Seminararbeit

Core-Based Trees und Protocol Independent Multicast (Sparse Mode)

vorgelegt am
Lehrstuhl für Praktische Informatik IV
Prof. Dr. Wolfgang Effelsberg
Universität Mannheim

im
Dezember 1999

von
Jürgen Dufner

Betreuer:
Dr. Thomas Fuhrmann

Inhaltsverzeichnis

Abkürzungsverzeichnis

CBT	Core Based Tree
DM	Dense Mode
DVMRP	Distance Vector Multicast Routing Protocol
IGMP	Internet Group Management Protocol
IP	Internet Protocol
MOSPF	Multicast Open Shortest Path First
OSPF	Open Shortest Path First
PIM	Protocol Independent Multicast
RFC	Request For Comment
RIP	Routing Information Protocol
SM	Sparse Mode
SPOF	Single Point Of Failure
TTL	Time To Live

Literaturverzeichnis

Ballardie, A.: *RFC 2189: Core Based Trees (CBTs version 2) Multicast Routing - Protocol Specification*, URL ftp://ftp.ietf.org/rfc/rfc2189.txt

Ballardie, A.: *RFC 2201: Core Based Trees (CBTs) Multicast Routing Architecture*, URL ftp://ftp.ietf.org/rfc/rfc2189.txt

Deering, S., Estrin, D., Farrinacci, D., Jacobsen, V., et al.: *Protocol Independent Multicast Version 2 Dense Mode Specification*, URL ftp://ftp.ietf.org/drafts/draft-ietf-pim-v2-dm-03.txt

Estrin, D., Farinacci, D., Helmy, A., Thaler, D., et al.: *RFC 2362: Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification*, URL ftp://ftp.ietf.org/rfc/rfc2362.txt

Fenner, W.: *RFC 2236: Internet Group Management Protocol, Version 2*, URL: ftp://ftp.ietf.org/rfc/rfc2236.txt

Wittmann, R., Zitterbart, M.: *Multicast - Protokolle und Anwendung*, dpunkt Verlag Heidelberg, 1999 ISBN 3-920993-40-3

1 Grundlagen der Gruppenkommunikation

1.1 Gruppenkommunikationsformen

Es wird in verschiedene Kommunikationsformen unterschieden:

- Unicast
- Multicast
- Concast
- Multipeer

Die Unicast-Kommunikationsform entspricht einer Punkt-zu-Punkt-Verbindung (1:1), d. h. es gibt genau einen Sender und einen Empfänger. Das ist die derzeit im Internet häufigste Verbindungsart.

Bei der Multicast-Kommunikation hat man einen Sender und n Empfänger (1:n), es ist also eine Erweiterung der Unicast-Kommunikation.

Concast (auch Concentration genannt) ist das Gegenteil der Unicast-Kommunikation, d. h. man hat m Sender und genau einen Empfänger (m:1).

Multipeer oder Mehrpunkt-Kommunikation entspricht der m:n-Kommunikation, es existieren also m Sender und n Empfänger. Multipeer ist die komplexeste dieser Kommunikationsformen und wird durch mehrfache Multicast-Kommunikationen simuliert.

Neben den vier genannten Kommunikationsformen gibt es noch das Broadcast. Broadcast ist keine Gruppenkommunikation, da hier die Gruppe der Empfänger nicht begrenzt ist. Die Daten werden einfach an alle möglichen Empfänger gesendet.

1.2 Gruppenaddressierung und -verwaltung im Internet

Die Gruppenaddressierung erfolgt in einer der folgenden Alternativen:

- Adressierung über Empfängerlisten
- Adressierung über einer spezielle Gruppenadresse

Bei der Verwendung einer Empfängerliste wird eine Liste aller Empfänger mit ihrer Unicastadresse geführt (bei Mailinglisten wird dies praktiziert). Diese Liste muss also bei Bei- oder Austritt aktualisiert werden.

Bei der Verwendung einer speziellen Gruppenadresse sendet der Sender die Daten genau einmal an die Gruppenadresse. Die netzinternen Zwischensysteme sind dafür verantwortlich, dass die Daten richtig versendet werden.

Die Verwaltung einer Gruppe (Bei- und Austritt, Abfrage aller Teilnehmer) erfolgt im Internet per **Internet Group Management Protocol** (IGMP). Die akuelle Versionsnummer ist 2 und ist im Request for Comment (RFC) 2236 spezifiziert. IGMP ist in das Internet Protocol (IP) integriert falls IP Multicast bereitstellen soll.

Grundsätzlich unterscheidet man in Anfragen (General-Membership-Query, Group specific Membership-Query, Leave-Query) und Antworten (Version-1-Membership-Report, Version-2-Membership-Report). Die General-Membership-Query holt Informationen über alle an das Teilnetz angeschlossenen Gruppen ein. Die spezifische Membership-Query fragt ab, ob noch Mitglieder einer speziellen Gruppe in dem angeschlossenen Teilnetz existieren. Die Membership-Reports informieren die Router über den Beitritt zu einer speziellen Gruppe. Zum Verlassen der Gruppe wird eine Leave-Query ausgeführt.

Die IGMP-Dateneinheiten sind in die IP-Dateneinheiten gekapselt, die mit einer Time to Live (TTL) von eins ausgestattet sind. Sie verlassen das lokale Subnetz nicht, sondern kommen nur bis zu dem nächsten Router. Für die weitere Verbreitung sind die Multicastprotokolle verantwortlich. Multicast-Router speichern die Zugehörigkeit zu einer Gruppe ab, nicht aber welche angeschlossene Systeme zu welcher Gruppe gehören.

1.3 Probleme des Multicast-Routings

Das Routing in einer Multicast-Gruppe muss so effizient wie möglich gestaltet werden, d. h. die Netzlast soll minimiert werden und Schleifen und Konzentrationspunkte sollen vermieden werden. Besondere Beachtung verdient die Dynamik der Gruppe. Wenn Teilnehmer die Gruppe verlassen oder der Gruppe beitreten, müssen die Routingalgorithmen den Pfad zwischen dem Sender und den Empfänger anpassen. Diese Veränderungen sollen inkrementell passieren, d. h. wenn ein Teilnehmer seinen Zustand bezüglich der Gruppe ändert, soll es nicht zu einer kompletten Neuberechnung der Routinginformationen kommen. Die Änderung soll vielmehr lokal behandelt werden.

Die Aufgabe der **Multicast-Algorithmen ist der Aufbau eines sog. Multicast-Baums**. Das ist ein Verteilbaum, der die Gruppenmitgliedschaft berücksichtigt und den Pfad vom Sender zu den Empfängern nach Kostengesichtspunkten (z. B. Routing-Metrik) optimiert. Der Algorithmus sollte dabei auch mit möglichst wenig Information in den Routern zurecht kommen. Diese Multicast-Algorithmen werden in drei grundlegende Klassen eingeteilt:

- empfängerbasiertes Routing
- Steiner-Bäume
- Bäume mit Rendezvous-Stellen

Auf empfängerbasiertes Routing und Steiner-Bäume wird hier nicht eingegangen. Core Based Trees (CBTs) und Protol Independent Multicast / Sparse Mode (PIM / SM) gehört zur Klasse der Bäume mit Rendezvous-Stellen, die nun kurz beschrieben werden.

Rendezvous-Stellen sind Router im Netz, die die Zusammensetzung der Gruppe kennen. Sie speichern lediglich die Information, ob unter den nachfolgenden Empfängern ein Gruppenmitglied ist. Ist kein Gruppenmitglied angeschlossen ist der Übertragungsabschnitt nicht aktiv.

Die Auswahl der Rendezvous-Stelle ist ein NP-vollständiges Problem und wird durch einfache Heuristiken geregelt. Ist eine Rendezvous-Stelle ausgewählt, melden sie die Gruppenmitglieder durch Versenden entsprechender Nachrichten an. Die Router, die auf dem Weg zur Rendezvous-Stelle liegen, leiten die Nachricht weiter. Dazu wird lediglich eine Information je Gruppe benötigt. Auf diesem Weg wird ein Spannbaum pro Gruppe erzeugt.

Der Vorteil der Verfahren mit Rendezvous-Stellen ist, dass sie die Verbreitung der Daten der Gruppenmitgliedschaft begrenzen. Der Nachteil ist, dass die Rendezvous-Stellen einen Single-Point-of-Failure (SPOF) darstellen. Dies kann dadurch behoben werden, dass mehrere Rendezvous-Stellen verwendet werden, was allerdings den Verwaltungsaufwand steigert.

2 Protocol Independent Multicast

Protocol Independent Multicast (PIM) ist insofern unabhängig, als es nicht auf die Verwendung eines bestimmten Unicast-Routingprotokolls festgelegt ist. Im Gegensatz dazu sind die im Teleseminar bereits vorgestellten Routingalgorithmen Distance Vector Multicast Routing Protocol (DVMRP) von dem Unicastprotokoll Routing Information Protocol (RIP) und Multicast Open Shortest Path First (MOSPF) von dem Unicastprotokoll Open Shortest Path First (OSPF) abhängig.

PIM wurde mit dem Ziel entworfen die Zustandshaltung in den Routern zu minimieren, den Verarbeitsungsaufwand von Kontroll- und Nutzdaten zu minimieren und die beanspruchte Bandbreite zu minimieren.

2.1 Modi im Protocol Independent Multicast

Für PIM existieren zwei unterschiedliche Modi: der Sparse Mode (SM) und der Dense Mode (DM). Die beiden Modi zielen auf unterschiedliche Anwendungsgebiete ab.

Der **Sparse Mode ist für leicht oder dünn besetzte Netze gedacht, während der Dense Mode für dicht besetzte Netze** gedacht ist. Im DM wird von grundsätzlich anderen Annahmen als im SM ausgegangen. Eine Annahme ist, dass in allen angeschlossenen Teilnetzen Empfänger vorhanden sind. Der Sender flutet also das gesamte Netz mit Dateneinheiten und wird dann durch Pruning auf Teilnetze beschränkt. Schneller Zutritt zur Gruppe enfolgt durch Graft-Dateneinheiten.

PIM / DM unterscheidet sich insofern von DVMRP und MOSPF, als es unabhängig von der für die Topologieerkennung genutzten Verfahren ist. Als Nachteil wird häufig der erhöhte Verkehr durch das Fluten in alle Netzbereiche genannt. Dieser Nachteil wird teilweise dadurch aufgehoben, dass PIM / DM nur dort verwendet werden sollte, wo eine hohe Gruppendichte besteht.

PIM / SM und PIM / DM sind zwei grundlegend unterschiedliche Konzepte bzw. Multicast-Routingprotokolle. DM ist für den Einsatz innhalb einer Domäne gedacht, wohingegen der SM auch größere Bereiche abdecken kann. Die folgenden Erklärungen beziehen sich alle auf den Sparse Mode.

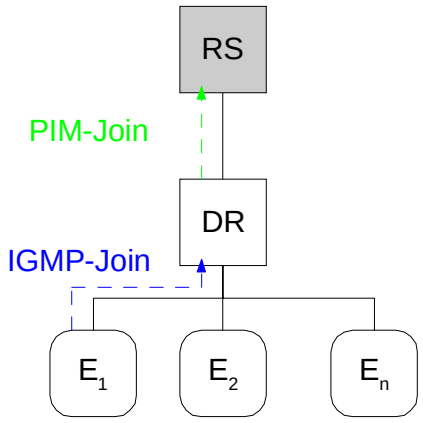
2.2 Der Sparse Mode (SM)

Der SM hat zwei explizite Forderungen:

- Forderung nach explizitem Gruppenbeitritt und
- Bereitstellung von Rendezvous-Stellen

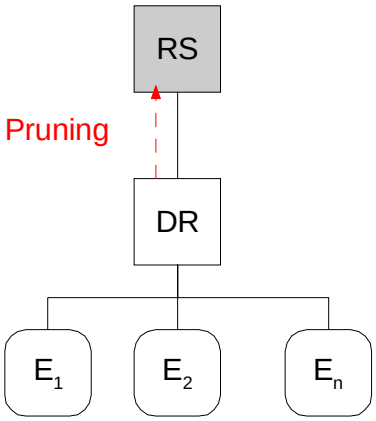
Im SM geht man von dünn besetzten Netzen aus. Die Gruppenmitglieder müssen sich selbst anmelden, weil man im SM am Anfang davon ausgeht, dass keine Gruppenmitglieder existieren. Im Gegensatz dazu geht man ja im DVMRP und MOSPF davon aus, dass sich überall Gruppenmitglieder befinden und den Multicast-Baum dann schrittweise durch Pruning reduziert. Im SM wird der Baum aber durch die Beitrittsmeldungen der Gruppenmitglieder aufgebaut, d. h. Multicast-Dateneinheiten werden nur dorthin versendet, wo auch eine explizite Beitrittsmeldung vorliegt.

Der **Beitritt eines Empfängers zu einer Gruppe** G wir durch eine IGMP-Nachricht veranlasst. Diese Nachricht wird vom Designated-Router (der Designated-Router ist der Router mit der niedrigsten IP-Adresse) des Empfängers an die Rendezvous-Stelle geschickt, die für die Gruppe G zuständig ist. Kann der Designated Router keine Rendezvous-Stelle finden, d. h. existiert die Gruppe nicht, wird kein Multicast-Eintrag erzeugt und die Dateneinheit wird verworfen.



Der Empfänger sendet diese Join-Dateneinheiten periodisch weiter solange er Mitglied in einer Gruppe G sein möchte. Falls nach einer Zeit T vom Router keine Join-Dateneinheiten mehr empfangen werden, schneidet sich der Router, falls keine weiteren Abhänigkeiten existieren, mittels einer Pruning-Dateneinheit vom Multicast-Baum ab.

In den Routern zwischen der Rendezvous-Stelle und dem Empfänger werden **Multicast-Routingeinträge** erzeugt. Hierbei sind drei Situationen zu unterscheiden:



- kein Eintrag zur Gruppe G bekannt
- Eintrag zur Gruppe G mit beliebigem Sender
- Eintrag zur Gruppe G mit speziellem Sender S

Im ersten Fall wird ein Routingeintrag der Form (*, *, RS) erzeugt. Die erste Komponente gibt die den Sender, die zweite die Gruppe und die dritte die Rendezvous-Stelle an. Hier sind also Sender und Gruppe unbekannt, nur die Rendezvous-Stelle ist bekannt und eingetragen.

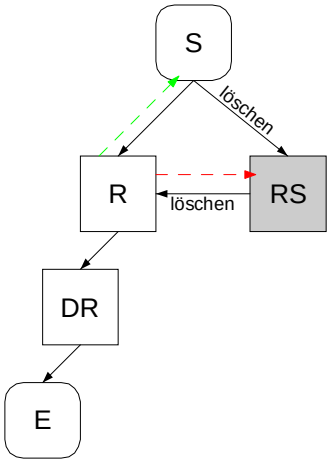
Im zweiten Fall ist die Gruppe bekannt und es wird ein Routingeintrag der Form (*, G) erzeugt.

Im dritten Fall sind sowohl die Gruppe als auch der Sender bekannt. Hier kann ein spezifischer Eintrag der Form (S, G) erzeugt werden.

Der **Sender schickt die Daten per Unicast-Verbindung zu der Rendezvous-Stelle**. Der Designated-Router des Senders kapselt die Multicastdateneinheiten in Unicastdateneinheiten, die an die Rendezvous-Stelle geschickt werden. Dort werden die Multicast-Dateneinheiten entkapselt und per Multicast-Baum an die Empfänger der Gruppe verteilt. Dieser Baum ist aber nicht notwendigerweise optimal.

Die Sender S einer Gruppe G senden zunächst alle auf dem gemeinsamen Baum über die Rendezvous-Stelle, die eine Eintrag der Form (*, G) hat. Ein Sender kann aber auch auf einem spezifischen Baum senden, so dass die Rendezvous-Stelle einen Eintrag der Form (S, G) hat.

Die **Umschaltung von einem gemeinsamen Baum zu einem spezifischen Baum** kann nur von der Rendezvous-Stelle und den Routern der lokalen Gruppe veranlasst werden. Dies wird aber erst ab einer signifikanten Anzahl gesenderter Pakete gemacht, um den Verwaltungsaufwand für unnötige Umschaltungen zu vermeiden. Es wird erst ab einer bestimmten Anzahl auf einen senderspezifischen Baum umgeschaltet, weil dann davon auszugehen ist, dass eine hohe Anzahl weitere Dateneinheiten folgt.



Ein Router aus dem gemeinsamen Baum sendet periodisch Join-Dateneinheiten an den Sender, solange bis er eine Rückmeldung vom Sender erhält. Der Router setzt ein Bit, das anzeigt, dass der Sender ab sofort auf einem spezifischen Baum sendet. Der Router schickt dann eine Pruning-Dateneinheit an die Rendezvous-Stelle, die dann ihren Routingeintrag in einen Eintrag der Form (S, G) abändert. Zusätzlich wird noch ein Bit gesetzt, das anzeigt, dass die Rendezvous-Stelle Teil des gemeinsamen Baums ist, aber nicht des spezifischen Baums. Der Designated-Router des Senders schickt dann auch keine Daten mehr zu der Rendevous-Stelle, sondern nur noch an den Router, der die Umschaltung zu dem spezifischen Baum initiierte. Die übrigen Router ändern ihre Routingeinträge ebenfalls in (S, G) ab. Der Eintrag wird erst dann gelöscht, wenn nach einer Zeit T keine Daten mehr über diesen Router gesendet werden.

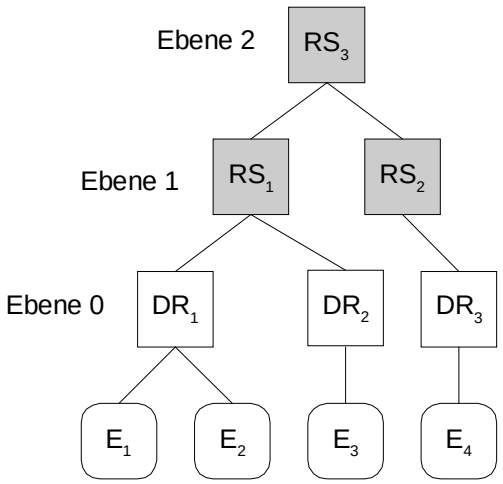
Im PIM / SM werden Dateneinheiten **grundsätzlich periodisch** versendet. Es werden daher keine Mechanismen für Quittungen in dem Protokoll benötigt, was einerseits den Protokollentwurf hinsichtlich der Kontrolle vereinfacht. Andererseits erhöht das aber den Kontrollverkehr im Netz, was sich vor allem in großen Netzen bemerkbar macht.

In den PIM-Domänen existiert noch ein sogenannter **Bootstrap-Router**, der für die Verbreitung von Information über die Rendezvous-Stelle verantwortlich ist. Hierfür werden periodisch dedizierte Bootstrap-Dateneinheiten versendet, die für die auch für die Bestimmung des Bootstrap-Routers verwendet werden. Innerhalb einer Domäne existieren jeweils mehrere Kandidaten für Bootstrap-Router und mehrere Kandidaten für Rendezvous-Stellen. In der Regel sind diese Kandidaten die gleichen Router. Die Information über die Rendezvous-Stelle wird per Unicast verbreitet. Sie kann von den Routern auch dazu verwendet werden, eine neue Rendevous-Stelle zu ermitteln.

2.3 Hierarchical Protocol Independent Multicast (HPIM)

Die Idee des hierarchischen PIM ist die Anordnung der Rendezvous-Stellen in einer Hierarchie. Die Designated-Router der jedes Mitglieds wird als Ebene 0 aufgefaßt, die erste Rendezvous-Stelle dahinter als Ebene 1 usw.

Tritt ein Mitglied einer Gruppe G bei, so wird die Join-Dateneinheit an die Rendezvous-Stelle der ersten Ebene verschickt. Dort wird die Join-Dateneinheit quittiert und an die Rendezvous-Stelle der nächsthöheren Ebene verschickt, bis die für die Gruppe G höchste Ebene erreicht ist.



Beginnt nun ein Sender zu senden, werden Registrier-Dateneinheiten bis zur ersten Rendezvous-Stelle verschickt, der die Gruppe G bekannt. Im schlimmsten Fall werden die Registier-Dateneinheiten bis zur höchsten Ebene geleitet. In der Registier-Dateneinheit wird jeweils vermerkt, welche Rendezvous-Stellen bereits durchlaufen wurden, um Schleifen zu vermeiden. Die nachfolgenden Dateneinheiten werden dann direkt an die Rendevous-Stelle der höchsten Ebene verschickt, um doppelte Wege zu vermeiden.

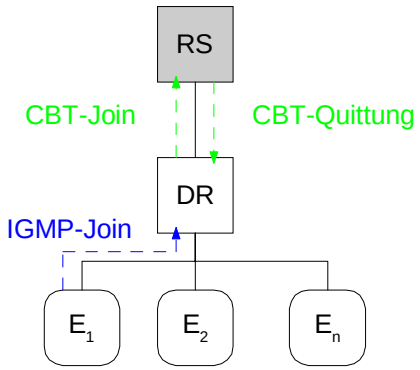
Ein weiteres Mittel um Schleifen zu vermeiden ist, dass die bisher durchlaufenen Router sich merken, für welche Gruppe an welche Rendezvous-Stelle die Dateneinheiten weiterzuleiten sind. Tritt eine Schleife auf und Router hat zuerst an eine Rendezvous-Stelle i die Dateneinheiten gesendet und soll jetzt die Dateneinheiten an eine Rendezvous-Stelle j schicken, die in der Hierarchie weiter oben liegt, wird der Router seinen Routingeintrag so abändern, dass fortan nur noch zur Rendezvous-Stelle j Dateneinheiten versendet werden, weil Rendezvous-Stelle j weiter oben in der Hierarchie liegt.

3 Core Based Trees (CBT)

Core Based Trees sind ebenso wie PIM / SM ein Multicast-Algorithmus von gemeinsamen Bäumen mit Rendezvous-Stellen (hier als Kerne bezeichnet). Die wesentlichen Unterschiede zu PIM / SM sind, dass CBT einen bidirektionalen Baum aufbaut und die Verwendung von Kontroll-Dateneinheiten

Der im PIM / SM erstellte Baum operiert nur in einer Richtung: vom Sender zum Empfänger. Im Gegensatz dazu kann in CBT der Baum auch in Richtung Empfänger zum Sender durchwandert werden.

Die **Ziele beim Entwurf von CBT** waren die Minimierung der Zustandsinformation und Senkung des Overheads durch Kontroll-Dateneinheiten. Das erste Ziel wird dadurch erreicht, dass der Multicast-Baum auch zur Verbreitung der Zustandsinformation genutzt wird. Der Nachteil ist allerdings die Verkehrskonzentration auf dem Multicast-Baum. Das zweite Ziel wird dadurch erreicht, dass jede Kontroll-Dateneinheit explizit quittiert wird. Das erhöht allerdings die Komplexität des Protokolls.



Der **Beitritt zu einer Gruppe** G erfolgt ebenfalls wie im PIM / SM durch einen expliziten Gruppenbeitritt. Konkret wird eine IGMP-Join-Dateneinheit an den Designated-Router versendet. Dieser leitet die Nachricht weiter an den Core. Der Core oder ein Router auf dem Weg dorthin, der bereits Mitglied der Gruppe G ist, muss jetzt, im Gegensatz zu PIM / SM, die Nachricht quittieren. Auf dem Weg der Nachricht vom Empfänger zum Core werden in den durchwanderten Routern ein transienter Multicast-Eintrag erzeugt. Auf dem Rückweg der Quittung vom Core oder Router, der bereits Mitglied der Gruppe ist, zum Empfänger wird der Multicast-Eintrag permanent gemacht.

Der **Austritt aus einer Gruppe** G erfolgt auch explizit, d. h., dass der Empfänger eine IGMP-Leave-Dateneinheit an seinen Designated-Router sendert. Der Router untersucht, ob weitere Abhängigkeiten existieren, falls ja, wird die Leave-Dateneinheit verworfen, ansonsten wird sie zum nächsten Router geschickt, um diesen Teil aus dem Multicast-Baum zu entfernen.

In CBT existiert ebensfalls ein **Bootstrap-Mechanismus** zur Auswahl des Cores. Dieser Mechanismus arbeitet wie PIM / SM mit dem periodischen Versenden von Bootstrap-Nachrichten. Es existiert zusätzlich noch ein Verfahren der manuellen Konfiguration.

4 Bewertung von PIM / SM und CBT

Positiv zu bewerten ist, dass im PIM / SM kein Fluten mit Dateneinheiten stattfindet. Es findet auch ein Fluten statt, aber nur mit Kontrolleinheiten, die periodisch über das Netz verschickt werden, um die Rendezvous-Stelle bekannt zu machen. Kontrolleinheiten sind zwar erheblich kleiner und von der Anzahl her geringer als die Dateneinheiten der empfängerbasierten Multicast-Algorithmen, aber es findet Fluten statt, das sich bei großen Netzen bemerkbar macht.

Das generelle Problem von Bäumen mit Rendezvous-Stellen ist der Single-Point-of-Failure, die Rendevous-Stelle selbst. Wenn die Rendezvous-Stelle ausfällt, muss sie so schnell wie möglich wieder in Betrieb genommen oder ersetzt werden. Dieses Problem wird dadurch in den Griff bekommen, dass stets mehrere Kandidaten für die Rendezvous-Stelle bereit stehen.

Abschließend ist noch zu sagen, dass derzeit der PIM / SM das am häufigsten verwendete Multicastprotokoll im Internet ist.