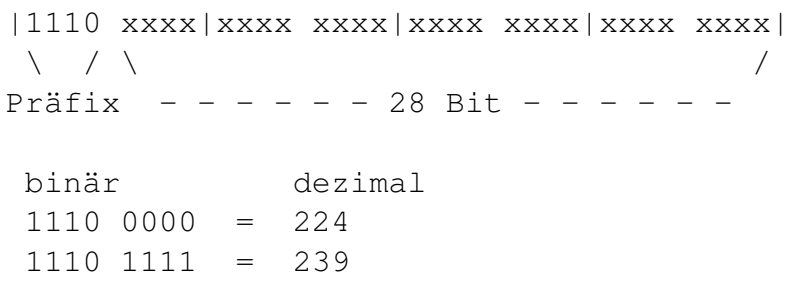
Multicast kann nativ oder in einem Overlay-Netzwerk realisiert werden. Nativ bedeutet, daß das Routing von Multicast-Paketen auf der Netzwerkschicht stattfindet. Ist die Multicast-Kommunikation in einem Overlay-Netzwerk realisiert, so ist auf Anwendungsebene ein Overlay-Netzwerk mit eigenem Namensraum auf dem nativen Netzwerk (dem Underlay) aufgesetzt. Hier wird der Multicast als Application Layer Multicast oder auch Overlay Multicast (OLM) bezeichnet.

Zur besseren Veranschaulichung, sind auf Abbildung 2.3 das OSI-Schichtenmodell (5; 6) und das DoD-Schichtenmodell (7) nebeneinander aufgezeigt.

2.1 IP-Layer Multicast

Der IP-Layer Multicast oder auch native Multicast ist in den TCP/IP Protokoll-Stack integriert. Er ist realisiert – wie der Name schon sagt – in dem IP-Layer. Bezieht man sich auf die feinere (und umständlichere) Unterteilung des OSI Schichtenmodells, so findet eine Multicast-Behandlung im Layer 3, dem Network Layer, innerhalb des IP-Protokolls statt.

Abbildung 2.4: IPv4 Multicast Adressen



224.0.0.0	reserviert
224.0.0.1	Alle Nodes innerhalb des Subnetzwerks
224.0.0.2	Alle Router in einem Subnetzwerk
224.0.0.4	Alle DVMRP Router in einem Subnetzwerk
224.0.0.9	Alle RIPv2 Router eines Subnetzwerks
224.0.1.1	Alle NTP Empfänger

Tabelle 2.1: Ausgewählte permanente IPv4 Multicast-Adressen

Auf Ebene des Transport-Layers wird nicht TCP, sondern das verbindungslose UDP verwendet. Würde TCP verwendet werden, käme es zur ACK-Implosion (8). Ein Sender von Multicast-Paketen würde von jedem Empfänger mit Empfangsbestätigungen überhäuft werden.

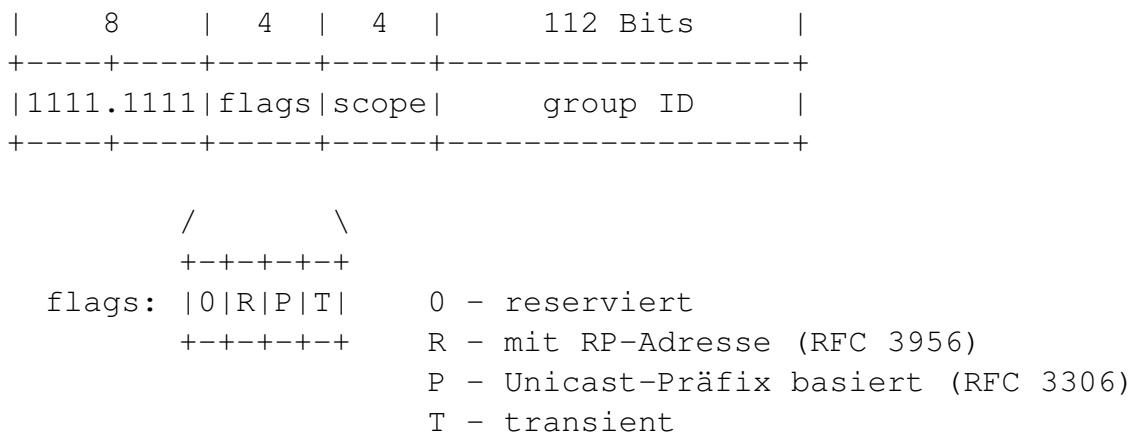
Für Multicast IPv4-Pakete sind die (32-Bit) Adressen von 224.0.0.0 bis 239.255.255.255 reserviert (9). Dieser Adressraum wird mit Class D bezeichnet. Die ersten vier Bits – das Präfix – einer Multicast-Adresse lauten also „1110“ (siehe Abbildung 2.4).

Im Gegensatz zu den Class A, Class B und Class C Adressen sind die restlichen Bits einer Class D Adresse nicht in Net-ID und Host-ID unterteilt. Eine Multicast Adresse gehört also nicht zu einem bestimmten Netzwerk. Die „eigentliche“ Multicast Adresse (ohne Präfix) hat eine Länge von 28 Bits. Demnach gibt es 268.435.456 ($= 2^{28}$) unterschiedliche Multicast Adressen.

Der IPv4-Multicast Adressraum wiederum ist unterteilt in permanente und dynamische Gruppen. Einer permanenten Gruppe ist eine Multicast-Adresse fest zugeordnet. Einige dieser Adressen und der dazugehörigen Gruppen sind in Tabelle 2.1 aufgelistet.

IPv6 Multicast-Adressen (10) sind dadurch gekennzeichnet, daß sie mit acht Einsen beginnen. Anschließend sind vier Bits für Flags und weitere vier Bits für den Scope – den Gül-

Abbildung 2.5: IPv6 Multicast Adressen



tigkeitsbereich – reserviert. Die verbleibenden 112 Bits bilden die Group-ID. Abbildung 2.5 stellt das Format der IPv6 Multicast Adressen dar.

Wie bei den IPv4 Multicast-Adressen gibt es in IPv6 permanente Multicast-Adressen. Sie weisen sich als solche dadurch aus, daß das Flag T auf 0 gesetzt ist. T steht für transient, also für nicht dauerhaft.

Wie wird nun eine logische Multicast IP-Adresse auf eine Layer 2 Adresse, die MAC-Layer Adresse zugeordnet? Ein erster Einfall wäre es, daß ein Router, der ein Multicast Paket verteilt, dieses Multicast-Paket an die MAC-Adressen der Empfänger des angrenzenden Subnetzes adressiert und entsprechend oft einzeln weiterleitet. Dies würde bedeuten, daß auf IP-Ebene die Multicast-Pakete effizient über einen Verteilbaum geroutet werden können. Aber auf MAC-Ebene würde wieder eine Unicast-Kommunikation stattfinden. Zudem müsste der Router die Informationen halten und pflegen, zu welchen Hosts bzw. an welche MAC-Adressen Multicast Pakete zuzustellen sind. Ein weiterer Einfall könnte sein, daß die Multicast-Pakete gebroadcastet werden, was aber aus Gründen der Effizienz auch nicht wünschenswert ist.

Stattdessen funktioniert es folgendermaßen: Ein Node empfängt nun eine Fülle von Paketen, auch die nicht an ihn adressierten. Damit der Treiber und somit der Prozessor des Rechners sich nicht mit den Paketen befassen muß, die nicht an den Node adressiert sind, ist ihm ein MAC Filter vorweggeschaltet. Dieser Filter verwirft alle Pakete, die nicht an den Node adressiert sind. Es findet eine Vorauswahl auf Hardwareebene statt, so daß die Rechenlast auf Softwareebene reduziert wird.

Soll der Router nun ein IPv4-Multicast Paket zustellen, so bildet er die Multicast IP-Adresse auf eine MAC-Adresse ab. Dieser Vorgang heißt Adress Mapping. Eine MAC-Adresse hat

eine Länge von 48 Bits. Das Präfix für eine IPv4-Multicast MAC-Adresse hat eine Länge von 25 Bits. Es verbleiben noch 23 Bits für die eigentliche Multicast Adresse. In diese wird der 23-Bit Postfix der IP-Multicast Adresse übernommen. Die fünf ersten Bits der „eigentlichen“ Multicast Adresse werden nicht übernommen, sie werden verworfen. Dies hat zur Folge, daß beispielsweise die beiden unterschiedlichen IP-Multicast Adressen 224.1.2.3 und 225.1.2.3 auf dieselbe MAC-Adresse abgebildet werden. Wenn ein Node Multicast-Pakete empfangen möchte, so wird zunächst der MAC-Filter des Interfaces so konfiguriert, daß sowohl an die Unicast MAC-Adresse als auch an die entsprechenden Multicast MAC-Adressen adressierte Pakete durchgelassen werden. Im Anschluß muß noch überprüft werden, ob nach Entfernen des MAC-Layer Anteils die richtige Multicast IP-Adresse enthalten ist.

2.1.1 Gruppenmanagement mit IGMP/MLD

Es gibt zwei unterschiedliche Multicast-Zustände:

ASM (*,G) ASM steht für Any Source Multicast. Jeder kann Multicast-Pakete an die Gruppe G versenden.

SSM (S,G) SSM bedeutet Source Specific Multicast. Nur Multicast-Pakete vom Sender S herstammend sind zulässig.

Multicast-Kommunikation erfolgt empfängerbasiert. Damit in einem Netzwerk ein Multicast-Routing etabliert werden kann, müssen die Router zunächst einmal wissen, *wohin* die Multicast-Pakete, die sie empfangen, überhaupt zuzustellen sind: Der Router benötigt ein Bild darüber, in welchen Multicast-Gruppen die Hosts in den angrenzenden Subnetzwerken beigetreten sind. So kann er über das mit dem entsprechenden Subnetzwerk verbundene Interface ein Multicast-Paket gezielt an die Hosts zustellen, die Mitglied der Gruppe des Multicast-Pakets sind. In das Routing müssen deshalb die Informationen über die Multicast Gruppenmitgliedschaften der Hosts (der End-Systeme) einfließen.

Dazu bedarf es zunächst einer Kommunikation zwischen Router und Host. Diese ist über das Internet Group Management Protocol (IGMP) für IPv4 bzw. über das Protokoll namens Multicast Listener Discovery (MLD) für IPv6 geregelt. IGMP in seiner aktuellen Version Nr. 3 ist in RFC 3376 (11) definiert, MLD in der aktuellen Version Nr. 2 in RFC 3810 (12). MLD ist ein Sub-Protokoll von ICMPv6 (13). Dies bedeutet, daß MLD-Nachrichten wie alle ICMPv6-Nachrichten in IPv6-Paketen durch ein vorhergehenden Next-Header Wert von 58 identifiziert werden.

Ein Interface kann bezüglich IGMP entweder die Rolle eines Routers (Multicast Router Part) oder die des Hosts, des Empfängers (Multicast Listener Part) annehmen. Ein Router hat üblicherweise mehrere Interfaces. Hier ist es durchaus möglich, daß ein Interface den Listener

Part und ein anderes den Router Part übernimmt. Dies ist z. B. beim IGMP/MLD Proxying der Fall, welches wir auf Seite 17 ff. näher betrachten.

Aufgabe des Hosts ist es, den oder die Router über seine Gruppenmitgliedschaften zu informieren, mit dem Ziel, Multicast-Pakete zu erhalten. Dafür verschickt er Reports. Reports sind selber Multicast-Pakete, adressiert an die Adresse 224.0.0.2 (IPv4, vgl. Tabelle 2.1) bzw. FF02:0:0:0:0:0:0:16 (IPv6), die im Datenblock das IGMP- bzw. ICMP-Paket (bei MLD) enthalten. Das TTL-Feld (IPv4) bzw. das Hop Limit Feld (IPv6) hat den Wert 1. Somit wird das Paket von Routern nicht weitergeleitet und ausschließlich von IP-Multicast Routern im Subnetzwerk angenommen.

Der Router merkt sich jeweils für seine Interfaces die Gruppenmitgliedschaften der Hosts der angrenzenden Subnetzwerke. Dabei muß er sich nicht merken, welcher Host welcher Gruppe angehört. Es reicht aus, wenn er sich lediglich merkt, welche Gruppenabbonements für ein angrenzendes Subnetz existieren. Insofern erfährt der Router das Subnetzwerk als ein Ganzes.

Die Mitgliedschaft in einer Multicast-Gruppe führt der Router als ein zeitlich begrenztes Abonnement: Wird eine Gruppenmitgliedschaft eines Subnetzwerks nicht innerhalb eines Timeouts, dem Group Membership Interval (Default: 260 Sekunden[1]), von einem Host aus dem Subnetzwerk durch einen Report erneuert, so werden zukünftige Multicast-Pakete, die an diese Gruppe adressiert sind, nicht mehr an dieses Subnetzwerk weitergeleitet.

Der Router verschickt Queries. Queries sind wie die Reports Multicast-Pakete, die nicht geroutet werden (TTL = 1 bzw. Hop Limit = 1) und im Datenteil das IGMP- bzw. ICMP-Paket (MLD) enthalten. Sie sind adressiert an die All-Systems Gruppe, d.h. an die Adresse 224.0.0.1 (IPv4) bzw. an die Link-Scope All-Nodes Multicast-Adresse FF02::1 (IPv6), oder, wenn sie sich auf eine bestimmte Multicast-Gruppe beziehen, eben die Adresse dieser Gruppe.

Es gibt drei Varianten von Queries. Das allgemeine Query (General Query; (*,*) Query) ist sozusagen die Frage „in die Runde“ (an die Hosts des Subnetzwerks gerichtet), ob es jemanden gibt, der Multicast-Nachrichten empfangen will. Dann gibt es ein gruppenspezifisches Query (Group-specific Query; (*,G) Query), eine Aufforderung an die Hosts der Gruppe G, ihre Mitgliedschaft bekannt zu geben. Und zuletzt gibt es eine noch feinere Anfrage, das gruppen- und senderspezifische Query (Group-and-Source-specific Query; (S,G) Query). Damit werden Hosts, die Multicast-Pakete einer bestimmten Gruppe von einem bestimmten Sender empfangen, aufgefordert, ihre Gruppenmitgliedschaft mitzuteilen.

[1] Group Membership Interval = Robustness Variable x Query Intervall + Query Response Interval
Defaults: Robustness Variable = 2, Query Intervall = 125 s, Query Response Interval = 10 s

Auf diese Queries antworten die Hosts mit Report-Nachrichten. Auch wenn sich etwas an ihren Gruppenmitgliedschaften geändert hat (Beitreten oder Austreten einer Gruppe), verschicken sie eine Report-Nachricht, um dies dem Router bzw. den Routern mitzuteilen.

Die Reports bestehen im Wesentlichen aus einem oder mehreren sogenannten Multicast Address Records. Das können Current-State Records oder State-Change Records sein.

In einem Current-State Record ist die Information über die Gruppenmitgliedschaft bezüglich einer Multicast-Adresse enthalten. Handelt es sich um eine SSM-Gruppe (einem Channel (1)) so sind auch Informationen über die Senderadressen, von denen Multicast-Pakete angenommen werden, enthalten.

Der State-Change Record enthält die für einen Router notwendigen Informationen zur Zustellung von Multicast-Paketen, welche an eine Multicast-Adresse adressiert sind, über einen geänderten Zustand bezüglich der Gruppenmitgliedschaft. Auch hier kann es sich um eine ASM- oder SSM-Mitgliedschaft handeln.

Erreicht einen Host eine Query-Nachricht, so zeigt der Querier damit an, daß er über die Gruppenmitgliedschaften, die im Subnetzwerk vorherrschen, informiert werden möchte. Der Host antwortet darauf, indem er in Current-State Records seine Gruppenmitgliedschaften kund tut.

Interessant ist, daß er nicht unmittelbar auf das Query reagiert, sondern eine zufällige Zeitspanne wartet, bevor er mit einer Report-Nachricht antwortet. Die Zeitspanne kann höchstens so lang sein wie die sogenannte Max Resp Time, welche aus dem mit dem Query übermittelten Max Resp Code abgeleitet wird. Hier findet das gleiche Prinzip wie bei CSMA/CA zur Vermeidung von Kollisionen Anwendung: Würden die Hosts unmittelbar auf ein Query reagieren, würden die zeitgleich verschickten Report-Nachrichten-Pakete einen Link-Stress Peak erzeugen. Dadurch, daß die Hosts unterschiedlich lange den Report verzögern, wird der Peak vermieden.

Man mag nun auf den Gedanken kommen, daß ein Host, sobald ein anderer Host, der derselben Multicast-Gruppe angehört, seinen Report nicht mehr verschicken braucht, da der Router das Subnetz ja „nur“ als ein Ganzes wahrnimmt und nun die Infomation erhalten hat, daß es (mindestens) einen Empfänger gibt. Alte Versionen von IGMP/MLD (9; 14; 15) haben dies auch so vorgesehen. Allerdings behindert das Unterdrücken von Reports das IGMP/MLD-Snooping (16). Deswegen soll jeder Host mit Reports auf Queries antworten, wenn er Mitglied entsprechender Multicast-Gruppen ist.

Tritt ein Host einer Multicast-Gruppe bei oder verläßt diese, so verschickt er einen Report, der durch State-Change Records die Veränderung ausdrückt.

Der Join, d.h. das Abonnieren von Multicast-Paketen einer bestimmten Gruppe entspricht einer Report-Nachricht, die einen Multicast Address Record enthält, genauer, einen State-Change Record, der anzeigt, daß zukünftig Pakete von der Gruppe erwünscht sind. Der besseren Verständlichkeit wegen sind die grafisch dargestellten IGMP/MLD Reports der Abbildung 2.6 auf Seite 19 und der nachfolgenden Schaubilder als Joins und nicht als Reports gekennzeichnet.

Gibt es in einem Subnetzwerk mehrere Router, so wird automatisch einer zum Querier bestimmt. Der Querier ist der einzige Router (genauer: das einzige Interface), der Query-Nachrichten verschickt. Der einfache wie elegante Mechanismus heißt Querier Election: Ein Router-Interface ist, sobald es seinen Betrieb aufnimmt, im Querier-Zustand. Empfängt ein Router ein Query von einer Absender IP-Adresse niedriger als der des eigenen Interfaces, so geht er in den Non-Querier Zustand über. Es werden nun keine Queries verschickt. Dieser Zustand ist mit einem Timeout versehen, dem Other Querier Present Interval (Default: 255 Sekunden[2]). Wird innerhalb des Intervalls ein Query von einem Absender niederer IP-Adresse empfangen, so wird der Timer wieder von vorne gestartet. Läuft das Timeout aus, wird der Router selber wieder zum Querier.

Es macht Sinn, daß es nur einen Querier gibt. Mehrere Querier würden zu redundantem Multicast-Traffic führen, was nicht erwünscht ist, da die Redundanzfreiheit den Multicast als solchen auszeichnet. Auch führt es zu redundanten Queries und Reports, eine Redundanz, die so nicht sinnvoll genutzt werden kann (da sie über die Anzahl der Router gesteuert würde, welche recht willkürlich ist).

Um Redundanz und damit Fehlertoleranz zu erreichen, gibt es die Robustness Variable. Sie gibt an, wie oft ein Host eine Report-Nachricht verschickt. Sie hat als Default den Wert 2. IGMP/MLD kann also den Verlust einer Nachricht verkraften, funktioniert dann immer noch korrekt.

Tabelle 2.2 vermittelt einen Überblick, wei über die Zeit das IP-Gruppenmanagement in seiner Funktionalität gewachsen ist. Weil ältere Versionen der Protokolle in Routern implementiert wurden oder sind, sind neuere Versionen abwärtskompatibel.

2.1.2 IGMP/MLD Proxying

In bestimmten gerouteten Netzwerken kann eine Multicast-Kommunikation ausschließlich anhand der IGMP/MLD-Informationen über die Gruppenmitgliedschaften der Hosts erreicht werden, ohne daß ein Multicast Routing-Protokoll zum Einsatz kommt. Hierfür ist es notwendig, daß die Multicast-Kommunikation (Status- sowie Daten-Pakete) über einen Spannbaum

[2] Other Querier Present Interval = Robustness Variable x Query Inteval + Query Response Interval / 2

Tabelle 2.2: Überblick über die Entwicklung des IP-Gruppenmanagements (IGMP/MLD)

	RFC	Protokoll	Features
August 1989	1112	**IGMP**	Gruppensignalisierung in IPv4 General Membership Query Group Specific Membership Query Version 1 Membership Report
November 1997	2236	**IGMPv2**	Low Leave Latency (Prune) Querier Election Robustness Version 2 Membership Report
Oktober 1999	2710	**MLD**	Gruppensignalisierung in IPv6 abgeleitet von IGMPv2 eingebettet im Protokoll ICMPv6
Oktober 2002	3376	**IGMPv3**	Source Filtering (Source Specific Multicast) SSM-Queries und -Reports
Juni 2004	3810	**MLDv2**	Übersetzung des IGMPv3 nach IPv6

erfolgt, es also keine Maschen gibt. Dies ist gleichbedeutend damit, daß es genau eine Möglichkeit gibt, ein Multicast-Paket von einem Sender zu einem Empfänger weiterzuleiten.[3] So erfüllt das Netzwerk in Abbildung 2.11 auf Seite 25 diese Bedingung nicht, da es mehrere Maschen unter der Vernetzung der dort abgebildeten Multicast-Router gibt.

Solch eine Architektur ist in RFC 4605 (1) beschrieben. Das Verfahren, den Multicast hierbei zu ermöglichen, wird IGMP/MLD Proxying genannt.

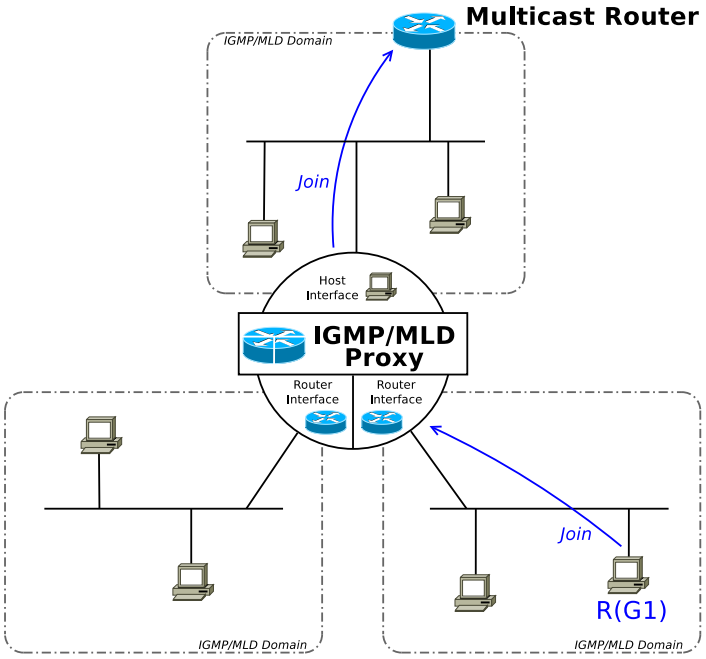
Der IGMP/MLD Proxy hat genau ein Upstream-Interface und kann mehrere Downstream-Interfaces haben. Das Upstream-Interface übernimmt den Group Member Part, die Downstream-Interfaces übernehmen den Multicast Router Part im IGMP/MLD-Protokoll.

Zur hierarchisch oberen IGMP/MLD-Domain verhält sich der Proxy über das Upstream-Interface demnach wie ein IGMP/MLD-Host. Deshalb wird das Upstream-Interface auch als Host-Interface bezeichnet. Nach unten übernimmt er die Rolle des IGMP/MLD-Routers – die Downstream-Interfaces werden auch Router-Interfaces genannt.

Abbildung 2.6 zeigt einen IGMP/MLD Proxy mit den angrenzenden IGMP/MLD Domains.

[3] Damit ist nicht ausgeschlossen, daß das Unicast-Netzwerk frei von Maschen sein muß.

Abbildung 2.6: IGMP/MLD Proxying – Signalisierung von Gruppenmitgliedschaften



Hier ist beispielhaft in der hierarchisch obersten IGMP/MLD-Domain ein Multicast-Router platziert.

Der IGMP/MLD Proxy merkt sich für jedes Downstream-Interface die Gruppenmitgliedschaften der über das Interface erreichbaren Hosts. Diese Informationen der unteren Domains gibt er an die obere Domain aggregiert weiter. Er abboniert somit stellvertretend für seine unteren Domains den Multicast-Traffic. Erhält der Proxy nun über das Upstream-Interface ein Multicast-Paket, so leitet er es an alle Downstream-Interfaces weiter, über welche Mitglieder erreichbar sind, die zu der Gruppe des Multicast-Pakets gehören. In Abbildung 2.6 wird solch eine Signalisierung einer Gruppenmitgliedschaft „nach oben“ dargestellt.

Gelangt ein Multicast-Paket über ein Downstream-Interface zum Router, so wird es neben den anderen Downstream-Interfaces, hinter denen Empfänger erreichbar sind, auch über das Upstream-Interface weitergeleitet. So wird ermöglicht, daß auch Empfänger in hierarchisch höheren Multicast-Domains das Paket erhalten.

Abbildung 2.7 zeigt unterschiedliche Fälle des Weiterleitens von Multicast-Paketen. Wichtig für uns ist die Tatsache, daß alle Informationen über die Gruppenmitgliedschaften (Joins, Prunes, States) und alle Multicast-Pakete zur Wurzel – zur hierarchisch höchsten IGMP/MLD Domain – gelangen.

Wie aus mehreren IGMP/MLD-Proxys eine Kaskade gebildet werden kann, zeigt Abbildung 2.8. Die hierarchisch höchste IMGP/MLD-Domain ist die Root-Domain.

Abbildung 2.7: IGMP/MLD Proxying – Weiterleiten von Multicast-Paketen

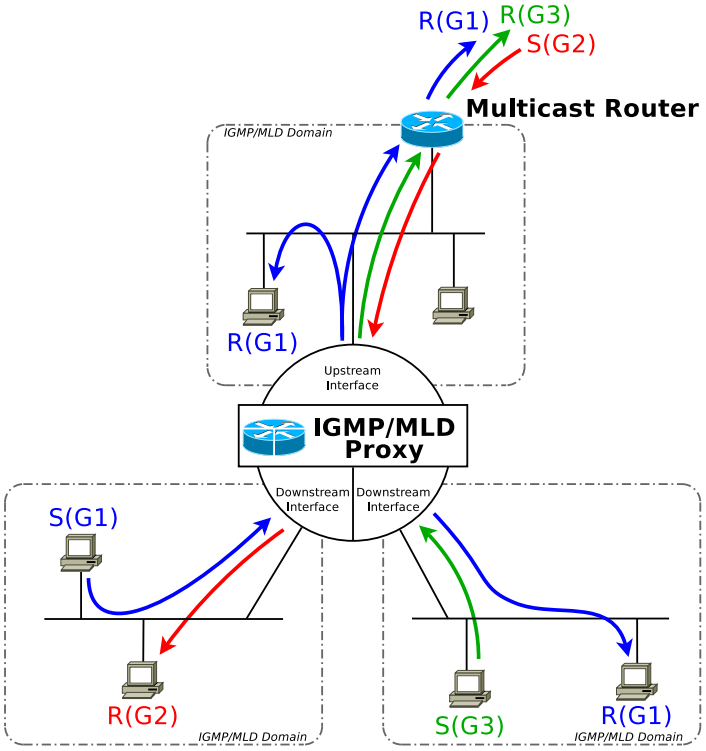


Abbildung 2.8: IGMP/MLD Proxying – Kaskadierung mehrerer IGMP/MLD Proxys mit Weiterleitung von Multicast-Paketen (grün) und einem Gruppen-Join (rot)

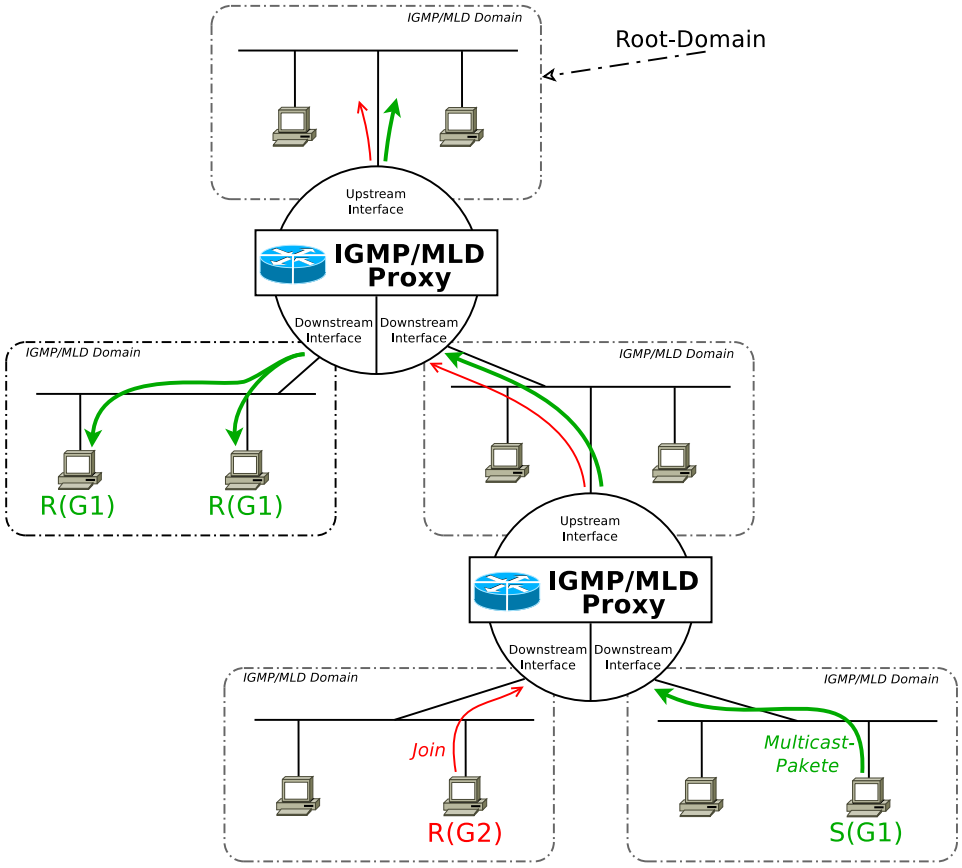
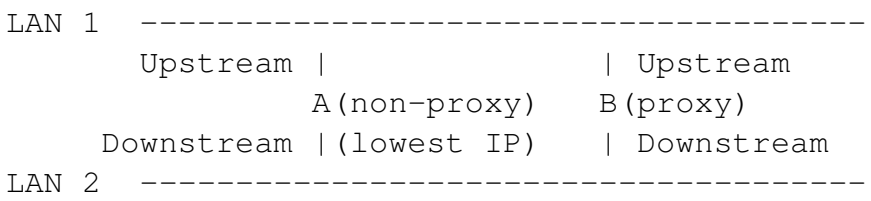


Abbildung 2.9: IGMP/MLD Proxying – Uniqueness, Quelle: (1)

Abbildung 2.10: IGMP/MLD Proxying – Upstream Loop, Quelle: (1)



Wenn es aus Gründen der Redundanz (um die Ausfallsicherheit zu erhöhen) mehrere IGMP/MLD-Proxys zwei IGMP/MLD-Domains miteinander verbinden (siehe Abbildung refmcastProxying:uniqueness), so kann, da die Router-Interfaces eine IGMP/MLD Querier Election abhalten, das Ergebnis dieser Wahl auch den aktiven IGMP/MLD Proxy bestimmen.

Allerdings muß man darauf achten, eine sogenannte Upstream Loop zu vermeiden, wie sie in Abbildung 2.10 dargestellt ist.

Wenn die Topologie des Netzes es zuläßt, stellt IGMP/MLD Proxying eine relativ einfache Möglichkeit für eine IP-Multicast Gruppenkommunikation dar. Weil die Konfiguration jedoch statisch erfolgt, können hierbei Fehler gemacht werden, so daß eine Upstream Loop entsteht. Auch können Änderungen in der Topologie eine manuelle Anpassung der Konfiguration der Proxys erfordern – eine weitere Möglichkeit Fehler zu machen.

2.1.3 Multicast Routing

Analog dem Unicast Routing beinhaltet das Multicast Routing das Ermitteln von optimalen Routen und Festhalten dieser in den Routingtabellen sowie das Forwarding von Multicast Paketen anhand der Routingtabellen.

2.1.3.1 PIM-SM

PIM steht für Protocol Independent Multicast. Es ist ein Routing Protokoll für Multicast Nachrichten, welches in zwei Ausprägungen existiert, dem Dense-Mode und dem Sparse-Mode. PIM Dense-Mode (PIM-DM) ist definiert in dem RFC 3973(17), PIM Sparse-Mode (PIM-SM) im RFC 4601(18).

Mit Protocol Independent ist gemeint, daß das verwendete Unicast Routing-Protokoll nicht festgelegt ist. Dies ist ein wichtiges Merkmal von PIM, womit es sich von bisherigen Multicast Routing Protokollen wie Distance Vector Multicast Routing Protocol (DVMRP, (19)) oder Multicast Extensions to Open Shortest Path First (MOSPF, (20)) unterscheidet.

PIM-DM ist für Netze mit hoher Teilnehmerdichte entworfen worden. Bei geringerer Teilnehmerdichte arbeitet es zunehmend ineffizient. Typischerweise haben aber viele Multicast Domains zur Zeit eine geringe Dichte an Teilnehmern, so daß uns dieses Protokoll hier nicht weiter interessiert.

PIM-SM hingegen ist für Netze konstruiert, bei denen die Teilnehmerdichte als relativ gering angenommen wird. Dieser Annahme entsprechend ist ein Hauptmerkmal von PIM-SM, daß die Weiterleitung von Multicast-Paketen erst erfolgt, nachdem mögliche Empfänger ihren Wunsch angezeigt haben, diese zu empfangen.

Bevor wir uns der Funktionsweise von PIM-SM widmen, führen wir zunächst ein paar spezifische Begriffe ein:

In einer PIM-SM Domain wird für eine Multicast Gruppe ein Router als der Rendezvous Point (RP) für diese Gruppe konfiguriert. Man kann für jede Multicast Gruppe einen anderen Router auswählen. So ist es auch möglich, daß es genau einen RP für alle Gruppen gemeinsam gibt. Solch ein RP ist sozusagen der Bezugspunkt zur Gruppe, an den sich neue Teilnehmer dieser Gruppe wenden, um den Empfang von Multicast Paketen zu initiieren.

Innerhalb einer Domain kann es unterschiedliche Subnetzwerke geben, die mit einem oder mehreren Routern verbunden sind. Durch eine automatisch ablaufende, in PIM-SM spezifizierte Wahl wird einer als der Designated Router (DR) dieses Subnetzes festgelegt. Der DR handelt als Stellvertreter für die Belange (Multicast-Gruppenmitgliedschaften) des Subnetzwerkes; er erwirkt den Empfang von Multicast Nachrichten und die Beendigung des Empfangs. Die Informationen über die Gruppenmitgliedschaften der Hosts in dem Subnetzwerk werden per IGMP/MLD ausgetauscht.

Kommen Router in einer Domain neu hinzu, so müssen sie erfahren, welcher Router für welche Multicast Gruppe der RP ist. Diese Information erhalten sie von den Bootstrap-Routern. Ein Bootstrap-Router ist ein Router, der per Flooding wiederkehrend diese Information verteilt. Auch die Bootstrap-Router können innerhalb von PIM-SM für ihre Aufgabe automatisch gewählt werden.

Um das Weiterleiten von Paketen zu regeln, die der Statuskommunikation oder des Verbreiten von Multicast Nachrichten dienen, führt ein PIM-SM Router mehrere Tabellen. Die Multicast Routing Information Base (MRIB) wird benutzt, um zu entscheiden, wohin Join- und Prune-Nachrichten weitergeleitet werden. Der Reverse Path Forwarding (RPF) Neighbour bezüglich einer IP-Adresse ist derjenige Router, an welchen an die IP-Adresse adressierte Pakete weitergeleitet werden. Genau diese Pakete sind die Join- und Prune-Nachrichten.

Ein Router merkt sich in der Tree Information Base (TIB), von welchem Router er Join- bzw. Prune-Nachrichten bekommen hat, und weiß damit, an wen Multicast Traffic weiterzuleiten ist; es bildet sich der Reverse Path Tree nach dem Prinzip des Reverse Path Forwarding Algorithmus (21). Des Weiteren fließen Informationen aus den Assert-Nachrichten herstammend und - sofern eine lokale Domain vorhanden - aus dem IGMP/MLD Protokoll in die TIB hinein.

Die Tree Information Base ist recht umfangreich. Um das Multicast Forwarding durchzuführen, wird aus der TIB eine „Essenz“ destilliert, die Multicast Forwarding Information Base (MFIB).

Es sei noch gesagt, daß die drei Phasen nicht zwingend in fester Reihenfolge ablaufen. Die Initiierung einer Phase geschieht von unterschiedlichen Entitäten.

Funktionsweise Die Funktionsweise von PIM-SM läßt sich in drei Phasen aufteilen:

1. Phase One: Bilden des Rendezvous-Point Tree, der initialen Multicast Infrastruktur; Multicast getunnelt per Registering.
2. Phase Two: Nativer Multicast über eine Kombination des Rendeszvous Point Tree mit dem Source-specific Tree.
3. Phase Three: Multicast über den Shortest-Path Tree.

Diese Phasen werden wir anhand eines Beispiels besprechen. Abbildung 2.11 zeigt eine PIM-SM Domain mit den Routern A bis H und fünf Subnetzwerken. Der Router A ist als Rendezvous Point konfiguriert. (Um das Beispiel einfach zu halten, betrachten wir nur eine Multicast-Gruppe. Und so gibt es hier auch nur einen RP.)

Phase One In den Subnetzen gibt es die Rechner R1 bis R5, die ihre DRs über ihre Multicast Gruppenmitgliedschaft für die Gruppe G informiert haben, beispielsweise per IGMP oder MLD (siehe Abbildung 2.12). Die DRs verschicken nun ein (*,G) Join an den Rendezvous Point, welche den RP Tree (RPT) etablieren.

Nun beginnt der Sender S mit dem Versand von Multicast Paketen an die Multicast Gruppe G (siehe Abbildung 2.13). Der Designated Router des Subnetzwerkes, in dem sich S befindet,

Abbildung 2.11: PIM-SM - Grundgerüst des Beispiels

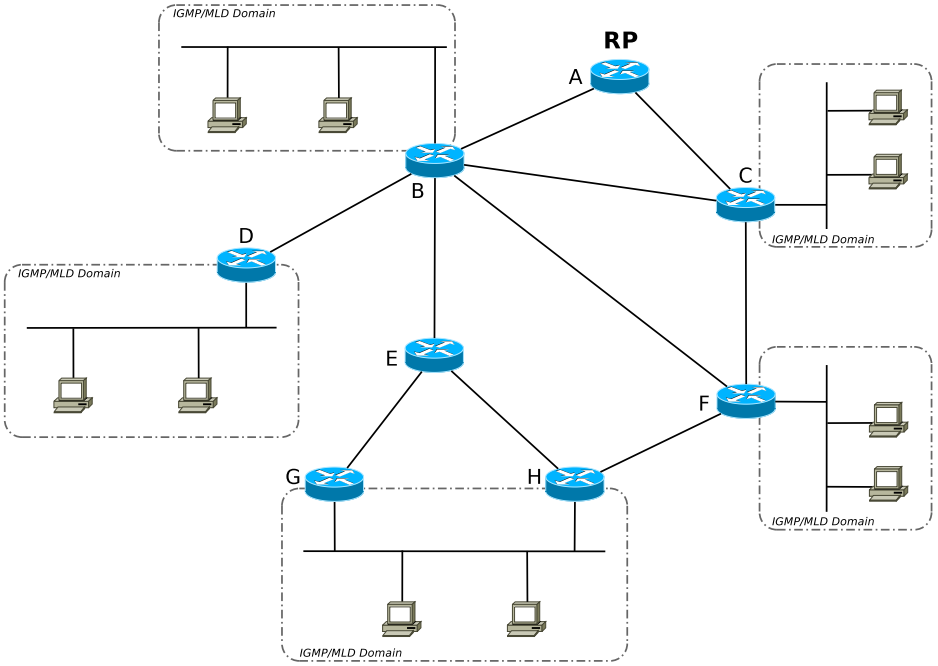


Abbildung 2.12: PIM-SM – Phase One - RP Tree

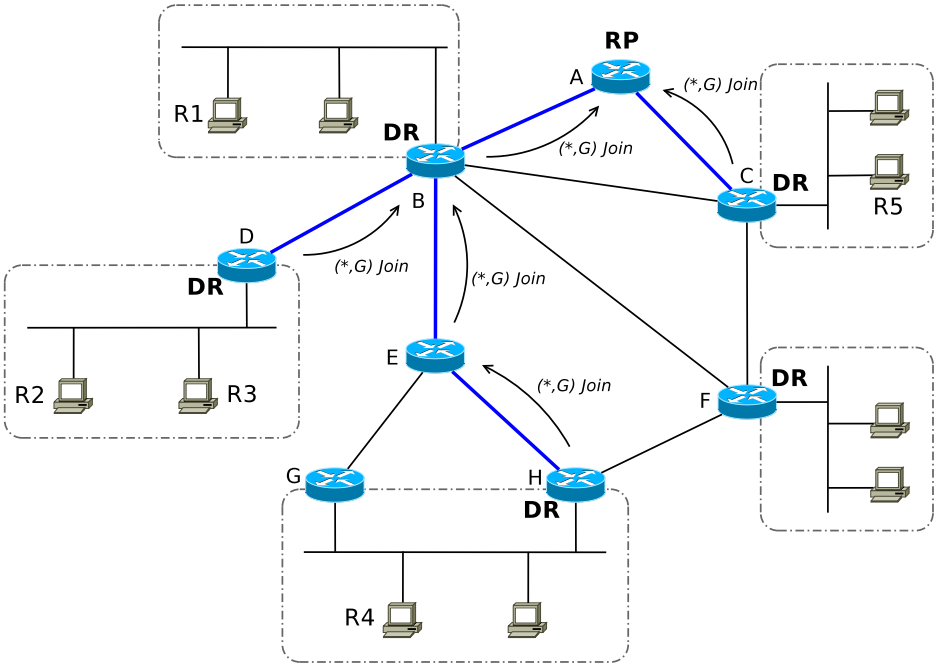
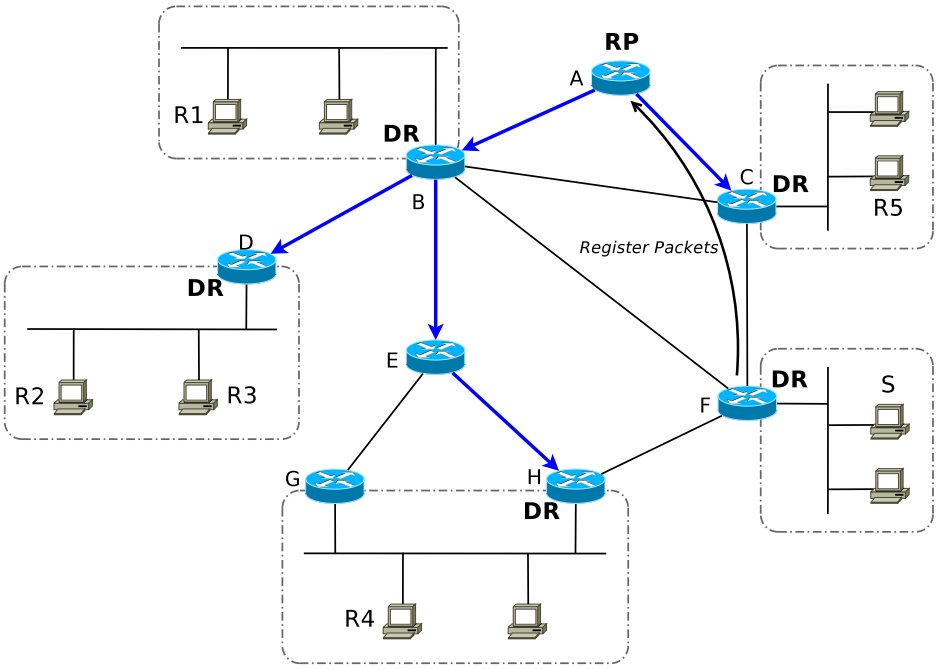


Abbildung 2.13: PIM-SM – Phase One - Registering



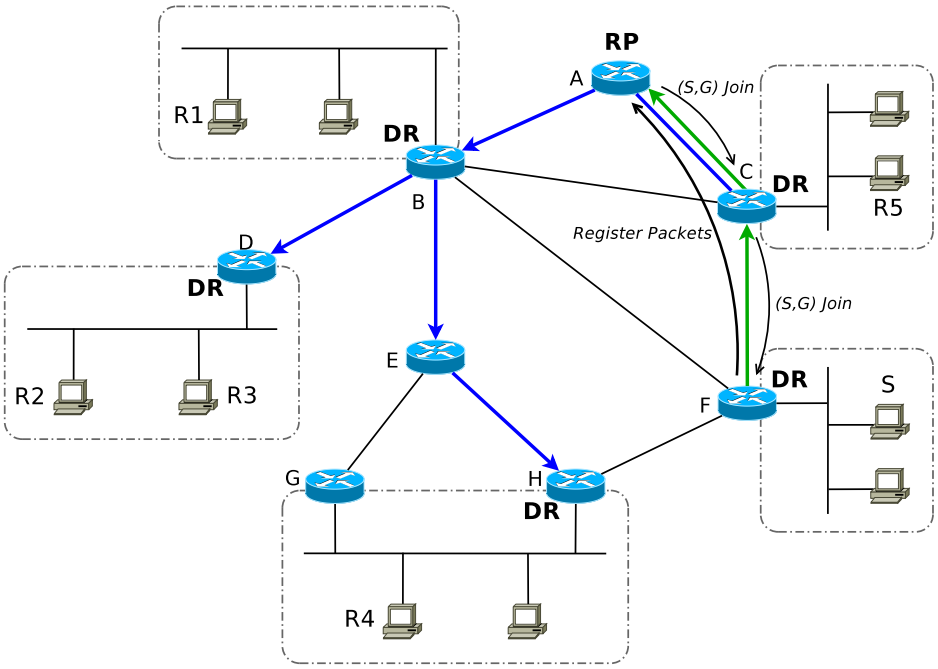
kapselt die Multicast Pakete in IP-Pakete und verschickt sie so per Unicast direkt an den Rendezvous Point. Die IP-Pakete heißen Register Packets, das Tunneln wird als Registering bezeichnet. Am RP angekommen werden die Register Packets entpackt und die enthaltenen Multicast Pakete über den RP Tree zugestellt.

Phase Two Das Registering hat zwei Nachteile. Zum einen ist das Tunneln für die Router recht aufwendig. Das Ver- und Entpacken der Multicast-Pakete kostet einiges an Rechenzeit. Zum anderen werden Pakete eventuell eine Strecke hin und wieder zurück geroutet, wie es auch im Beispiel der Fall ist. So führt der Pfad der Register Packets in Abbildung 2.13 über die Strecke CA, und über dieselbe Strecke wird wiederum das Multicast-Paket zurückgeleitet, welches beim Empfänger R5 ankommt.

Deshalb sendet der RP, sobald er Register Packets von einem DR erhält, einen (S,G) Join - eine senderspezifische Anmeldung für die Multicastgruppe - an den DR gerichtet aus (siehe Abbildung 2.14). Wenn der Join den DR erreicht hat, ist ein senderspezifischer Pfad (Source-specific Tree) vom DR zum RP etabliert, dem entlang auch unmittelbar Multicast Pakete vom Sender S folgen. Sobald der RP von C nativ Multicast-Pakete erhält sendet er sie nicht wieder zurück zu Router C. Schließlich würde dies zu doppelten Paketen führen.

Alsbald erhält der RP die Multicast Pakete doppelt, einmal als Register Packets und einmal über den Source-specific Tree. Nun verschickt er eine Register-Stop Nachricht an den DR,

Abbildung 2.14: PIM-SM – Phase Two - (S,G) Join



der, sobald er diese Nachricht erhalten hat, das Senden der Register Packets einstellt (siehe Abbildung 2.15).

Am Ende der zweiten Phase werden die Multicast Pakete zunächst entlang des Source-specific Tree und anschließend entlang des RP Tree verteilt, wie es in Abbildung 2.16 gezeigt ist. Der Multicast Traffic fließt nun vollständig nativ und nicht mehr getunnelt.

Phase Three Wenn die Multicast Sitzung von relativ langer Dauer und hohem Traffic-Aufkommen ist, mag sich der Aufwand lohnen, vom kombinierten Baum weg zu einem optimalen Baum - dem Source-specific Shortest Path Tree (SPT) - hin zu gelangen. So erreichen die Multicast Pakete den Router H erst nach fünf Hops, obwohl der sendende DR (Router F) nur einen Hop entfernt ist.

Hierbei geht die Initiative von den DRs aus, die in ihren angeschlossenen Subnetzwerken Empfänger haben. Diese senden ein (S,G) Join an den sendenden DR (siehe Abbildung 2.17), welcher entlang des kürzesten Pfads vom emfangenden zum sendenden DR gelangt und somit über diesen neuen Pfad ebenfalls den Empfang der Multicast Pakete bewirkt.

Jetzt erhalten die empfangenden DRs die Multicast Pakete doppelt. Einmal über den SPT und über den RPT. Um den Empfang über den RPT abzubestellen, wird ein (S,G,rpt) Prune an den RP ausgesandt (siehe Abbildung 2.18).

Abbildung 2.15: PIM-SM – Phase Two - Register Stop

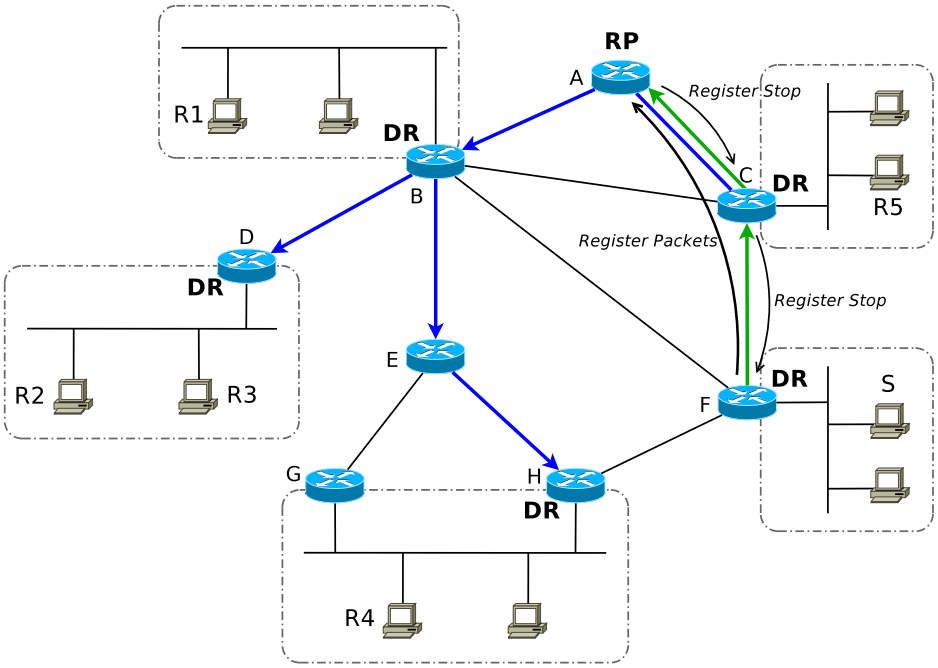


Abbildung 2.16: PIM-SM – Phase Two - Nativer Fluss

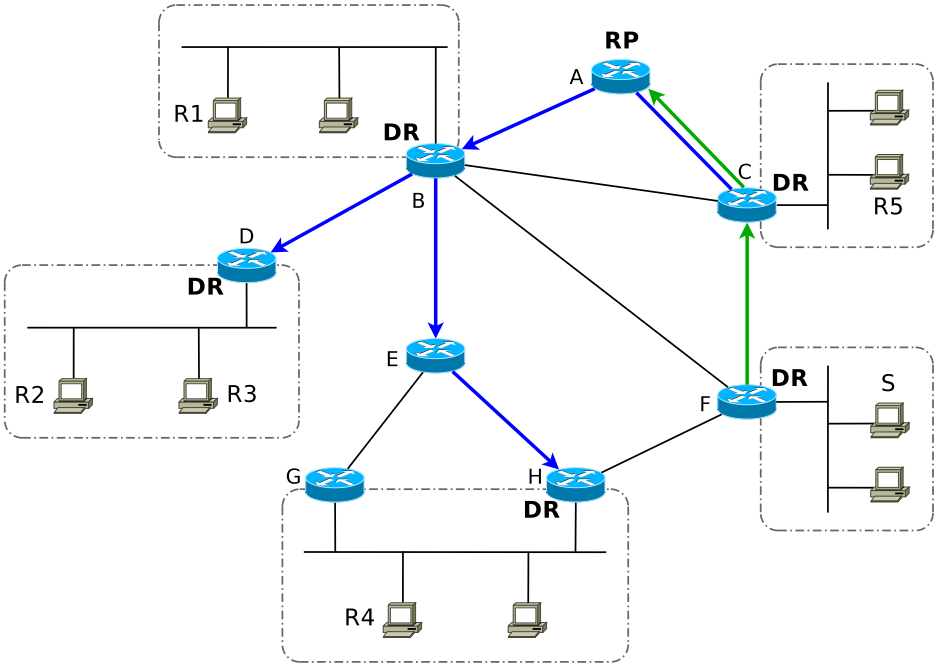


Abbildung 2.17: PIM-SM – Phase Three - (S,G) Join

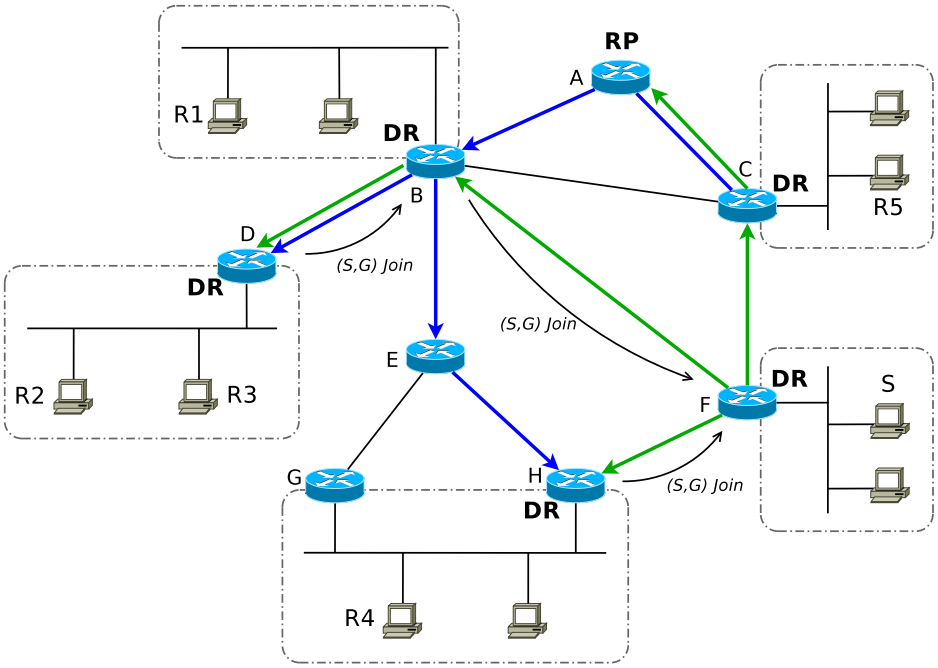


Abbildung 2.18: PIM-SM - Phase Three - (S,G,rpt) Prune

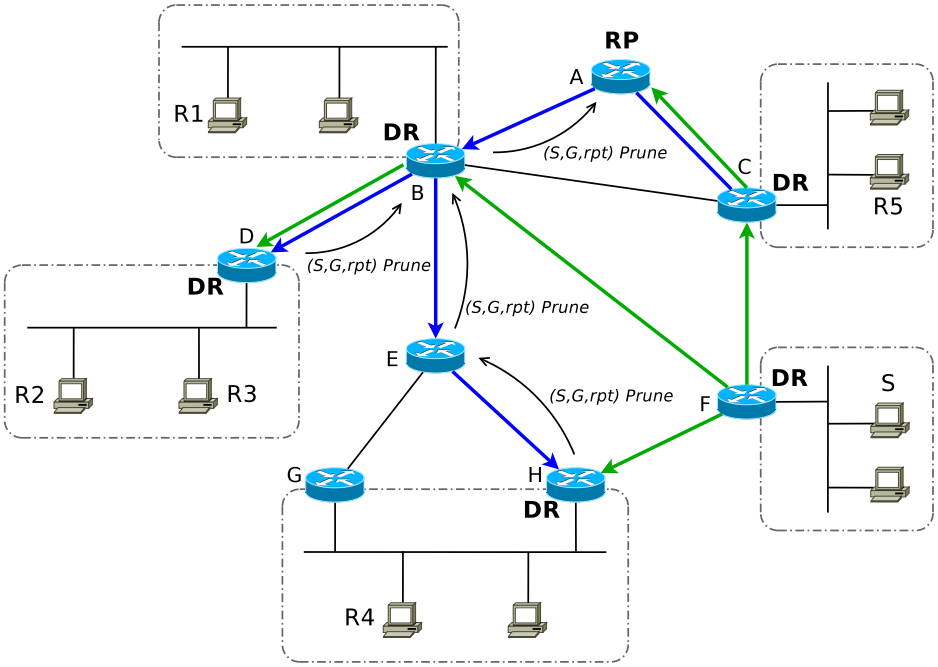


Abbildung 2.19: PIM-SM – Phase Three - Shortest-Path Tree

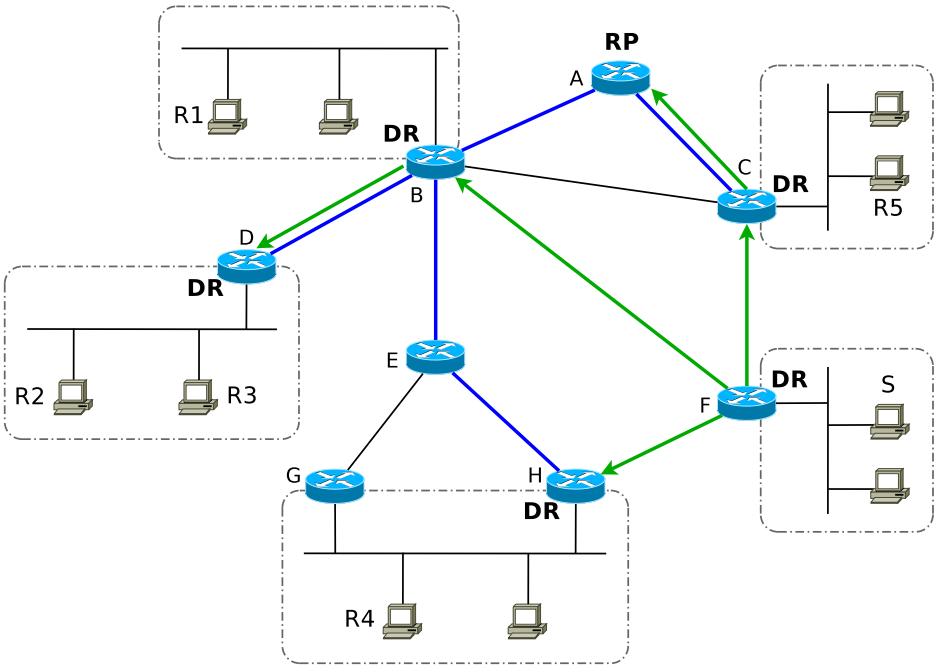


Abbildung 2.19 zeigt den entgültigen Source-specific Shortest-Path Tree.

2.1.3.2 PIM-SSM

PIM-SSM ist ein vereinfachter Modus von PIM-SM und ist ebenfalls in dem PIM-SM RFC 4601 (18) beschrieben (in Abschnitt 4.8). Er zeichnet sich dadurch aus, daß es weder Any Source-Multicast Gruppenmitgliedschaften noch Any Source Routing States (*,G) gibt. Es gibt ausschließlich den Source Specific Multicast (S,G) und somit nur Source Specific States.

Konkret hat das zur Folge, daß es keinen Rendezvouspoint und somit keinen Shared Tree wie den RP Tree von PIM-SM gibt, sondern ausschließlich Source-specific Shortest Path Trees.

PIM-SSM ist eher für den Fall weniger allgemein bekannter Sender geeignet. Dann hat es gegenüber PIM-SM zum einen den Vorteil, daß das Routing vereinfacht ist, weil der Aufwand des RP Tree wegfällt, und zum anderen, daß Multicast Pakete ausschließlich über den optimalen SPT verteilt werden.

2.1.3.3 BIDIR-PIM

Bidir-PIM ist definiert in RFC 5015 (22). Einen anschaulichen Einblick in die Funktionsweise bieten zwei Dokumente von der Firma Cisco Systems (23), (24).

Pakete einer Multicast-Gruppe werden ausschließlich entlang eines gemeinsam genutzten bidirektionalen Baumes (bidirectional shared tree), dem RP Tree, geroutet. Dessen Wurzel ist der Rendezvous Point (RP) der Gruppe. Bidirektional bedeutet, daß Multicast-Pakete sowohl zur Wurzel (aufwärts) als auch von der Wurzel weg (abwärts) weitergeleitet werden.

Anders als bei PIM-SM ist der RP virtualisiert. Er wird identifiziert über ein IP-Adresse. Diese muß nicht zwingend einer Instanz, einem existierenden Router zugewiesen sein; es reicht aus, daß sie zu einem Subnetz gehört, welches innerhalb der Bidir-PIM Domain erreichbar ist.

Wie bei PIM-SM werden die Informationen über Multicast Gruppenmitgliedschaften über explizite Join- bzw. Prune-Nachrichten übermittelt. Traffic von Sendern wird ohne Bedingungen (unabhängig von den Gruppenmitgliedschaften möglicher Empfänger) nativ den RP Tree aufwärts bis zum RP weitergeleitet. Anschließend und wenn es vorher Verzweigungen zu Empfängern gibt, werden die Multicast-Pakete abwärts entlang des RP Tree weitergeleitet.

Gegenüber PIM-SM fällt nun die Notwendigkeit und somit der Aufwand weg, daß ein Designated Router (DR) per Register Messages bzw. per Source-specific Shortest Path Tree (SPT) Multicast Pakete zum RP weiterreicht. Register-Messages führen zu Prozessorlast des DR beim Kapseln der Multicast-Pakete in Unicast-Pakete. Und je größer die Anzahl der Sender ist, um so größer wird der Rechenaufwand für das Entpacken dieser Pakete, wenn sie beim RP angekommen sind. Viele SPTs vieler Sender bedeuten viele (S,G)-Einträge in den Routingtabellen der beteiligten Router und können ein Ressourcen-Problem darstellen. So ist die Aufgabe des RPs – nämlich die Wurzel des gemeinsamen Spannbaumes zu identifizieren – rein passiv.

Eine weitere Funktion des DRs in PIM-SM ist es sicherzustellen, daß genau ein Router von mehreren möglichen Routern, die an ein Subnetz angrenzen, Multicast-Pakete von Sendern des Subnetzes Domain-weit verfügbar macht und empfangene Multicast-Pakete an die Empfänger des Subnetzes weiterleitet.

Diesen Part übernimmt in Bidir-PIM der Designated Forwarder (DF). Auch der DF ist innerhalb eines Subnetzes einmalig. Er wird unter den Routern, die an das Subnetz angrenzen, durch eine Designated Forwarder Election bestimmt. Dabei gewinnt der Router mit der besten Unicast-Metrik zum RP.

Das Routen von Multicast-Paketen aufwärts zum RP erfolgt nach dem Reverse Path Forwarding Prinzip. Dazu werden die Multicast-Pakete jeweils über das RPF-Interface eines

Bidir-PIM Routers zum Reverse Path Forwarding (RPF) Neighbor in Richtung zum RP weitergeleitet.

Bidir-PIM ist nicht ein „besseres“ Multicast-Routing als PIM-SM oder PIM-SSM. Zwar fällt der höhere Aufwand durch senderspezifische Routing-Zustände weg, noch muß ein Tunneling von Multicast-Paketen betrieben werden. Aber dafür gibt es keinen optimalen Source-specific Shortest Path Tree.

So zielt Bidir-PIM auf Anwendungsfälle mit vielen Sendern und Empfängern ab (many-to-many). PIM-SSM hingegen ist für den Fall weniger Sender oder nur einem Sender gedacht – auch hier entfällt der Aufwand des Tunnelns von Multicast-Paketen. PIM-SM ermöglicht generell ein Multicast-Routing, z. B. wenn man sich nicht auf viele oder wenige Sender festlegen kann.

2.2 Application-Layer Multicast

Beim Application-Layer Multicast (ALM) werden die drei Aufgabenbereiche der Gruppenkommunikation, Gruppenmanagement, Pflege des Verteilbaums und Weiterleiten von Multicast-Paketen, nicht mehr inhärent im Netzwerk behandelt. Stattdessen sind sie auf Anwendungs-Ebene angesiedelt. Die am ALM beteiligten Knoten erzeugen ein virtuelles Netzwerk (Overlay), abstrahiert vom darunter liegenden nativen Netzwerk (Underlay). Die Links des Overlays werden hierbei aus den Ende-zu-Ende Pfaden (auf Ebene der Netzwerk-Schicht) zwischen den Knoten gebildet, d.h. die Knoten kommunizieren auf IP-Layer Ebene (lediglich) per Unicast-Verbindungen. Für die Realisierung des Overlays kommen verschiedene Peer-to-Peer Netzwerke infrage.

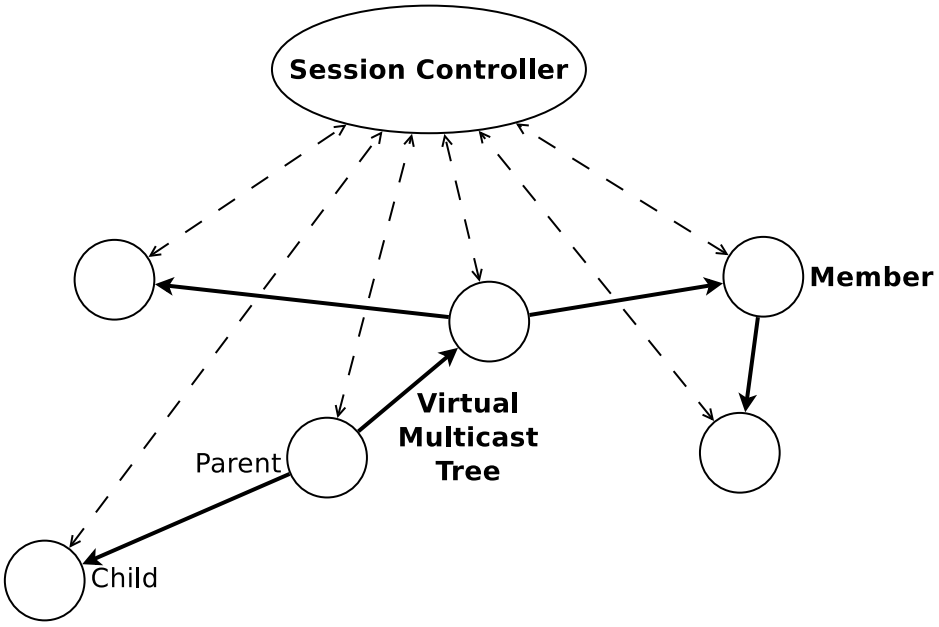
Die folgende Einteilung der ALM-Implementierungen leitet sich aus der Klassifizierung der verwendeten Peer-to-Peer Netzwerke ab. Diese lassen sich in unstrukturierte und strukturierte Peer-to-Peer Netzwerke unterteilen.(25)

2.2.1 Unstrukturierte Overlays

2.2.1.1 Zentralisierte Architekturen

Zentralisierte Application-Layer Multicast Routing-Architekturen zeichnen sich dadurch aus, daß das Gruppenmanagement und die Erzeugung und Pflege des Multicast-Routing Baumes von einer zentralen Instanz aus gesteuert werden. Das Weiterleiten der Multicast-Pakete geschieht im Sinne eines effizienten Routings nicht über eine zentrale Instanz, sondern über den Multicast-Routing Baum, der sich über die Gruppenmitglieder aufspannt.

Abbildung 2.20: Vereinfachte ALMI-Architektur. Nachgezeichnet und erweitert; Quelle: (2)



Application Level Multicast Infrastructure Ein Beispiel für den zentralisierten Ansatz ist die Application Level Multicast Infrastructure (ALMI) (26).

Abbildung 2.20 zeigt das Prinzip der ALMI-Architektur. Das Weiterleiten von Multicast-Nachrichten findet über einen gemeinsam genutzten Baum statt (Shared Tree). Dieser Baum ist bidirektional, d.h. Multicast-Nachrichten können über die Links des Baumes in beide Richtungen weitergeleitet werden. In ALMI ist die zentrale Instanz der sogenannte Session Controller. Diese ist, um einen Flaschenhals zu vermeiden, nicht Teil des gemeinsamen Multicast-Routing Baumes. Nur Gruppenmitglieder können Multicast-Nachrichten versenden.

Ein Gruppenmitglied kommt hinzu, indem es sich beim Session-Controller anmeldet (JOIN-Nachricht). Es bekommt als Antwort vom Session Controller eine Member-ID zugewiesen, das ist der Identifier im Overlay. Zudem wird dem neuen Knoten mitgeteilt, welcher andere Knoten sein Parent Node ist (somit steuert der Session Controller den Aufbau des Baumes). An diesen sendet er nun eine sogenannte GRAFT-Nachricht um die Ports für eine bidirektionale Kommunikation zu erhalten. Um eine Multicast-Sitzung zu beenden, sendet ein Node ein LEAVE an den Session Controller.

Den Shared Tree erzeugt der Session Controller, indem er anhand einer Metrik der Linkkosten zwischen den Knoten (z. B. Verzögerung) einen minimalen Spannbaum berechnet, der alle Gruppenmitglieder erreicht. Jedem Knoten wird dabei der Parent Node (Upstream) zugeteilt. So wird gewährleistet, daß von einem beliebigen Knoten des Baumes versende-

te Multicast-Nachrichten an alle anderen Multicast-Knoten genau einmal (mindestens und höchstens einmal) zugestellt werden.

Wenn nun jeder der n Knoten des Baumes die Linkkosten zu den anderen Knoten ausmessen und das Ergebnis dem Session Controller mitteilen würde, ergäbe das einen Overhead von $O(n^2)$. Deshalb ist in ALMI die Anzahl der Nachbarknoten begrenzt, zu denen die Linkkosten ausgemessen werden. Diese Lösung führt nicht zu einem minimalen Spannbaum. Weil aber die Knoten fortwährend Nachbarknoten mit Links, die große Kosten aufweisen, durch Knoten mit günstigeren Links austauschen, wird der Multicast-Baum über die Zeit optimiert.

Die Vorteile einer zentralisierten Architektur liegen in ihrer Einfachheit der Implementierung und der hohen Kontrolle über die Topologie des Overlays. Fehlerhafte Knoten können aufgrund der zentralen Kontrolle leicht entdeckt und umgangen werden. Wie jeder zentralen Architektur ist ihr der Nachteil zu eigen, daß das ganze System nicht mehr funktioniert, sobald die zentrale Instanz (der Session Controller) ausfällt. Dieses Problem kann man durch Backup-Instanzen lösen oder zumindest mindern, die im Fehlerfall für den Session Controller einspringen. Ein weiteres Problem der zentralen Architektur ist jedoch nicht gelöst, sie skaliert nicht. Je mehr Gruppenmitglieder es gibt, um so größer ist die Netzlast aufgrund der Kommunikation der Knoten mit dem Session Controller.

2.2.1.2 Vollständig verteilte Architekturen

In vollständig verteilten Application-Layer Multicast Routing-Architekturen gibt es keine zentrale Instanz. Sämtliche Aufgaben des Gruppenmanagements, der Koordinierung sowie Bildung und Pflege des Routing-Baumes werden von den Knoten übernommen.

Narada Ein Beispiel für diese Architektur ist End System Multicast (ESM) (27). Während in „traditionellen“ Netzwerk-Architekturen zwischen Hosts (End Systems) und den Entitäten des eigentlichen Netzwerks (Router, Switche) unterschieden wird, werden sämtliche Aufgaben der Multicast-Kommunikation von den Hosts übernommen. Das ebenfalls in (27) vorgestellte Protokoll Narada realisiert den Ansatz von ESM.

Narada ist ein Mesh-First Multicast Routing-Protokoll. Zunächst bilden die Knoten des Overlays einen redundanten Graphen (Mesh). Danach werden senderspezifische Multicast-Bäume ((S,G) Trees) gebildet, welche Teilgraphen des Mesh sind. Eine Zielsetzung von Narada ist es, ein effizientes Overlay-Netzwerk zu erhalten; effizient bezüglich des Underlays und effizient gegenüber der Anwendung, die die Multicast-Infrastruktur nutzt. Ersteres bedeutet, daß die Redundanz und der daraus resultierende Linkstress auf den physikalischen Links des Underlays eines Pfades zweier Knoten möglichst minimal ist. Effizient gegenüber

der Anwendung bedeutet beispielweise bei Audio-Konferenz oder Gruppen-Messaging Anwendungen, daß eine geringe Verzögerung gewährleistet wird. Bei einer Video-Konferenz Anwendung muß das Overlay zudem eine hohe Bandbreite gewähren, während bei einer Videoübertragung (nur eines Senders) die Verzögerung eine geringere Rolle spielt, es bei der Bildung des Baumes hauptsächlich nur auf die Bandbreite der Pfade ankommt.

Das Gruppenmanagement ist auf alle Knoten gleich verteilt. Jeder Knoten verfolgt die Präsenz aller anderen Gruppenmitglieder, einerseits um die Robustheit des Protokolls zu erhöhen, andererseits um eine effiziente Wartung des Mesh zu ermöglichen. Dies wird erreicht, indem jeder Knoten in regelmäßigen Abständen eine Broadcast-Nachricht (Heartbeat Message) an alle Gruppenmitglieder über das Mesh verschickt. Setzt der Herzschlag von einem Knoten aus, so wird dies von einem anderen Knoten bemerkt und dieser sendet eine Anfrage an den verstummten Knoten, ob dieser noch im Overlay präsent ist. Bleibt nun eine Antwort aus, so wird er aus der Liste der aktiven Gruppenmitglieder ausgetragen.

Neue Knoten kommen hinzu, indem sie in Kontakt mit einer kleinen Anzahl von bereits im Overlay vorhandenen Knoten treten. Somit stellen sie initial eine Verbindung zum Overlay her und können eine Heartbeat-Message broadcasten, so daß alle anderen Knoten von ihrer Existenz erfahren. Die Identifier (ID im Overlay) und die Locator (IP-Adresse im Underlay) der initialen Knoten müssen separat, beispielsweise per E-Mail Benachrichtigung kommuniziert werden.

Das Overlay ist dynamischen Änderungen unterworfen. Die Linkkosten der Pfade, nach denen das Mesh bzw. die Multicast-Bäume gebildet werden, können sich mit der Zeit ändern. Knoten kommen hinzu oder verschwinden aus dem Overlay (entweder weil sie ihre Sitzung beenden oder aufgrund eines Fehlers). Deswegen wird das Mesh wiederkehrend bewertet und neu aus den Pfaden zwischen den Knoten gebildet.

Der Algorithmus, um das Mesh zu bilden, ist ein Distance-Vektor Algorithmus. Die senderspezifischen Verteilbäume werden nach dem Reverse-Path Forwarding Prinzip (wie es auch in PIM zur Anwendung kommt) etabliert.

Narada ist durch seine verteilte Eigenorganisation relativ robust. Weil jeder Knoten dem anderen in seiner Funktionalität gleicht („echte“ Peers), es keine Knoten mit „besonderen“ Funktionen gibt, ist der Konfigurationsaufwand relativ gering – Narada stellt eine ready-to-deploy Lösung dar. Ein weiterer Vorteil von Narada ist, daß das Protokoll nicht auf eine bestimmte Metrik festgelegt ist und somit für unterschiedliche Anwendungsfälle optimiert werden kann. Allerdings skaliert Narada nicht für große Gruppen, hauptsächlich aufgrund des großen Umfangs an Routing-Nachrichten, die zwischen den Knoten ausgetauscht werden.

2.2.2 Strukturierte Overlays

Unstrukturierte Application-Layer Multicast-Architekturen skalieren nicht unbeschränkt, weil mit zunehmender Gruppenzahl die Netzlast durch den Overhead zu stark ansteigt. Der Anforderung, auch bei großer Anzahl von tausenden von Gruppenmitgliedern eine Overlay Multicast-Kommunikation zu ermöglichen, versuchen die Strukturierten Peer-to-Peer Netzwerke gerecht zu werden.

Strukturierte Peer-to-Peer Netzwerke zeichnen sich dadurch aus, daß die Knoten in einem virtuellen geordneten Raum platziert sind. Daraus resultiert, daß es eine Ordnung auf die Bezeichner der Knoten – den Peers – gibt. Man sagt, sie sind strukturiert.

Es wird unterschieden, ob die Weiterleitung von Multicast-Nachrichten durch direktes Fluten oder durch Bildung eines Verteilbaumes erfolgt.

2.2.2.1 Flooding

Tanenbaum (28) beschreibt das Fluten (Flooding) als einen statischen Algorithmus, bei dem ein an einem Knoten über eine Kante eingehendes Paket über jede andere Kante weitergeleitet wird.

Um die „Flut“ von Paketen einzudämmen und (Endlos-) Schleifen zu unterbinden, ist eine zusätzliche Maßnahme erforderlich: Ein bereits weitergeleitetes Paket wird, wenn es erneut eintrifft, nicht wiederholt weitergeleitet sondern verworfen. Alternativ zu diesem Vorgehen oder in Kombination dazu kann ein Paket verworfen werden, wenn der Hop-Counter des Pakets den Durchmesser des Netzes überschreitet. (Dies berücksichtigt den Worst-Case, wenn der Urheber des Pakets am Rand des Netzwerkes liegt.)

Flooding ist kein Routing-Algorithmus, bei dem Aufwand betrieben wird, um den oder die ausgehenden Links für die Weiterleitung von Paketen auszuwählen. Vielmehr ist es eine Form des gerouteten Broadcast, d.h. ein Algorithmus zur Verteilung von Nachrichten (bzw. Paketen) an alle Empfänger in einem (möglicherweise virtuellen) Netzwerk.

Content-Addressable Network Dieses Prinzip des Forwardings wendet das Content-Adressable Network (CAN) (29; 30) für das Weiterleiten von Multicast-Nachrichten an.

CAN verwendet einen d-dimensionalen endlichen kartesischen Koordinatenraum. Dieser wird dynamisch in unterschiedliche Zonen unterteilt, die jeweils eine endliche Anzahl zusammenhängender Koordinaten jeder der d Dimensionen umfasst. Über eine Hash-Funktion wird jede Zone einem Knoten des Overlays zugeordnet. Zwei Knoten sind Nachbarn in einer

Dimension i, wenn sich ihre Zonen in d-1 Dimension überlappen und in genau der i-ten Dimension unterscheiden. Beim Routen von Unicast-Nachrichten wird eine Nachricht an den Nachbarn, der näher am Zielknoten liegt, weitergeleitet, bis es schließlich am Empfänger ankommt. Ein neuer Knoten kommt hinzu, indem von einem sogenannten Bootstrap-Knoten koordiniert ein anderer Knoten einen Teil seiner Zone an den neuen Knoten abtritt. Anders herum wird beim Weggang eines Knotens dessen Zone mit der eines Nachbarknotens vereint.

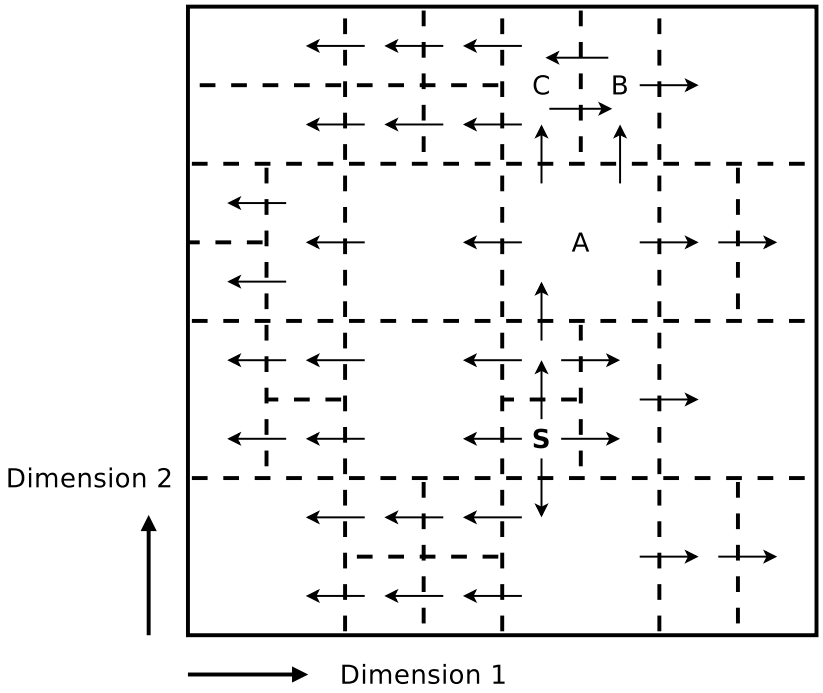
Multicast-Kommunikation (31) wird ermöglicht, indem zunächst die Knoten, welche Mitglied einer Gruppe G sind, ein „Mini“-CAN bilden. Über eine Hash-Funktion wird die Gruppenadresse G auf eine Koordinate des d-dimensionalen Koordinatensystems abgebildet. Der Knoten (des „Haupt“-CANs – er ist nicht notwendiger Weise Mitglied von G), in dessen Zone die Koordinate liegt, wird der Bootstrap-Knoten des „Mini“-CANs. Die Knoten der Gruppe G werden anschließend nach dem CAN-Algorithmus in das „Mini“-CAN hinzugefügt, bekommen entsprechend Zonen zugeteilt.

Multicast-Nachrichten können nun über das „Mini“-CAN geflutet werden. Für das Weiterleiten der Nachrichten gelten folgende Regeln:

1. Der Sender leitet eine Multicast-Nachricht an alle Nachbarknoten weiter.
2. Erhält ein Knoten eine Multicast-Nachricht von einem Nachbarn der Dimension i, so leitet er sie an alle Nachbarn der Dimensionen $1 \ldots (i-1)$ und an die Nachbarn der Dimension i, die in entgegengesetzter Richtung angrenzen, aus der das Paket herstammt, weiter. An Nachbarn höherer Dimension als i wird die Nachricht nicht weitergeleitet.
3. Empfängt ein Knoten eine Multicast-Nachricht, die mindestens die Hälfte der Strecke einer Dimension des Koordinatenraumes durchlaufen hat, so wird sie nicht mehr weitergeleitet.
4. Jeder Knoten merkt sich die Sequenznummer einer weitergeleiteten Nachricht und kann daran ein weiteres Eintreffen dieser Nachricht erkennen, um ein wiederholtes Weiterleiten zu unterdrücken.

Die ersten beiden Regeln stellen sicher, daß jeder Knoten der Gruppe G mindesten einmal eine Nachricht erhält. Mit der dritten Regel werden Schleifen vermieden. Abbildung 2.21 zeigt ein Beispiel für das Fluten einer Multicast-Nachricht vom Sender S in einem zweidimensionalen „Mini“-CAN. Wenn der Koordinatenraum nicht optimal unterteilt ist, kann es dazu kommen, daß ein Knoten mehrere Kopien von einer Nachricht von unterschiedlichen Nachbarknoten erhält. So erhält im Beispiel der Knoten mit der Zone B und der Knoten mit der Zone C jeweils zweimal die Multicast-Nachricht.

Abbildung 2.21: Beispiel für den Versand von Multicast-Nachrichten in einem zweidimensionalen CAN. Nachgezeichnet und leicht abgeändert; Quelle: (2)



2.2.2.2 Verteilbäume

BIDIR-SAM BIDIR-SAM steht für Bidirectional Scalable Adaptive Multicast. (32). Das Weiterleiten von Multicast-Nachrichten erfolgt über einen Präfix-basierten virtuellen Verteilbaum, dessen Knoten dynamisch auf die Knoten der darunter liegenden verteilten Hash-Tabelle (Distributed Hash Table, DHT) wie etwa Pastry (33) abgebildet werden.

2.2.3 Benennung in Peer-to-Peer Netzwerken

Ein Problem, das nicht nur in Peer-to-Peer Systemen anzutreffen ist, sondern im Grunde in allen Bereichen, in denen auf Entitäten (z.B. Ressourcen, Funktionseinheiten oder Standorte) zugegriffen wird, ist das Problem der Benennung. (34)

Eine Entität (z. B. eine Webpage) hat einen oder mehrere Zugriffspunkte (z. B. HTTP-Server). Solch ein Zugriffspunkt hat eine Adresse (z. B. IP-Adresse). Auf die Entität verweist genau ein Name, der Bezeichner (z. B. URI). Der wesentliche Unterschied zwischen der Adresse und dem Bezeichner ist der, daß der Bezeichner fest mit der Entität verknüpft ist, also immer auf die gleiche Entität verweist. Eine Adresse eines Zugriffspunktes auf eine Entität

hingegen kann sich ändern. Wie beim Aufruf einer Webseite fällt es oft schwer, zwischen der Entität und dem Zugriffspunkt auf diese Entität zu unterscheiden (ganz genau betrachtet ist der Zugriffspunkt selber auch eine Entität). Bei der Benennung geht es darum, Entitäten (genauer: Zugriffspunkte auf die Entitäten) anhand ihrer Bezeichner zu lokalisieren, d.h. zu einem Bezeichner die (aktuelle) Adresse zu erfahren.

Die große Herausforderung in Peer-to-Peer Netzwerken hierbei ist es, das Nachschlagen (den Lookup) einer Adresse von einem Bezeichner in vollständig dezentraler Weise mit skalierbarem Aufwand zu erreichen.

Es gibt folgende Lösungsansätze für Algorithmen, um den Lookup in Peer-to-Peer Systemen zu bewerkstelligen:

Zentrale Datenbasis Es gibt eine zentrale Datenbasis der Zuordnungen von den Bezeichner auf die Knoten. Dieser Ansatz skaliert nicht, denn mehr Knoten bedeuten auch mehr Zugriffe auf diese eine Datenbasis. Auch bedeutet dieser zentrale Ansatz einen Single Point of Failure (und zugleich, daß die Kontrolle über das gesamte System auf einem Punkt konzentriert ist). Sobald der Zugriff auf die Datenbasis nicht mehr möglich ist, können keine Knoten mehr gefunden werden. Napster basiert auf diesem Ansatz.

Broadcast an alle Nachbarknoten Eine Anfrage zur Auflösung eines Bezeichners wird an alle bekannten Nachbarknoten weitergeleitet. Auch dieser Ansatz skaliert nicht, denn je mehr Knoten und somit Suchanfragen entstehen, desto größer ist die Netzlast hierdurch. Gnutella hat dieses Verfahren eingesetzt, was auch die Ursache dafür war, daß es unter seiner selbst erzeugten Netzlast zusammengebrochen ist, sobald eine genügend große Anzahl von Usern Dateien geshared haben.

Feste Hierarchie Die Knoten sind in einer festen Hierarchie angeordnet (vergleichbar der Hierarchie beim Domain Name System (DNS). Auch dieser Ansatz hat ähnlich dem Ansatz einer zentralen Datenbasis den Nachteil, daß, je hierarchisch höher und somit näher an der Wurzel ein Knoten platziert ist, zum einen eine höhere Traffic Last entsteht und zum anderen ein Ausfall des Knotens sich auf größere Ausfälle des Systems auswirkt.

Vollständig verteilter Lookup Hier wird ein Lookup von Knoten zu Knoten geroutet, bis der referenzierte Knoten erreicht ist. Die Datenbasis der Zuordnungen ist über die Knoten verteilt. Jeder Knoten verwaltet dabei eine begrenzte Anzahl von Zuordnungen. Es gibt keine Hierarchie unter den Knoten – jeder Knoten ist gleichberechtigt. Ein Knoten heißt daher Peer. Die Organisation, also das Hinzukommen von Peers in die Peer-to-Peer Struktur und das Verschwinden wird mit relativ geringem Aufwand ebenfalls vollständig dezentral erreicht. Freenet, als ein Vertreter des vollständig verteilten Lookups, hat zusätzlich als hehres Ziel Anonymität. Um dies zu erreichen, muß es ausgeschlossen sein, eine Verbindung zwischen Dokument und dessen Herkunft im Netzwerk herzu-

stellen. Das hat zur Folge, daß selten getauschte Dokumente nicht immer zuverlässig gefunden werden können.

CAN (29; 30), Chord (35), Pastry (33) und Tapestry (36) sind neuere Ansätze für einen vollständig verteilten Lookup. Ihre Eigenschaften wie Skalierbarkeit sind gut untersucht worden. Ihnen ist gemeinsam, daß sie auf verteilten Hashtabellen aufbauen.

Bei einer verteilten Hashtabelle, auf englisch Distributed Hash Table (DHT) werden die Zuordnungen von Bezeichner auf Knoten über die teilnehmenden Knoten selber verteilt. Die Bezeichner heißen Keys. Ein Key wird über eine Hashfunktion auf ein Datum gebildet. Dieses Datum kann eine (binäre) Datei sein oder auch die IP-Adresse von Knoten.

Für das Nachschlagen einer Adresse eines Knotens zu einem Key stellt eine DHT folgende Operation bereit: `lookup(key)`; die DHT ist somit eine Lösung des Lookup-Problems.

Um dies zu erreichen ist folgendes zu berücksichtigen:

- Keys werden auf Nodes gemappt. Dies muß load-balanced erfolgen (vermeiden einer Hierarchie), indem eine gute Hash-Funktion verwendet wird.
- Ein Lookup, der von einem Knoten nicht direkt beantwortet werden kann, ist an einen Knoten weiterzuleiten, der dem aufzulösenden Knoten näher ist.
- Zu diesem Zweck pflegt ein Knoten eine Routing Tabelle in der zu einem Key der *Successor*, ein dem Zielknoten näherer Knoten, nachgeschlagen werden kann.

Die Algorithmen der oben erwähnten Ansätze unterscheiden sich stark in der Art und Weise, wie sie ihre Routing-Tabellen bilden und pflegen und wie das Hinzufügen und Entfernen von Knoten geregelt ist.

2.3 Hybrider Multicast

Der hybride Multicast ist eine Kombination von IP-Layer- und Application-Layer Multicast. Es gibt Domains, in denen ein nativer Multicast möglich ist. Über die Domain-Grenzen hinweg jedoch findet kein IP-Layer Multicast statt. Um einen Multicast zu ermöglichen, der die einzelnen Inseln – das Underlay – überspannt, werden die Inseln durch eine grenzüberschreitende Overlay Multicast (OLM) Domain – dem Overlay – miteinander verbunden. Das Overlay hat die Funktion eines Routing Backbones. Sender und Empfänger befinden sich in den nativen Domains.[4] Die OLM Domain, ist ein Peer-to-Peer Netzwerk, auf dem die Funktionalität eines Multicast aufsetzt. In jeder IP-Layer Multicast Domain gibt es einen Peer. Er handelt als Stellvertreter nach außen hin, d.h. als Proxy, indem er die lokalen Multicast-Zustände

[4] Es ist auch denkbar, daß es Multicast-Empfänger im Overlay gibt. (37)

(Sender und Empfänger) erfährt, in das Overlay gegebenenfalls weiterreicht und Multicast-Pakete entsprechend in die lokale Domain bzw. in das Overlay weiterleitet. Mit anderen Worten, wenn Multicast-Sender und -Empfänger in unterschiedlichen IP-Layer Domains liegen, werden die Multicast-Pakete über das Overlay von der Domain des Senders zur Domain des Empfängers weitergereicht.

2.3.1 Motivation

In vielen Edge Domains ist ein IP-Layer Multicast bereits vorhanden. Aber es mangelt an einem Internet-weiten IP-Layer Multicast. Gegenüber dem Application-Layer Multicast hat der IP-Layer Multicast den Vorteil des deutlich geringeren Overheads. Der IP-Layer Multicast ist effektiver hinsichtlich der Netzlast als der Application-Layer Multicast. So ist es sinnvoll, den IP-Layer Multicast dort zu nutzen, wo er verfügbar ist.

2.3.2 Hybrid Shared Tree Architektur

Die Hybrid Shared Tree (HST) Architektur (4; 3) ist ein Vorschlag, einen hybriden Multicast zu realisieren. Als Overlay dient der auf Pastry aufsetzende Scalable Adaptive Multicast on Bi-directional Shared Trees (BIDIR-SAM) (4).

Kennzeichnend für die HST-Architektur ist, daß das Overlay für Sender und Empfänger transparent ist. Ein netzwerkübergreifender Multicast wird somit ermöglicht, ohne daß Sender und Empfänger an ihren Implementierungen des nativen Multicast Änderungen vorzunehmen hätten.

Das Inter-domain Multicast Gateway Das Herzstück der HST-Architektur ist das Inter-domain Multicast Gateway (IMG). Abbildung 2.22 zeigt beispiehaft voneinander abgegrenzte Domains, in denen jeweils IP-Layer Multicast-Kommunikation verfügbar ist. Über Border Router sind sie mit dem Internet Backbone verbunden, so daß Unicast-Kommunikation zwischen den Domains möglich ist. Die IMGs spannen ein Overlay auf.

Das IMG ist der Vermittlungspunkt zwischen dem Overlay und dem Underlay. In jeder IP-Layer Multicast Domain gibt es ein IMG. Dessen Aufgabe ist es, die IP-Layer Multicast Domain mit ihren Multicast Zuständen in der OLM Domain zu vertreten. Im Peer-to-Peer-Netzwerk, dem Overlay, sind die IMGs die Peers; Multicast Pakete, die über das Overlay geleitet werden, werden also zwischen den IMGs verschickt. Das IMG handelt als ein Proxy für die lokale Domain: Den Traffic von lokalen Sendern macht das IMG im Overlay für Empfänger in anderen IP-Layer Domains verfügbar. Auch sorgt das IMG dafür, daß Empfänger der lokalen Domain gegebenenfalls Multicast Pakete von Sendern aus anderen IP-Layer

Abbildung 2.22: Beispielhafte Darstellung der Hybrid Shared Tree Architektur mit dem von den IMGs abstrahierten Overlay. Quelle: (3)

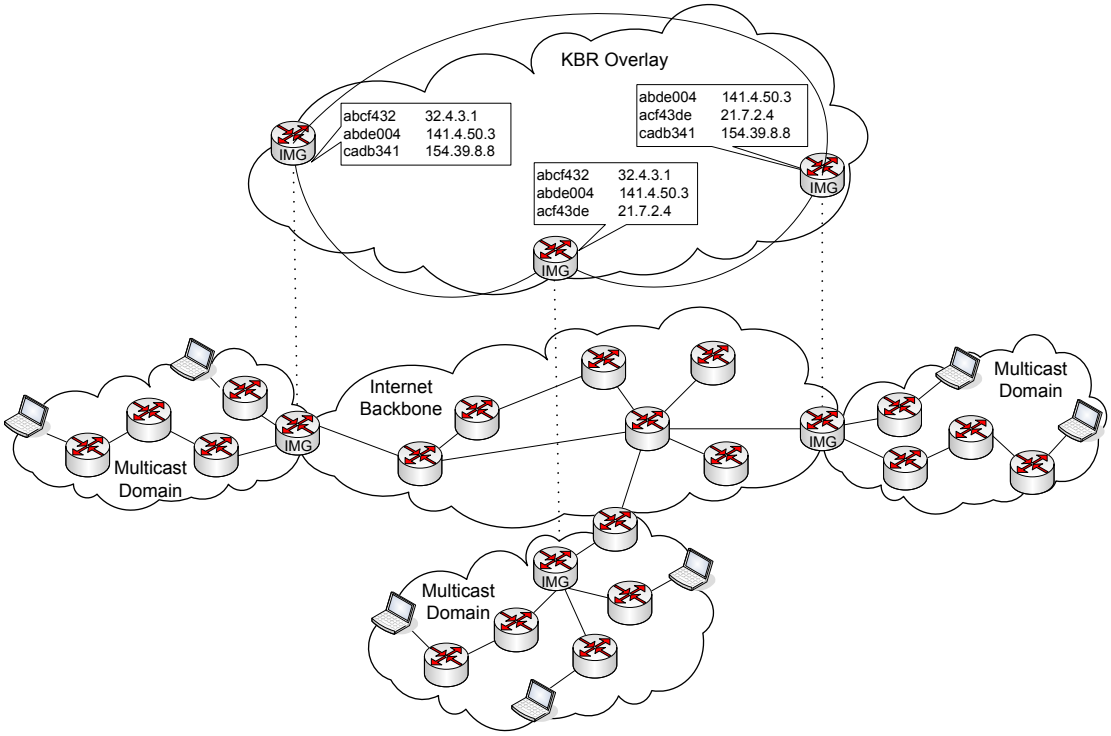
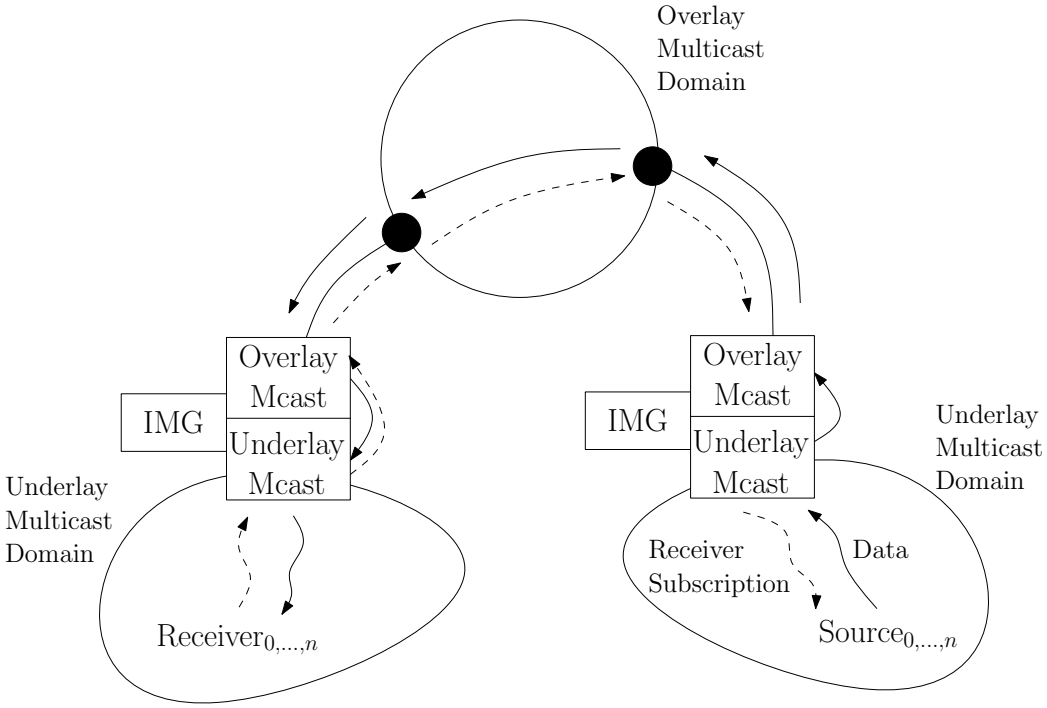


Abbildung 2.23: Schematische Sicht eines HST-Multicast Szenarios Quelle: (3)



Domains erhalten. Abbildung 2.23 zeigt den Versand bzw. Empfang von Multicast-Paketen, wobei Sender in einer anderen IP-Layer Multicast Domain liegen als Empfänger. Multicast Pakete werden von der Ursprungsdomain über die Overlay Multicast Domain an die der Empfänger weitergereicht.

Bezüglich der Interaktion des IMG mit der lokalen Domain wird unterschieden, ob ein IP-Layer Multicast Routing-Protokoll in der Domain Anwendung findet oder nicht. Eine Domain ohne ein Multicast Routing-Protokoll wird als Small-Size Domain bezeichnet. Weist die Domain ein Routing-Protokoll auf, so handelt es sich um eine Large-Size Domain.

3 Platzierung des Inter-domain Multicast Gateways

3.1 Das Service-Placement Problem

Wir widmen uns der Fragestellung, wo das IMG innerhalb einer IP-Layer Multicast Domain optimal platziert werden kann.

Dazu vergleichen wir unterschiedliche Netzwerkaufstellungen miteinander, indem wir die genutzten Verbindungen – die Links – anhand von Kosten-Metriken bewerten.

Es gibt unterschiedliche Kosten-Metriken, nach denen eine optimale Platzierung erfolgen kann. Unter anderem gibt es folgende Metriken:

- Verzögerung
- Anzahl von Hops
- monetäre Kosten

Eine (Link-) Kosten-Metrik ist eine Abbildung der Kosten eines Links (einer Verbindung) auf einen numerischen Wert. Damit sind die Linkkosten untereinander vergleichbar, und sie können miteinander verrechnet werden.

Für die Linkkosten gilt:

1. Linkkosten fallen pro verschicktem Paket pro Link an.
2. Je größer das Nachrichtenpaket ist, desto größer sind die Linkkosten.
3. Für unterschiedliche Links können für ein Paket unterschiedliche Linkkosten anfallen.
4. Wird ein Paket über mehrere Links übertragen, so können die jeweiligen Linkkosten zu den Gesamtlinkkosten aufsummiert werden.

Ziel für eine optimale Platzierung ist es, eine Netzwerkaufstellung zu finden, in der die Gesamtlinkkosten minimal sind. Insbesondere fragen wir uns bei einem gegebenen Netzwerk, an welchem Ort für das IMG wir die geringsten Gesamtlinkkosten erhalten.

Linkkosten in einer IP-Layer Multicast Domain Um die Linkkosten in einer IP-Layer Multicast Domain bei einer HST-Architektur zu erfassen, die beim Versand von Multicast-Paketen entstehen, kann man zunächst unterscheiden, ob sich ein Sender innerhalb oder außerhalb der Domain befindet. Aus Sicht des IMG unterscheiden wir also zwischen eingehenden und ausgehenden Multicast-Traffic. Es sei gesagt, daß es mehrere Sender geben kann – die Betrachtung muß dann für jeden Sender separat erfolgen.

Liegt der Sender innerhalb der Domain, so abonniert das IMG (sobald es von dem Sender erfährt) den vom Sender erzeugten Traffic. Das IMG ist aus Sicht der IP-Layer Multicast Domain ein Multicast-Empfänger. Nun werden, sofern es Empfänger außerhalb der Domain gibt, die Pakete in das Overlay weitergeleitet. Gibt es kein Abonnement auf die Multicast-Pakete außerhalb der Domain, so werden auch keine Pakete in das Overlay weitergereicht (das IMG abonniert aber dennoch den Traffic, siehe Abschnitt 2.3.2).

Bei lokalen Empfängern, und wenn sich der Sender außerhalb der Domain befindet, gelangen die nativen Multicast-Pakete über den Border-Router (BR) als OLM-Pakete gekapselt in die Domain und erreichen das IMG. Das IMG entkapselt dann die Multicast-Pakete und reicht sie an die Domain weiter, entsprechend dem vorherrschenden Verteilungsmechanismus (direkt im Subnetz, IGMP/MLD-Proxying, PIM-SM, PIM-SSM oder Bidir-PIM).

3.2 Small-Size Domain

Eine Small-Size Domain besteht aus einem IP-Netzwerk. Sie ist dadurch gekennzeichnet, daß kein Multicast-Routing Protokoll zum Einsatz kommt.

Gruppenmanagement wird ermöglicht durch IGMP/MLD. Das IP-Netzwerk entspricht einer IGMP/MLD-Domain. Innerhalb dieser Domain signalisieren die Empfänger ihre Gruppenzugehörigkeiten über IGMP/MLD-Reports. Diese sind adressiert an eine bestimmte Multicast-Adresse, so daß (potentiell) vorhandene IGMP/MLD-Router die lokalen Gruppenmitgliedschaften verfolgen können. Das IMG übernimmt bezüglich IGMP/MLD den Router-Part.

Mehrere IGMP/MLD-Domains können durch IGMP/MLD-Proxys miteinander verbunden sein, ohne daß ein IP-Layer Multicast Routing-Protokoll verwendet wird. Auch hier übernimmt das IMG den IGMP/MLD Router-Part. Platziert in der Wurzel-Domain bzw. beim hierarchisch höchsten IGMP/MLD-Proxy kann es die Mitgliedschaften aller Empfänger verfolgen.

Abbildung 3.1: IMG platziert in einer Small-Size Domain ohne IGMP/MLD Proxying

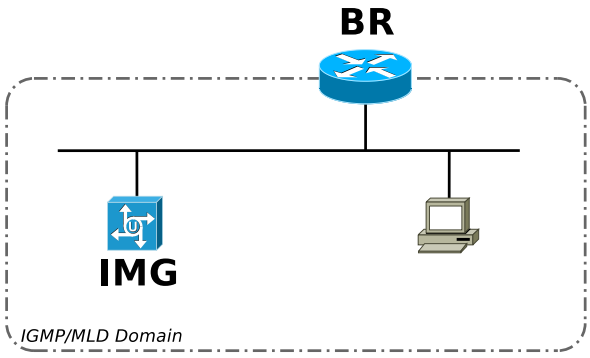
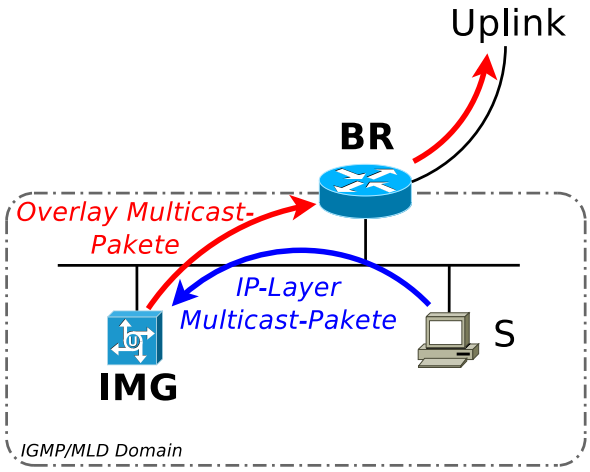


Abbildung 3.2: Small-Size Domain – Weiterleitung von lokalem Multicast-Traffic in das Overlay



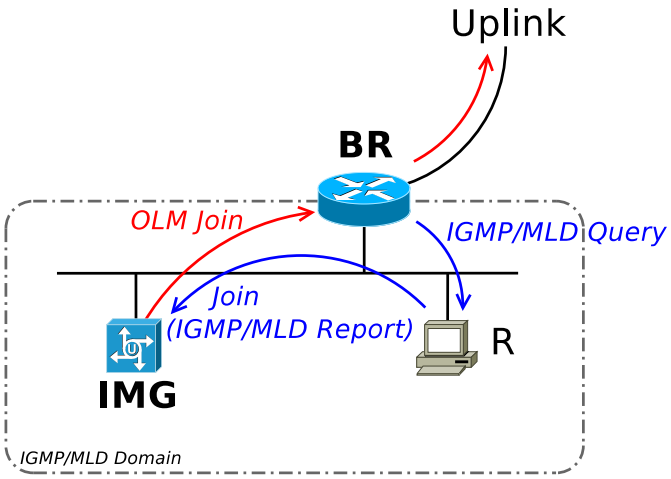
3.2.1 Kein IGMP/MLD-Proxying

Abbildung 3.1 zeigt eine aus einer einzelnen IGMP/MLD-Domain bestehende Small-Size Domain. Hier gibt es nur die Möglichkeit, das IMG beim Border Router (BR), d.h. innerhalb derselben IGMP/MLD-Domain zu platzieren.

Ausgehender Verkehr Bei einem lokalen Sender greift das IMG die IP-Layer Multicast-Pakete auf und reicht sie als Overlay-Multicast Pakete in das Overlay weiter, sofern es im Overlay ein Abonnement auf die Multicast-Pakete gibt (siehe Abbildung 3.2).

Eingehender Verkehr Wir betrachten den Empfang von Multicast-Nachrichten aus dem Overlay. Abbildung 3.3 zeigt die Gruppensignalisierung auf. Der Empfänger versendet ein

Abbildung 3.3: Small-Size Domain – Signalisierung



Join als Antwort auf das Query des Router-Interfaces. Dieses Join gelangt über den IGMP/MLD-Proxy weiter zum IMG. Das IMG abonniert daraufhin im Overlay den Traffic der entsprechenden Multicast-Gruppe (OLM-Join).

Nun erhält das IMG über den BR Multicast-Pakete aus dem Overlay, entpackt diese und reicht sie nativ weiter, so daß der Empfänger R sie über das IGMP/MLD-Proxy zugestellt bekommt (siehe Abbildung 3.4).

3.2.2 Mit IGMP/MLD-Proxying

Ausgehender Verkehr Besteht die Small-Size Domain aus mehreren per IGMP/MLD-Proxys miteinander verbundenen IGMP/MLD-Domains, so muß das IMG in der Wurzel-IGMP/MLD Domain platziert sein. Denn, wie wir in Kapitel 2.1.2 festgestellt haben, wird sämtlicher Multicast-Traffic (Multicast-Signalisierungen und -Pakete) zur Root-Domain weitergeleitet. In Abbildung 3.5 ist ein Szenario einer Small-Size Domain, bestehend aus mehreren IGMP/MLD-Domains, dargestellt. Das IMG ist in der Wurzel-IGMP/MLD Domain platziert. Es bezieht den Multicast-Traffic über den IGMP/MLD Proxy vom Senders S, der sich in einer hierarchisch niederen IGMP/MLD Domain befindet. Wäre das IMG in der IGMP/MLD-Domain A platziert, so würde es keine Multicast-Pakete des Senders S aus der IGMP/MLD-Domain B erhalten. In die Root-IGMP/MLD Domain aber werden die Multicast-Pakete vom IGMP/MLD-Proxy weitergeleitet, auch wenn es dort keinen Host gibt, der Multicast-Pakete von S abonniert hat.

Abbildung 3.4: Small-Size Domain – Empfang von Multicast-Paketen aus dem Overlay

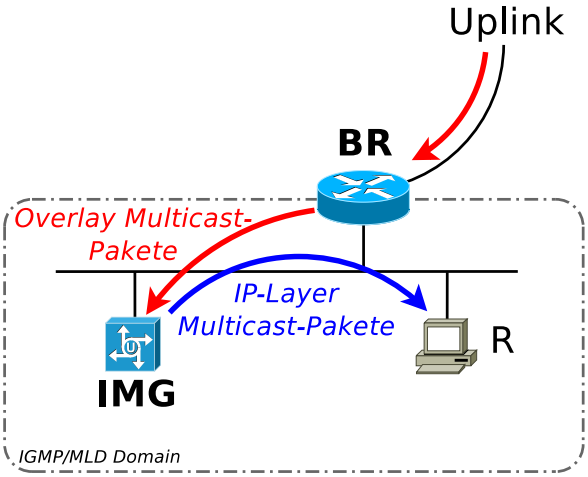


Abbildung 3.5: Das IMG ist in einer Small-Size Domain mit IGMP/MLD Proxying in der Wurzel-IGMP/MLD Domain platziert.

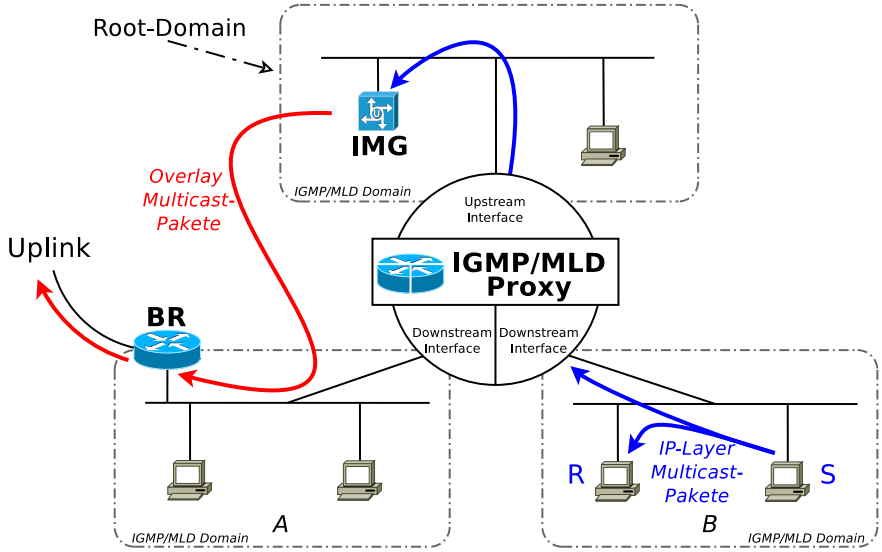
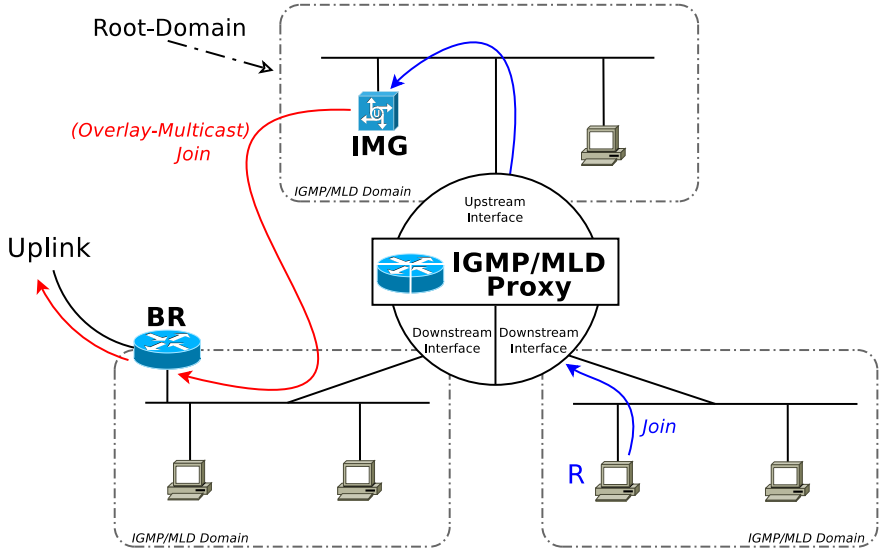


Abbildung 3.6: IGMP/MLD-Proxying – Abonnement von Multicast-Traffic aus dem Overlay



Eingehender Verkehr Der Empfänger R, welcher in einer IGMP/MLD-Domain niederer Hierarchie liegt, signalisiert seine Gruppenmitgliedschaft durch ein IGMP/MLD-Join. Dieses leitet das IGMP/MLD-Proxy weiter in die Root-Domain. Das dort platzierte IMG greift das Join auf und abonniert über das Overlay die entsprechende Multicast-Gruppe (siehe Abbildung 3.6).

Sofern über das Overlay Multicast-Pakete der abonnierten Gruppe verschickt werden, werden diese an das IMG weitergeleitet. Das IMG leitet sie dann nativ als IP-Layer Multicast-Pakete weiter. Über das IGMP/MLD-Proxy gelangen erreichen sie schließlich den Empfänger R (siehe Abbildung 3.7).

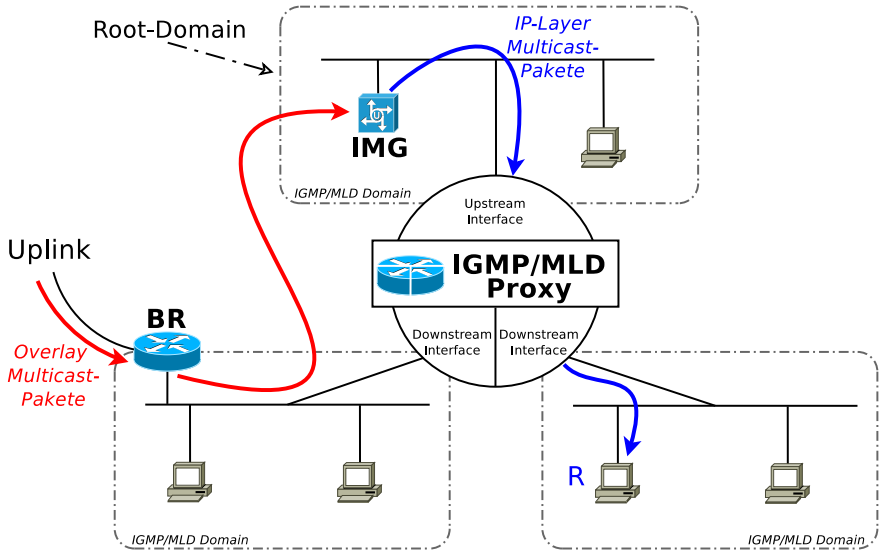
3.3 Large-Size Domain

Eine Large-Size Domain zeichnet sich dadurch aus, daß IP-Layer Multicast-Pakete anhand eines IP-Layer Multicast Routing-Protokolls weitergeleitet werden. Auch hier muß das IMG von allen Gruppenmitgliedschaften und Sendern in der Domain erfahren. Wie dies dem IMG ermöglicht wird, hängt von dem verwendeten Routing-Protokoll ab.

3.3.1 PIM-SM

In PIM-SM signalisiert der Rendezvous-Point (RP) einer Multicast-Gruppe dem IMG, ob ein lokaler Sender aktiv ist. Daraufhin abonniert das IMG die Multicast-Pakete des Senders, in-

Abbildung 3.7: Das IMG ist in einer Small-Size Domain mit IGMP/MLD Proxying in der Wurzel-IGMP/MLD Domain platziert.



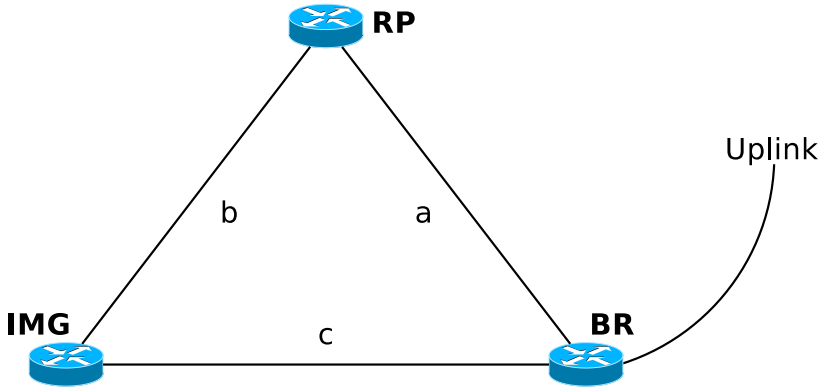
dem es der Multicast-Gruppe beitritt. Das IMG wird ebenfalls vom RP über lokale Empfänger einer Multicast-Gruppe informiert. So kann es gegebenenfalls im Overlay der entsprechenden Multicast-Gruppe beitreten, um den aus dem Overlay empfangenen Multicast-Traffic transparent an die lokalen Sender weiterzuleiten.

Ausgehender Verkehr Folgendes Szenario betrachten wir nun: Der Sender befindet sich innerhalb der PIM-SM Domain. Er verschickt seine Pakete per Tunnel oder per Source-specific Shortest Path Tree (SPT) an den Rendezvous-Point (RP). Es gibt externe Empfänger, so daß das IMG den Traffic vom Sender empfängt, kapselt und in das Overlay über den Border Router (BR) weiterleitet.

Als erstes ist eine allgemeine Topologie zu bilden, die alle möglichen konkreten Topologien abdeckt. Da der Sender S Multicast-Pakete ausschließlich über den RP verschickt, bezieht das IMG die Pakete vom RP. Diese werden dann gekapselt und an den BR weitergeleitet. Relevant ist demnach die Platzierung des IMG relativ zum RP und zum BR. So betrachten wir einen vollständigen Graphen, der die drei Entitäten als Knoten hat. Abbildung 3.8 zeigt diesen vollständigen Graphen.

Die Kanten sind mit den Linkgewichten a, b und c versehen. $a, b, c \in \{0, 1, 2, \cdots; \infty\}$. a beispielsweise ist ein Maß für die Linkkosten, die entstehen, wenn ein Paket vom RP zum BR weitergeleitet wird, ohne daß es über das IMG geroutet wird:
$Linkkosten(RP \rightarrow BR, direkt) = a * l$. Ist a = 0, so bedeutet dies, daß der RP und BR auf demselben Knoten liegen. Dann gilt auch: b = c.

Abbildung 3.8: Vollständiger Graph mit RP, IMG und BR als Knoten und den Linkgewichten a, b und c.



Hat a den Wert ∞, so gibt es keine Verbindung zwischen dem RP und dem BR, die nicht über das IMG führt. Die drei Knoten bilden dann eine Achse, bei der der Knoten mit dem IMG zwischen dem RP und dem BR liegt. Für b und c gelten diese Eigenschaften entsprechend.

Das Routen der Multicast-Pakete vom Sender zum RP geschieht unabhängig vom RP-Tree. Die Linkkosten hierfür werden als konstant angenommen, denn sie sind unabhängig von den Linkkosten der Verteilung über den RP-Tree an die Empfänger (einschließlich dem IMG).

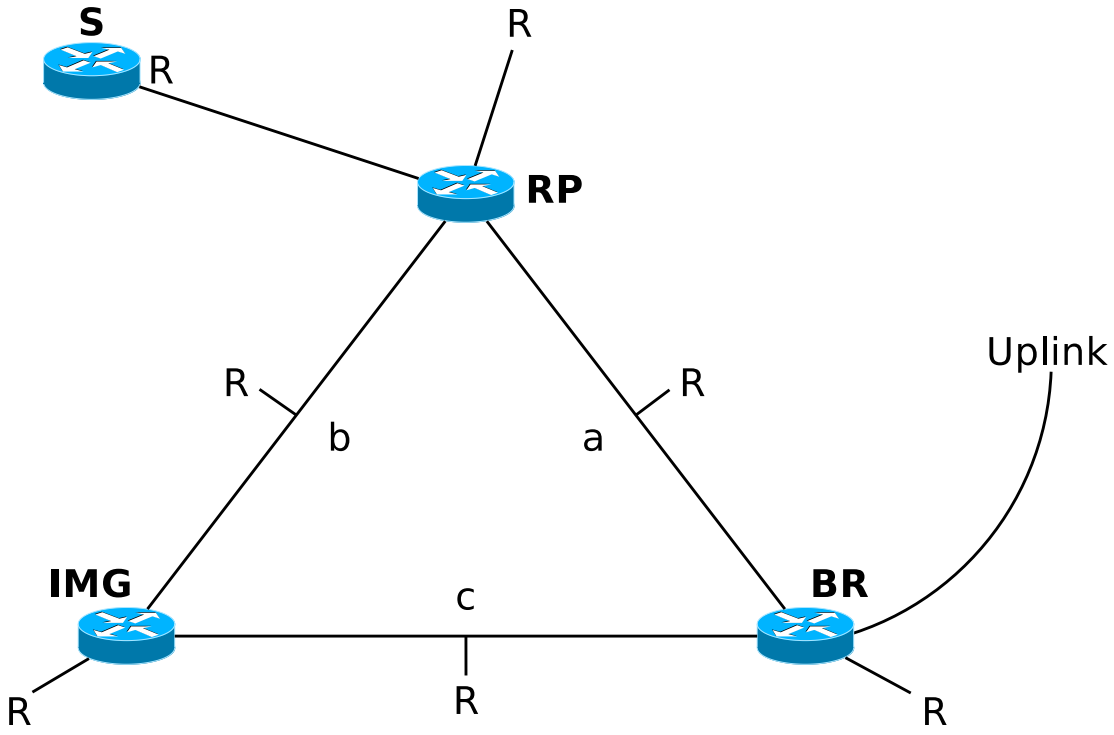
Empfänger bzw. Router, hinter denen Empfänger liegen (Empfänger-Subnetze), können bei den Knoten und zwischen den Knoten (an den Kanten) liegen. Abbildung 3.9 zeigt dies auf. Für die Linkkosten ist es nicht entscheidend, ob in einem Empfänger-Subnetz mehrere Empfänger liegen, oder nur ein Empfänger liegt.

Die Linkkosten der Empfänger betrachten wir nicht im Einzelnen, sondern nur als eine gesamte, von der Anzahl der Empfänger abhängige Größe. Da wir diese Kosten bereits in den Linkkosten der Empfänger erfasst haben, werden sie nicht bei den Linkkosten des IMGs angerechnet.

Wenn $b < a + c$ ist, werden IP-Layer Multicast-Pakete vom RP zum IMG über den Pfad mit den Linkkosten a geroutet. Teilen sich lokale Empfänger sich den Pfad mit dem IMG (komplett oder anteilig), so verringert dies die Linkkosten des IMG. Der Faktor, um den die Lnkkosten reduziert werden nennen wir u. Wegen der gerade beschriebenen Eigenschaft gilt: $0 \leq u \leq 1$.

Der Pfad vom IMG zum BR ist mit einem Overlay-Faktor o gewichtet. Wegen des Overheads des Application-Layer Multicast gilt für den Overlay-Faktor: $o < 1$.

Abbildung 3.9: Wie Abbildung 3.8, erweitert um einen Sender S und mögliche Platzierungen von Empfängern.



Zusammengefasst gilt für die Linkkosten L_{IMG} des IMG dann:

$$L_{IMG} = u * b + o * c. \tag{3.1}$$

Diese Gleichung wird minimal, wenn c gegen 0 geht.

Demnach ist das IMG beim BR optimal platziert.

3.3.2 Bidir-PIM

Bei Bidir-PIM gelangen alle IP-Layer Multicast-Pakete lokalen Ursprungs über den bidirektionalen Baum zur Wurzel, dem Rendezvous-Point (RP). Auch die Informationen über die Gruppenmitgliedschaften der lokalen Empfänger gelangen entlang des bidirektionalen Baumes zur Wurzel. Daher muß das IMG im selben Subnetz liegen wie der RP.

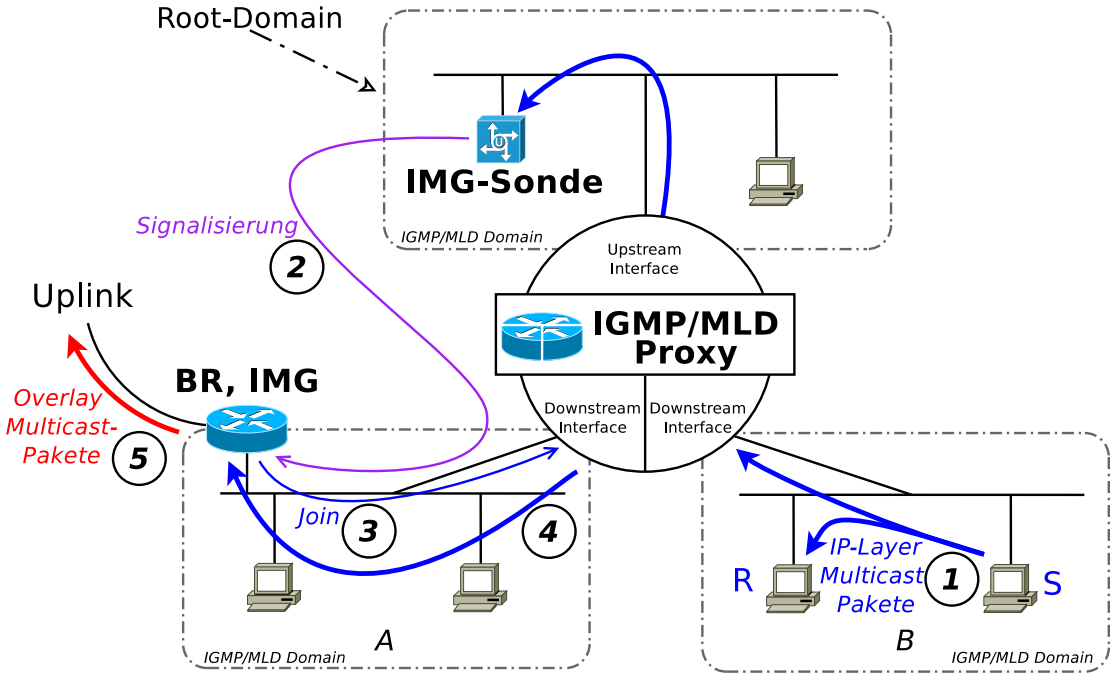
Im Grunde ist das die gleiche Situation wie beim IGMP/MLD-Proxying: Gruppensignalisierungen und Multicast-Pakete gelangen über einen gemeinsamen Baum zur Wurzel des Baumes.

3.4 Ausblick

Beim IGMP/MLD-Proxying und bei Bidir-PIM könnte man eine IMG-Sonde an der Wurzel und das eigentliche IMG beim Border Router (BR) platzieren. Die IMG-Sonde weiß, welche lokalen Sender und Empfänger es gibt. Diese Informationen signalisiert sie weiter an das eigentliche IMG.

Gibt es einen lokalen Sender, so abonniert das IMG nativ den IP-Layer Multicast. Abbildung 3.10 stellt diese Situation in einer Small-Size Domain mit IGMP/MLD-Proxying dar.

Abbildung 3.10: Das IMG ist zweigeteilt, sowohl beim BR als auch in der Wurzel-IGMP/MLD Domain platziert. IMG beim BR abonniert lokalen Traffic.



Literaturverzeichnis

[1] 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.

[2] R. Steinmetz and K. Wehrle, Eds., *Peer-to-Peer Systems and Applications*, ser. LNCS. Berlin Heidelberg: Springer-Verlag, 2005, vol. 3485.

[3] M. Wählisch and T. C. Schmidt, “Multicast Routing in Structured Overlays and Hybrid Networks,” in *Handbook of Peer-to-Peer Networking*, X. Shen, H. Yu, J. Buford, and M. Akon, Eds. Berlin Heidelberg: Springer Verlag, January 2010, to appear. [Online]. Available: http://www.springer.com/computer/communications/book/978-0-387-09750-3

[4] ——, “Between Underlay and Overlay: On Deployable, Efficient, Mobility-agnostic Group Communication Services,” *Internet Research*, vol. 17, no. 5, pp. 519–534, November 2007. [Online]. Available: http://www.emeraldinsight.com/10.1108/10662240710830217

[5] J. D. Day and H. Zimmermann, “The osi reference model,” *Proceedings of the IEEE*, vol. 71, no. 12, pp. 1334–1340, 1983. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1457043

[6] J. Day, “The (un)revised osi reference model,” *SIGCOMM Comput. Commun. Rev.*, vol. 25, no. 5, pp. 39–55, 1995.

[7] R. Braden, “Requirements for Internet Hosts - Communication Layers,” IETF, RFC 1122, October 1989.

[8] R. Wittmann and M. Zitterbart, *Multicast Communication*. Morgan Kaufmann, San Francisco, Calif., 2001.

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

[10] R. Hinden and S. Deering, “IP Version 6 Addressing Architecture,” IETF, RFC 4291, February 2006.

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

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

[13] A. Conta, S. Deering, and M. Gupta, “Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification,” IETF, RFC 4443, March 2006.

[14] W. C. Fenner, “Internet Group Management Protocol, Version 2,” IETF, RFC 2236, November 1997.

[15] S. E. Deering, W. C. Fenner, and B. Haberman, “Multicast Listener Discovery (MLD) for IPv6,” IETF, RFC 2710, October 1999.

[16] M. Christensen, K. Kimball, and F. Solensky, “Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches,” IETF, RFC 4541, May 2006.

[17] A. Adams, J. Nicholas, and W. Siadak, “Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised),” IETF, RFC 3973, January 2005.

[18] B. Fenner, M. Handley, H. Holbrook, and I. Kouvelas, “Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised),” IETF, RFC 4601, August 2006.

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

[20] J. Moy, “Multicast Extensions to OSPF,” IETF, RFC 1584, March 1994.

[21] Y. K. Dalal and R. M. Metcalfe, “Reverse path forwarding of broadcast packets,” *Commun. ACM*, vol. 21, no. 12, pp. 1040–1048, 1978.

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

[23] *Bidirectional PIM Deployment Guide*, Cisco Systems, Inc., 2008. [Online]. Available: http://www.cisco.com/en/US/prod/collateral/iosswrel/ps6537/ps6552/ps6592/prod_white_paper0900aecd80310db2.pdf

[24] *Bidirectional PIM [Cisco IOS Software Releases 12.1 T]*, Cisco Systems, Inc., 2000. [Online]. Available: http://www.cisco.com/en/US/docs/ios/12_1t/12_1t2/feature/guide/dtbipim.pdf

[25] K. Katrinis and M. May, “Application-layer multicast,” in *Peer-to-Peer Systems and Applications*, ser. LNCS, R. Steinmetz and K. Wehrle, Eds., vol. 3485. Berlin Heidelberg: Springer-Verlag, 2005.

[26] D. Pendarakis, S. Shi, D. Verma, and M. Waldvogel, “Almi: An application level multicast infrastructure,” in *3rd USENIX Symposium on Internet Technologies and Systems (USITS ’01)*.

[27] Y.-H. Chu, S. G. Rao, and H. Zhang, "A case for end system multicast," in *SIGMETRICS '00: Proceedings of the 2000 ACM SIGMETRICS international conference on Measurement and Modeling of Computer Systems*. New York, NY, USA: ACM Press, 2000, pp. 1–12.

[28] A. S. Tanenbaum, *Computer Networks, Fourth Edition*. Prentice Hall PTR, August 2002. [Online]. Available: http://www.worldcat.org/isbn/0130661023

[29] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker, "A Scalable Content-Addressable Network," in *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*. New York, NY, USA: ACM, 2001, pp. 161–172.

[30] S. P. Ratnasamy, "A Scalable Content-Addressable Network," Ph.D. dissertation, University of California, Berkeley, October 2002.

[31] S. Ratnasamy, M. Handley, R. M. Karp, and S. Shenker, "Application-level multicast using content-addressable networks," in *NGC '01: Proceedings of the Third International COST264 Workshop on Networked Group Communication*. London, UK: Springer-Verlag, 2001, pp. 14–29.

[32] M. Wählisch, T. C. Schmidt, and G. Wittenburg, "BIDIR-SAM: Large-Scale Content Distribution in Structured Overlay Networks," in *Proc. of the 34th IEEE Conference on Local Computer Networks (LCN)*, M. Younis and C. T. Chou, Eds. Piscataway, NJ, USA: IEEE Press, October 2009, pp. 372–375.

[33] 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.

[34] H. Balakrishnan, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica, "Looking up data in p2p systems," *Commun. ACM*, vol. 46, no. 2, pp. 43–48, 2003.

[35] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Protocol for Internet Applications," *IEEE/ACM Trans. Netw.*, vol. 11, no. 1, pp. 17–32, 2003.

[36] B. Y. Zhao, J. D. Kubiatowicz, and A. D. Joseph, "Tapestry: An Infrastructure for Fault-tolerant Wide-area Location and," University of California at Berkeley, Technical Report UCB/CSD-01-1141, April 2001.

[37] M. Wählisch, T. C. Schmidt, and S. Venaas, "A Common API for Transparent Hybrid Multicast," individual, IRTF Internet Draft – work in progress 01, October 2009. [Online]. Available: http://tools.ietf.org/html/draft-waehlisch-sam-common-api

Versicherung über Selbständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit im Sinne der Prüfungsordnung nach §24(5) ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 8. Januar 2010	
Ort, Datum	Unterschrift