Universität Karlsruhe
Fakultät für Informatik
Institut für Telematik
76128 Karlsruhe

Netzwerk-Management und Hochgeschwindigkeits-Kommunikation

Teil XVIII

Seminar SS 1998

Herausgeber:
Roland Bless
Stefan Dresler
Günter Schäfer
Claudia Schmidt
Hajo Wiltfang

Universität Karlsruhe
Institut für Telematik

Interner Bericht 17/98
ISSN 1432–7864

Zusammenfassung

Der vorliegende Interne Bericht enthält die Beiträge zum Seminar „Netzwerk-Management und Hochgeschwindigkeits-Kommunikation“, das im Sommersemester 1998 zum achtzehnten Mal stattgefunden hat.

Die Themenauswahl kann grob in folgende 3 Blöcke gegliedert werden:

1. Der erste Block ist Fragen der effizienten Kommunikation mittels ATM (Asynchronous Transfer Mode) gewidmet. Dabei stehen Vorschläge für eine Verbesserung der Handhabung von Gruppenkommunikation sowie Verfahren für sogenannte parallele Pfade im Vordergrund. Ein weiterer Beitrag in diesem Block erläutert ein integriertes Modell zur benutzergerechten Unterstützung der Dienstgüte in ATM-Netzen. Weiterhin wird eine Technik vorgestellt, um IP-basierte Kommunikation mittels ATM-Switching effizienter zu gestalten.

2. Ein zweiter Block behandelt Verfahren zur Unterstützung der Dienstgüte im IP-basierten Internet. Ein Ansatz beschäftigt sich mit der Problematik, dienstgüteunterstützende Mechanismen auch im LAN-Bereich zur Verfügung zu stellen. Ein weiterer Ansatz definiert Konzepte, um eine möglichst einfach und schnell zu realisierende Unterstützung von Diensten mit unterschiedlicher Charakteristiken zu erreichen.

3. Der dritte Block umfaßt den Themenbereich Sicherheit im Internet. Es werden mehrere Protokolle vorgestellt und verglichen, die zur Schlüsselverwaltung in der IP-Sicherheitsarchitektur dienen.

Abstract

This Technical Report includes student papers produced within small lessons called seminar of “Network Management and High Speed Communications”. For the eightteenth time this seminar has attracted a large number of diligent students, proving the broad interest in topics of network management and high speed communications.

The topics of this report may be divided into three blocks:

1. The first block is devoted to ATM technology. Two contributions cover aspects of group communication and so-called multi-path schemes, resp. Furthermore, an integrated architecture for user-oriented support of quality of service in ATM networks and a method for speeding up IP-based communications by using ATM switching techniques are presented.

2. A second block deals with support for quality of service in the IP-based Internet. One contribution describes an approach for providing quality of service for integrated services in local area networks. Another article presents current efforts to provide simple mechanisms and fast deployment of services with differentiated characteristics in the Internet.

3. The third block is devoted to security aspects in the Internet. Some protocols for key management in the IP-security framework are presented and compared.

Inhaltsverzeichnis

Vorwort

Das Seminar „Netzwerk-Management und Hochgeschwindigkeits-Kommunikation“ erfreute sich auch in diesem Semester wieder großer Beliebtheit. Gerade heutzutage sind Stichworte wie „ATM“, „Quality of Service“, oder „Internet“ in aller Munde. Daher sind die Forschungsgebiete in diesen Bereichen auch von allgemeinem Interesse, so daß sie eine derartige Vielzahl von innovativen Arbeiten aufweisen können, deren Behandlung in anderen Lehrveranstaltungen so detailliert nicht möglich ist.

Jetzt liegt auch der nunmehr achtzehnte Seminarband als Interner Bericht vor. Durch die engagierte Mitarbeit der beteiligten Studenten konnte so zumindest ein Ausschnitt aus dem komplexen und umfassenden Themengebiet klar und übersichtlich präsentiert werden. Für den Fleiß und das Engagement der Seminaristen sei daher an dieser Stelle recht herzlich gedankt.

Die ausgesprochen gute Resonanz bei den Studenten hat uns veranlaßt, auch im Wintersemester 1998/99 ein derartiges Seminar – natürlich mit geändertem aktuellem Inhalt – durchzuführen, so daß bald ein weiterer Interner Bericht mit neuen Forschungsergebnissen aus innovativen Seminarbeiträgen erscheinen wird. Doch vorerst sollen im vorliegenden Band folgende Themengebiete vorgestellt werden:

ATM-Mehrpunkt-zu-Mehrpunkt-Kommunikation mit dem UNI 4.0 und SEAM

Bei der sogenannten Mehrpunkt-zu-Mehrpunkt-Kommunikation (auch M:N- oder Multipeer-Kommunikation) können mehrere Sender einer Gruppe an mehrere Empfänger dieser Gruppe senden. Im Zusammenhang mit ATM ist dabei u.a. Sorge zu tragen, daß zum einen festgelegt werden ist, wie ein Endsystem in eine Gruppe eintreten bzw. aus ihr austreten kann, und zum anderen, wie sichergestellt werden kann, daß in ATM-Zellen segmentierte Datenpakete unterschiedlicher Sender bei den Empfängern wieder korrekt zusammengesetzt und ggf. dem Sender zugeordnet werden können.

In der Ausarbeitung werden zuerst generelle Aspekte der Gruppenkommunikation sowie der speziell im Zusammenhang mit ATM auftretenden Probleme dargelegt werden. Es folgt eine Beschreibung der vom User-Network Interface (UNI) 4.0 vorgesehenen Mechanismen zur Realisierung von Mehrpunkt-zu-Mehrpunkt-Kommunikation, etwas des Leaf Initiated Joins (LIJ). Anschließend werden weitergehende Aspekte der Skalierbarkeit und Effizienz dieser Kommunikationsform anhand des SEAM-Ansatzes vorgestellt.

Parallele Kommunikationspfade in ATM-Netzen

Ein Verfahren zur Vermeidung von Engpässen bei der Übertragung von Daten liegt in der Verwendung mehrerer Kommunikationsverbindungen. Das sogenannte „Striping“ dient dabei nicht nur zu einer möglichen Erhöhung des Durchsatzes (wie es durch Kanalbündelung erreicht werden kann), sondern kann auch der Lastverteilung sowie weiteren Zielgrößen dienen.

Die Ausarbeitung zu diesem Thema stellt zuerst das Konzept des Striping vor, verbunden mit den erhofften Vorteilen und möglichen Realisierungsformen. Es folgen Aspekte

der Lastverteilung und der notwendigen Synchronisationsmaßnahmen, um eine reihenfolgetreue Auslieferung sicherstellen. Anschließend werden die mit den Protokollen erzielten Ergebnisse dargelegt.

Ein Ansatz zur benutzergerechten Nutzung der Dienstgüte in ATM-Netzen

Am Center for Telecommunications Research (CTR) der Columbia Universität in New York wurde ein objektorientierter Ansatz für das Management von ATM-Netzwerken entwickelt. Dieser Ansatz ist durch eine sehr umfangreiche Modellierung gekennzeichnet. Das Integrierte Referenzmodell (IRM) betrachtet dabei nur Breitbandnetzwerke. Eine Erweiterung auf Multimedia-fähige Endsysteme liefert dann das Erweiterte Referenzmodell (XRM), das im Detail noch in die Teilmodelle R-, G- und B-Modell aufgegliedert wird. Schließlich sorgt das Binding-Modell für eine objektorientierte Abstraktion, die ein einfaches und flexibles Erstellen von Diensten zur Verwaltung dieser Breitbandnetzwerke ermöglichen soll. Der vorliegende Beitrag liefert einen Überblick über die Vielzahl der am CTR definierten Modelle und arbeitet vor allem den Bezug zwischen diesen Modellen heraus. Als praktischer Aspekt zur Verdeutlichung und Anwendung der vorgestellten Modelle wird hierbei das Management der Dienstgüte betrachtet.

IP-Switching anstelle von Routing

Die zunehmende Größe und Bedeutung rechnergestützter Kommunikation stellt ständig steigende Anforderungen an die Netzwerke und die Komponenten zur Kopplung verschiedenartiger Netzwerke. Neben der klassischen Lösung basierend auf Brücken und Routern sind vor allem die leistungsfähigen Switches in den letzten Jahren verstärkt in Netzwerken eingesetzt worden. In Kombination mit der ATM-Technologie als schneller und flexibler Übertragungstechnik benötigen die heutigen Lösungen jedoch in größeren Szenarien immer Router, um IP-basierte Netzwerke miteinander zu verbinden. Eine vielversprechende Alternative hierzu stellt das IP-Switching dar, das vom Hersteller Ipsilon entwickelt worden ist. Dieser Beitrag stellt diese Lösung im Detail vor und vergleicht sie mit dem klassischen Routing in bezug auf ATM-Netzwerke.

Vergleich aktueller Key-Management-Protokolle für die IP-Sicherheitsarchitektur

Die Internet Engineering Task Force standardisiert derzeit mit *IPSec* eine Sicherheitsarchitektur für das Internet Protokoll, die für die Aushandlung von Sitzungsschlüsseln ein sogenanntes *Key-Management-Protocol* vorsieht. In dem Beitrag *Vergleich aktueller Key-Management-Protokolle für die IP-Sicherheitsarchitektur* werden drei für diese Aufgabe entwickelte Protokolle vorgestellt und miteinander verglichen.

Unterstützung integrierter Dienste im LAN-Bereich

Die Arbeitsgruppe „Integrated Services“ der IETF (Internet Engineering Task Force) hat ein Modell zur Bereitstellung von integrierten Diensten im Internet entwickelt. Im wesentlichen wurden ein Rahmenwerk, zwei neue Dienste („Controlled Load“ und „Guaranteed Service“) sowie ein Signalisierungsprotokoll RSVP zur Ressourcenreservierung definiert. Da Endsysteme im Internet häufig an lokale Netze (LAN – Local Area Network) angeschlossen sind, bilden LANs oft den ersten bzw. letzten Abschnitt auf dem Weg der Daten durch das Internet. Traditionell bieten LANs jedoch nur eine geringe Unterstützung für Dienstqualitäten, da ein entsprechendes Ressourcenmanagement fehlt, wie z.B. beim weit verbreiteten Zugriffsverfahren CSMA/CD des Ethernets.

Die IETF-Arbeitgruppe „ISSLL“ (Integrated Services over Specific Link Layers) hat es sich unter anderem zur Aufgabe gemacht, Mechanismen zur Unterstützung qualitätsbasierter Dienste in LAN-Technologien zu integrieren. Dazu werden zum einen Prioritätenmechanismen eingesetzt, welche durch die IEEE-802-Arbeitsgruppe für einige LAN-Typen bereits definiert wurden, zum anderen werden eine Architektur und ein Protokoll zur Signalisierung von Ressourcenanforderungen definiert (analog zum Ressourcenreservierungsprotokoll RSVP, das auf IP-Ebene arbeitet). In diesem Beitrag werden die durch die ISSLL-Arbeitsgruppe erarbeiteten Vorschläge zur Unterstützung integrierter Dienste in LANs vorgestellt.

Differentiated Services – oder wie das Internet schnell mit Dienstklassen ausgerüstet wird!

Das Internet der nächsten Generation soll für eine Vielzahl von Anwendungen mit unterschiedlichen Anforderungen an die Qualität der Kommunikationsdienste attraktiv sein. Die Realisierung von mehreren Dienstklassen, sogenannten Differentiated Services, soll dabei nicht durch ein komplexes Management erfolgen, sondern über eine Reihe einfacher Mechanismen innerhalb des Netzwerkes. Dazu bildete sich im Herbst 1997 die neue Arbeitsgruppe „Differentiated Services“ der IETF, welche eine generelle Architektur für die Differentiated Services entwerfen möchte. Weitere Punkte, die von der Arbeitsgruppe adressiert werden, sind die Modifikation der Paketformate von IPv4 und IPv6 sowie das Verhalten der Netzwerkknoten. Wichtig ist dabei, daß innerhalb des Netzwerkes nur die Dienstklassen bekannt sind und damit keine Zusatendinformationen pro Datenstrom von den Netzknoten verwaltet werden müssen.

ATM-Mehrpunkt-zu-Mehrpunkt-Kommunikation mit dem UNI 4.0 und SEAM

Sven Tropf

Kurzfassung

Diese Seminararbeit stellt unterschiedliche Verfahren der Gruppenkommunikation über ATM-Netze dar. Dabei wird zwischen der Punkt-zu-Mehrpunkt-Kommunikation (Multicast) und Mehrpunkt-zu-Mehrpunkt-Kommunikation (Multipeer) unterschieden. Bei erstem wird untersucht, wie ein Knoten einer Gruppe beitreten bzw. aus ihr austreten kann. Bei zweitem wird im wesentlichen auf Probleme im Zusammenhang mit der ATM-Technik eingegangen. Es muß hier Vorsorge getroffen werden, daß in einem Switch Zellen unterschiedlicher Sendeströme nicht auf einer ausgehenden Verbindung vermischt werden, da sonst die einzelnen Pakete nicht wiederhergestellt werden können. Als Ansatz für die Punkt-zu-Mehrpunkt-Kommunikation werden die Mechanismen des User Network Interface (UNI) 4.0 beschrieben, wie etwa der Leaf Initiated Join. Für die Realisierung der Mehrpunkt-zu-Mehrpunkt-Kommunikation wird der SEAM-Ansatz vorgestellt, vor allem auch im Hinblick auf Aspekte wie Skalierbarkeit und Effizienz.

1 Einleitung

Die klassische Kommunikation erfolgt zwischen zwei Stationen über eine direkte Verbindung. Diese Punkt-zu-Punkt-Verbindung wird üblicherweise Unicast genannt. Damit aber mehrere Empfänger und Sender miteinander kommunizieren können, wie es zum Beispiel bei Videokonferenzsystemen der Fall ist, muß diese Form der Kommunikation erweitert werden.

Eine Punkt-zu-Mehrpunkt-Verbindung kann durch mehrere Unicast-Verbindungen zwischen Sender und Empfänger realisiert werden. Dies ist jedoch bei einer großen Anzahl von Empfängern sehr ineffizient. Es hat zum einen den Nachteil, daß eine höhere Netzlast entsteht, da jedes Paket für n Empfängern auch n mal repliziert wird. Zum anderen müssen alle Empfänger dem Sender bekannt sein. Bei einer echten Multicast-Verbindung hingegen können alle Empfänger über eine einheitliche Gruppenadresse adressiert werden. Hier ist es nicht nötig, daß dem Sender alle Gruppenmitglieder bekannt sind. Es können in diesem Fall auch Verfahren vorhanden sein, die das selbständige Beitreten und Verlassen von Empfängern zu einer Gruppe regeln. In ATM wurde dies in der UNI 4.0 Spezifikation des ATM-Forums durch den Leaf Initiated Join beschrieben (siehe Abschnitt 2).

Um nun eine Mehrpunkt-zu-Mehrpunkt-Kommunikation zu realisieren, also eine Gruppe mit mehreren Sendern und Empfängern, kann man einfach mehrere Punkt-zu-Mehrpunkt-Verbindungen, ausgehend von den Sendern, aufbauen. Somit hat man bei n Sendern auch n Bäume, was unter anderem den Aufwand gegenüber einer echten Multipeer-Verbindung erhöht, wenn ein Empfänger einer Gruppe beitritt. Daneben gibt es in ATM den Vorschlag des Multicast-Servers. Hier wird ein Baum vom Multicast-Server zu den Empfängern aufgespannt, wobei die Sender über Unicast-Verbindungen mit dem Server verbunden sind. Bei diesem Verfahren ist der Server ein Flaschenhals, da alle Verbindungen über ihn laufen. Zudem würde bei einem Server-Ausfall die Kommunikation zum Erliegen kommen.

Der Ansatz nach SEAM (Scalable and Efficient ATM Multipoint-to-Multipoint Communication) geht hingegen einen ganz anderen Weg (siehe Abschnitt 3). Er macht sich das Konzept des Core Based Trees zunutze. Hierbei wird pro Gruppe nur noch ein Baum aufgebaut, der sich vom Kern (Core) aus erstreckt. Alle Sender dieser Gruppe senden dann ihre Dateneinheiten über den Core.

2 Punkt-zu-Mehrpunkt-Kommunikation mit ATM

Die Kommunikation über ATM ist üblicherweise eine Punkt-zu-Punkt-Verbindung. Die Multicast-Fähigkeit (Punkt-zu-Mehrpunkt) wird erreicht, indem eine Zelle, die auf einer eingehenden Leitung eines ATM-Switches kommt, repliziert wird und den Switch auf mehreren Leitungen wieder verläßt.

Interessanter ist hier aber die Fragestellung, wie Empfänger zu einer (eventuell) bestehenden Multicast-Verbindung hinzukommen. Eine Möglichkeit ist der Sender-gesteuerte Beitritt. Hierbei wird eine Verbindung von dem Sender zu einem der Empfänger auf die übliche Weise aufgebaut, und anschließend werden weitere Empfänger durch jeweils die Nachricht ADD-PARTY hinzugefügt (siehe Abschnitt 2.1). Eine andere Möglichkeit ist der Empfänger-gesteuerte Beitritt, in UNI 4.0 Leaf Initiated Join genannt (siehe Abschnitt 2.2), der ohne die Mitwirkung des Senders erfolgen kann. Dies hat den Vorteil, daß der Sender der Multicast-Gruppe nichts über die Mitglieder seiner Gruppe wissen muß und den Mitgliederbestand auch nicht ständig nachzuverfolgen braucht. Dies bedeutet die Einsparung von Prozessorleistung und Speicherplatz. Zum anderen kann ein Beitritt zu einer Gruppe, die schon einen Baum aufgesetzt hat, an dem Punkt beendet werden, wo ein neuer Zweig angefügt wird. Es muß somit nicht der ganze Baum bis zum Sender hochgelaufen werden. Dies führt zur Einsparung von Bandbreite und Prozessorleistung in einzelnen Knoten sowie einer geringeren Verzögerung beim Eintritt eines neuen Empfängers. Somit ist ein Empfänger-gesteuerter Beitritt bei einem skalierbaren Multicast-Dienst, wie zu Beispiel Video on Demand, geeigneter.

2.1 Behandlung von Mehrpunktverbindungen

Die Behandlung von Mehrpunktverbindungen wird in der ITU-T Festlegung Q.2971 [ITU-96b] unter der Bezeichnung *Dynamic Add/Drop of Endpoints* definiert. Der Ursprung einer Punkt-zu-Mehrpunkt-Verbindung wird dabei als Wurzel (Root) und die Kommunikationspartner als Blätter (Leafs) bezeichnet. Durch Add/Drop-Prozeduren

können hier Wurzeln neue Blätter hinzufügen oder entfernen. Erst ab UNI 4.0 können auch Blätter einen Beitritt in eine Kommunikationsverbindung initiieren (siehe Abschnitt 2.2). Um nun eine Mehrpunktverbindung zu erzeugen, wird zuerst eine normale Verbindung zu einem der gewünschten Empfänger aufgebaut. Dabei wird für die Signalisierung die ITU-T Empfehlung Q.2931 [ITU-96a] an der Teilnehmer/Netz-Schnittstelle (User Network Interface) verwendet. Der Ablauf entspricht dem in Abbildung 1. Dabei entspricht SETUP einem Verbindungsaufbauwunsch, CONNECT der Annahme der Verbindung, CALL PROCEEDING der Anzeige, daß ein Verbindungsaufbau eingeleitet wurde (lokal), und CONNECT ACKNOWLEDGE der Bestätigung der Verbindungsannahme (lokal).

Für jeden neuen Endpunkt, der in eine existierende Mehrpunktkonfiguration aufgenommen werden soll, wird eine ADD-PARTY-Nachricht gesendet. Der genaue Ablauf wird in Abbildung 4 ab dem 3. Schritt verdeutlicht. Die ADD-PARTY (3)-Nachricht enthält neben der Zieladresse dieselbe *Call Reference* wie die ursprüngliche Verbindung, damit die Mehrpunktverbindung eindeutig gekennzeichnet ist. Es wird allerdings ein neuer virtueller Kanal (VCI) für die Verbindung benötigt. Falls das Netz in der Lage ist, die Verbindung bereitzustellen, wird eine SETUP (4)-Nachricht an die Zielstation gesendet, andernfalls wird die ADD-PARTY-Nachricht vom Netz durch eine ADD-PARTY-REJECT-Nachricht zurückgewiesen (hier nicht dargestellt). Bei einer Verbindungsannahme der Zielstation sendet diese ein CONNECT (6), das als ADD-PARTY-ACKNOWLEDGE (8) an die rufende Station gesendet wird.

Einzelne Verbindungen können von der rufenden Station durch DROP-PARTY-Nachrichten wieder ausgelöst werden, indem die virtuelle Kanalnummer und die Endpunkt-Referenz mitgegeben werden. Bei einer DISCONNECT-Nachricht von einer sendenden Endeinrichtung werden alle Verbindungen zu existierenden Mehrpunktverbindungen ausgelöst.

2.2 Leaf Initiated Join

Der Beitritt eines Empfängers (Blatt) zu einer Multicast-Gruppe wird in der UNI 4.0 Spezifikation als Leaf Initiated Join bezeichnet. Hierbei werden wie anfangs motiviert, zwei verschiedene Arten des Beitritts unterstützt:

1. der *Network LIJ*, bei dem neue Blätter automatisch zu Multicast-Gruppen hinzugefügt werden, nachdem sie den Wunsch dazu geäußert haben. Die Anfrage wird hier, wenn die Punkt-zu-Mehrpunkt-Verbindung schon existiert, nicht bis zur Wurzel weitergeleitet, sondern vom Netz bearbeitet.

2. der *Root LIJ*, bei dem der Sender (Wurzel) neue Knoten durch die Punkt-zu-Mehrpunkt-Prozeduren aus dem vorigen Abschnitt manuell hinzufügt

Beim LIJ unterscheidet man zwischen einem aktiven und einem inaktiven Ruf. Der inaktive Ruf unterscheidet sich vom aktiven dadurch, daß hier das Blatt einen Ruf zu einer Mulicast-Gruppe absetzt, die noch nicht existiert.

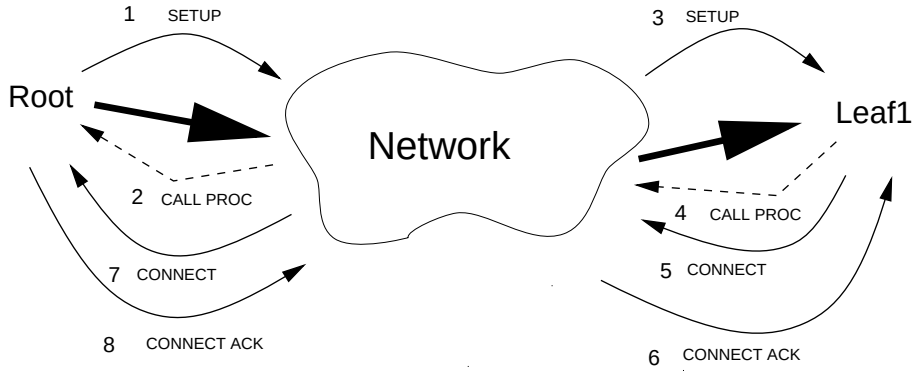

Abbildung 1: Erzeugung eines Network LIJ-Rufs von der Wurzel

2.2.1 Network LIJ

Abbildung 1 zeigt, wie die Wurzel einen Network LIJ-Ruf erzeugt. Die einzelnen Schritte, die hier dargestellt werden, sind zum Großteil identisch mit der Erzeugung eines üblichen Punkt-zu-Mehrpunkt-Rufs. Der wesentliche Unterschied ist, daß die SETUP-Nachricht (1) hier neben der Adresse der gerufenen Station noch zusätzliche Informationselemente enthält. Die *LIJ-Parameter* werden benutzt, um Optionen für den Ruf zu setzen, wie Beispiel Verkehrsparameter oder Quality of Service. Der *LIJ-Rufidentifikator* wird vom Netz benötigt, um die verschiedenen LIJ-Rufe zu unterscheiden, die von einer Wurzel kommen. Innerhalb des Netzes wird der LIJ-Rufidentifikator zusammen mit der Adresse der Wurzel verbunden, um global den LIJ-Ruf von anderen Rufen zu unterscheiden, da es auch möglich wäre, daß mehrere Sender den gleichen LIJ-Rufidentifikator benutzen. Wenn der Rufidentifikator fehlte, entspräche dies der Erzeugung eines Root LIJ-Ruf, bei dem anfragende Blätter nicht ohne die Beteiligung der Wurzel in die Gruppe aufgenommen werden könnten.

Die LIJ-Parameter und der Rufidentifikator können auch in der SETUP-Nachricht (3) vom Netz zum Empfänger gesendet werden, damit dieser genug Informationen hat, um einen Ruf anzunehmen. Die übrigen Schritte sind identisch zu einem üblichen Punkt-zu-Mehrpunkt-Ruf. Nachdem der Ruf abgeschickt wurde, kann die Wurzel weitere Blätter hinzufügen, indem sie die Punkt-zu-Mehrpunkt-Prozeduren aus Q.2971 verwendet. Sie können sich lediglich dadurch unterscheiden, daß die neuen LIJ-Parameter und der LIJ-Rufidentifikator mitgesendet werden.

Abbildung 2 zeigt die Schritte, die folgen, wenn ein Knoten von sich aus einem Network LIJ-Ruf beitritt. Das Blatt erzeugt zuerst eine LEAF SETUP REQUEST (1)-Nachricht, welche die Wurzel-Adresse, den LIJ-Rufidentifikator und eine Blatt-Sequenznummer enthält. Im nächsten Schritt antwortet das Netz mit einer SETUP-Nachricht (2), welche die gleiche Blatt-Sequenznummer beinhaltet. Dieses Echo ermöglicht dem Blatt, die Nachricht vom Netz mit seiner SETUP REQUEST-Nachricht in Verbindung zu bringen. Die nachfolgenden Schritte sind dieselben wie bei einem üblichen Beitritt zu einer Gruppe. Die Wurzel wird hier jedoch nicht vom Beitritt neuer Blätter benachrichtigt. Sie kann also weder bestimmen, wer ihre Übertragungen bekommt, noch kann sie einzelne Knoten, die selbst beigetreten sind, von den Übertragungen ausschließen.

Abbildung 3 zeigt die Schritte, wenn ein neues Blatt zu einem noch nicht existierenden Ruf beitreten will. Wie oben sendet das Blatt eine LEAF SETUP REQUEST (1)-

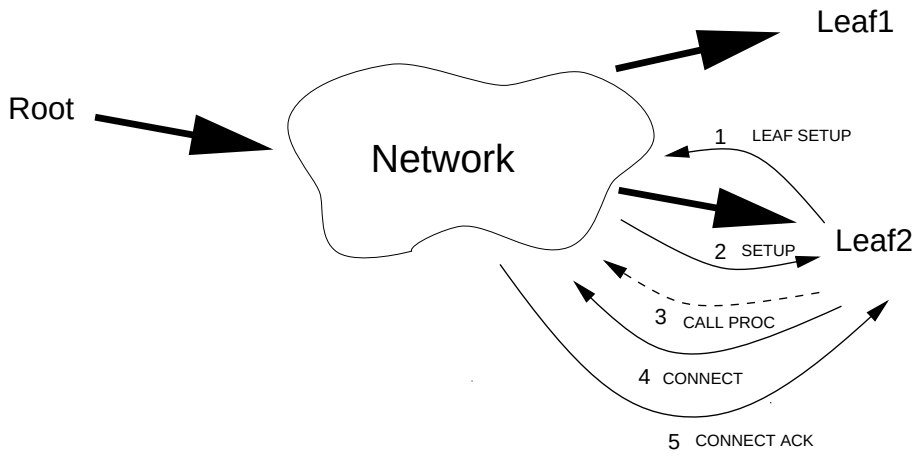

Abbildung 2: Leaf initiated join aufgrund eines Network LIJ-Rufs

Nachricht, und das Netzwerk setzt die Wurzel-Adresse und den LIJ-Rufidentifikator zusammen, um den Ruf zu identifizieren. In diesem Fall, da der Ruf noch nicht existiert, leitet das Netzwerk den Ruf weiter zu der Wurzel. Nachdem die Wurzel die LEAVE SETUP REQUEST (2)-Nachricht erhalten hat, entscheidet sie, ob sie einen neuen Ruf an das Blatt absetzt oder einen Fehlercode zurückschickt mit einer LEAF SETUP FAILURE-Nachricht (hier nicht dargestellt). Der mögliche Rufaufbau folgt dann den existierenden Punkt-zu-Mehrpunkt-Prozeduren. Die SETUP (3)-Nachricht muß die Blatt-Sequenznummer enthalten, die das Blatt in der LEAF SETUP REQUEST-Nachricht benutzt hat. Optional kann die SETUP-Nachricht der Wurzel die LIJ-Parameter und den Rufidentifikator enthalten, damit ein Network LIJ-Ruf erzeugt wird, wie in Abbildung 1 gezeigt wurde.

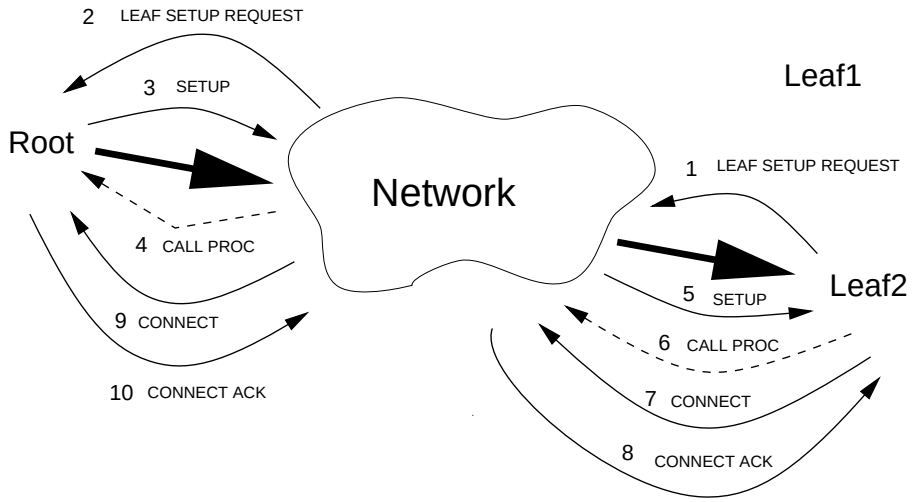

Abbildung 3: Leaf initiated join zu einem inaktiven Ruf mit anschließender Verbindung von der Wurzel zum Blatt

In beiden Fällen (Abbildung 2 und Abbildung 3), in denen ein Blatt zu einer Punkt-zu-Mehrpunkt-Verbindung beitreten will, benötigt es den LIJ-Rufidentifikator. Wie es zu diesem Rufidentifikator kommt, liegt außerhalb der UNI-Spezifikation. Er kann zum Beispiel durch einen Verzeichnisdienst oder durch eine wohlbekannte ID gewonnen werden.

2.2.2 Root LIJ

In Abbildung 4 werden die Interaktionen gezeigt, wenn ein Blatt einem Root LIJ-Ruf beitreten will. Die LEAF SETUP REQUEST (1)-Nachricht des Blattes wird vom Netz bis zur Wurzel weitergeleitet. Die Wurzel schickt daraufhin entweder ein LEAF SETUP FAILURE oder fügt den Blattknoten hinzu, indem sie eine ADD-PARTY (3)-Nachricht verschickt. In diesem Fall finden die Standardprozeduren statt (siehe Abschnitt 2.1), mit der Ausnahme, daß die Sequenznummer aus der LEAF SETUP REQUEST-Nachricht in die ADD-PARTY Nachricht eingefügt wird.

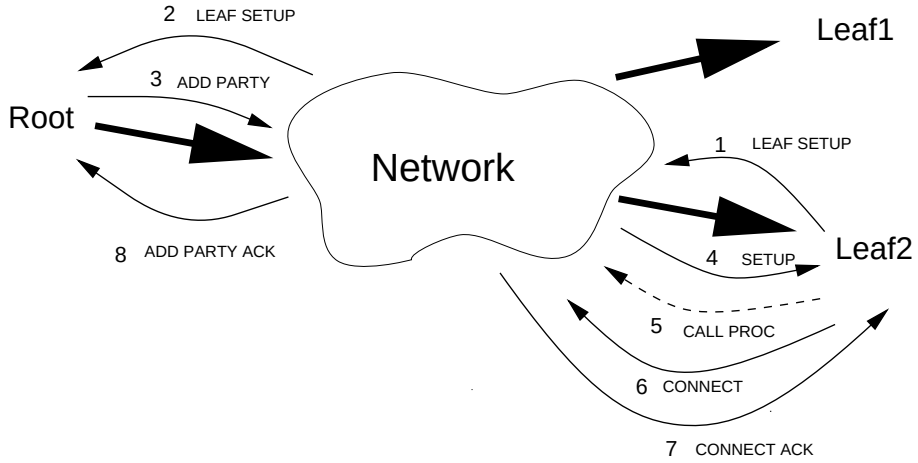

Abbildung 4: Leaf initiated join zu einem aktiven Root LIJ-Ruf mit anschließender Verbindung von der Wurzel zum Blatt

3 Mehrpunkt-zu-Mehrpunkt-Kommunikation mit ATM

Wie eine Punkt-zu-Mehrpunkt-Verbindung in ATM realisiert wird, wurde im letzten Abschnitt dargestellt. Oft werden aber auch Verbindungen benötigt, die aus einer Gruppe mit mehreren Sendern bestehen. Eine naheliegende Erweiterung hin zu dieser Mehrpunkt-zu-Mehrpunkt-Kommunikation wäre es, pro Sender einen Baum zu allen Empfängern aufzubauen. SEAM (*Scalable And Efficient Multipoint-to-Multipoint Communication*) basiert hingegen auf einem einzelnen geteilten Baum zwischen allen Mitgliedern der Gruppe. Dabei ist der *Core* die Wurzel des Baums. Er ist nicht unbedingt ein ATM-Switch, sondern kann auch ein Endgerät sein. Ein einzelner Virtueller Kanal (VC) pro Link wird hier benötigt, um Zellen von allen Sendern der Multicast-Gruppe zu den Empfängern in der Gruppe zu schicken.

Der Vorteil dieses Ansatzes liegt vor allem im Hinblick auf die Effizienz und Skalierbarkeit. In [GrRa97a] werden anhand einer Simulation die unterschiedlichen Kosten des Beitritts eines neuen Knotens zu einer Gruppe geschildert. Dabei schneiden die geteilten Bäume (SEAM) im Gegensatz zu den Sender-Baumgruppen viel besser ab. Dies liegt hauptsächlich daran, daß die Dauer des Beitritts eines neuen Knotens (Sender oder Empfänger oder beides) in SEAM wesentlich kürzer ist. Falls der neue Knoten ein Empfänger ist, muß er allen Sender-Bäumen beitreten. Daher ist es wahrscheinlich,

daß mindestens einer der Beitritte bei einer Baumgruppe mit mehreren Sendern länger dauert als in einem geteilten Baum. Falls andererseits der neue Knoten ein Sender ist, muß im Gegensatz zu SEAM ein komplett neuer Baum aufgebaut werden. Bei Senderbasierten Bäumen wachsen die Kosten somit proportional zur Gruppengröße, während bei SEAM die Kosten mit der Gruppengröße abnehmen, da ein neu beitretender Knoten eine große Gruppe schon früher trifft.

3.1 Signalisierung in SEAM

Wenn ein Knoten einer Multicast-Gruppe beitreten will, sendet er dem Core eine Nachricht. Entsprechend hoch ist die Bedeutung des Core beim Aufbau des Multicast-Baumes. Die Wahl eines geeigneten Core wird oft durch einen Initiator eingeleitet, der verantwortlich ist, den Core zu definieren und seine Existenz zu verbreiten. Die Auswahl des geeigneten Core liegt aber nicht unbedingt in seiner Verantwortung, sondern kann auch als ein Dienst des Netzes angeboten werden. Der Initiator versorgt dabei den nächstgelegenen Switch mit einer Reihe von möglichen oder aktuellen ATM-Adressen und bekommt von ihm die Adresse des Cores zurückgeliefert. Falls die Adresse des Cores aus irgendeinem Grund schon bekannt ist, z.B. von einer früheren Gruppe, dann kann diese Phase wegfallen. Der Initiator erfragt dann die Gruppenadresse (*group handle*) vom Core. Dies ist ein einheitlicher Identifikator, der aus der Adresse des Cores und einem zusätzlichen Identifikator besteht. Dadurch, daß die Adresse des Cores global eindeutig ist, muß nur darauf geachtet werden, daß der zusätzliche Identifikator ebenso eindeutig ist. Dieser kann derselbe sein wie der *LIJ-Rufidentifikator*, der in UNI 4.0 verwendet wurde, um verschiedene LIJ-Rufe der Wurzel unterscheiden zu können (siehe Abschnitt 2.2.1). Der Initiator versorgt seinerseits den Core mit der initialen Liste der Mitglieder, falls vorhanden, und Flags, die sie als Sender oder Empfänger identifizieren. Somit kann der Core die Mitglieder der Gruppe durch einen Core-gesteuerten Beitritt zum Setup-Zeitpunkt hinzufügen. Die Verteilung der Gruppenadresse an die Mitglieder der Gruppe liegt aber in der Verantwortlichkeit des Initiators und ist nicht Teil von SEAM. Sie kann zum Beispiel durch einen Namensdienst erfolgen oder durch ein direktes Kontaktieren der Gruppenmitglieder.

Ein wichtiger Teil von SEAM ist ein Signalisierungskonzept, das als *Short-Cutting* bezeichnet wird. Damit soll vermieden werden, daß alle Übertragungen erst über den Core gehen, bevor sie die Empfänger erreichen. Die Verzögerungszeit der Zellen wird somit kleiner und die Zellen überqueren jeden Link nur einmal. Um dies zu erreichen, werden die Routingtabellen in den Switches derart modifiziert, daß in jedem Switch *Reverse Path Forwarding* (RPF) emuliert wird, ein Routing-Verfahren, auf das auch DVMRP (Distance Vector Multicast Routing Protocol) aufsetzt. Ein Unterschied dazu ist, daß die Zellen nur auf die Ausgänge weitergeleitet werden, die das *Receiver-Downstream* (RD)-Bit gesetzt haben. Dieses Flag zeigt an, daß auf diesem Port noch Empfänger abwärts zu finden sind. Zusätzlich gibt es für jeden Port auch noch ein *Sender-Downstream* (SD)-Bit, ein Flag, das anzeigt, daß Sender der Gruppe sich abwärts befinden. Es wird bei einem Sender-Beitritt auf dem eingehenden Port gesetzt. Zukünftige Zellen, die auf Ports mit gesetztem SD-Bit hereinkommen, werden dann auf alle Ports, die ein RD-Bit gesetzt haben, umgeleitet.

Wie oben erwähnt, sendet ein Knoten dem Core eine Nachricht, wenn er der Gruppe beitreten will. Diese Anfrage wird auf dem kürzesten Weg in Richtung Core weiterge-

leitet, bis sie einen Knoten erreicht, der schon auf dem entsprechenden Gruppen-Baum aufsitzt. Hier entsteht dann ein neuer Ast von dem anfragenden Teilnehmer der Gruppe zu dem erreichten Switch. Diese Prozedur entspricht dem LIJ in UNI 4.0, wird hier aber auch auf sendende Teilnehmer erweitert. Bei einem LEAF SETUP REQUEST wird hier die Gruppenadresse als Informationselement mitgegeben. Wenn ein neuer Empfänger an einem Port ankommt, an dem das RD-Bit nicht gesetzt ist, wird dieses Bit gesetzt, damit zukünftige Pakete an den Empfänger weitergeleitet werden. Danach wandert die Anfrage weiter Richtung Core. Dieser Prozeß setzt sich solange fort, bis ein Switch erreicht wird, an dem das RD-Bit an mindestens einem Port gesetzt ist, da dies bedeutet, daß Pakete an diese Gruppe ohnehin bis zu diesem Switch gelangen.

In SEAM können auch mehrere Gruppenmitglieder in einem Schritt hinzugefügt werden, indem der Beitritt durch den Core initiiert wird. In einem Modell, in dem pro Sender ein Baum existiert, ist dies natürlich nicht möglich. Der Core-gesteuerte Beitritt kann ein enormer Leistungszuwachs für Anwendungen bedeuten, die davon abhängen, daß eine zentrale Instanz schnell einen Baum aufbaut. Ein Beispiel hierfür ist zum Beispiel ein Telekonferenzsystem für Gruppen mit einem zentralen Server.

3.2 Cut-Through

Durch die Mehrpunkt-zu-Mehrpunkt-Konfiguration des SEAM-Systems werden oft Pakete mehrerer Sender gleichzeitig an die Gruppe der Empfänger gesendet. Somit kommen in einem Switch mehrere Sendeströme unterschiedlicher virtueller Kanäle an, die auf einen ausgehenden virtuellen Kanal weitergeleitet werden sollen. Wenn dies ohne Vorkehrungen geschieht, so daß die Zellen unterschiedlicher Pakete auf dem ausgehenden Kanal miteinander vermischt werden, resultiert dies in der Zerstörung der einzelnen Pakete. Denn in ATM wird zugrundegelegt, daß die Daten auf einem virtuellen Kanal geordnet sind. Somit ist es üblicherweise nicht nötig, die Zellen derart zu identifizieren, daß sie zu einem bestimmten Paket gehören. Dies hat zur Folge, daß die einzelnen Pakete auf dem ausgehenden virtuellen Kanal nicht mehr bei den Empfängern unterschieden werden können.

Um dieses Problem zu umgehen nutzt man das EOP (*End of Packet*)-Zeichen von AAL 5 (ATM Adaptation Layer), das Teil des ATM-Zellkopfes ist. Genauer gesagt ist in diesem Fall das erste Bit im Feld PTI (Payload Type Indication) auf 1 gesetzt. Es gibt an, daß alle vorhergehenden Zellen, die auf diesem virtuellen Kanal empfangen worden sind, zu demselben Paket gehören. Wenn nun mehrere Sender Pakete auf den gleichen virtuellen Kanal schicken, müssen diese in ihrer Gesamtheit und geordnet weitergeleitet werden, um die Unversehrtheit der Pakete zu garantieren. Dies wird erreicht, indem man eine Funktion realisiert, die *Cut-Through* genannt wird. Die Switches leiten die Zellen eines komplettes Paket weiter, während sie die anderen Zellen solange puffern, bis eine EOP-Zelle des weitergeleiteten Paketes ankommt. Wenn die EOP-Zelle durchgeschaltet wurde, kann eine anderer Eingangsport seine Zellen weiterleiten. Dieser Port wird nach dem Round-Robin Verfahren ausgesucht, einem prioritätengestützten Mechanismus. In Abbildung 5 wird das zugrundeliegende Prinzip verdeutlicht. Die Zellen des Pakets Y werden durch den Switch durchgeschaltet, während Paket X solange gepuffert wird, bis die Zelle 5Y mit dem EOP für Paket Y durchgelaufen ist. Anschließend können die Zellen von Paket X durch den Switch laufen.

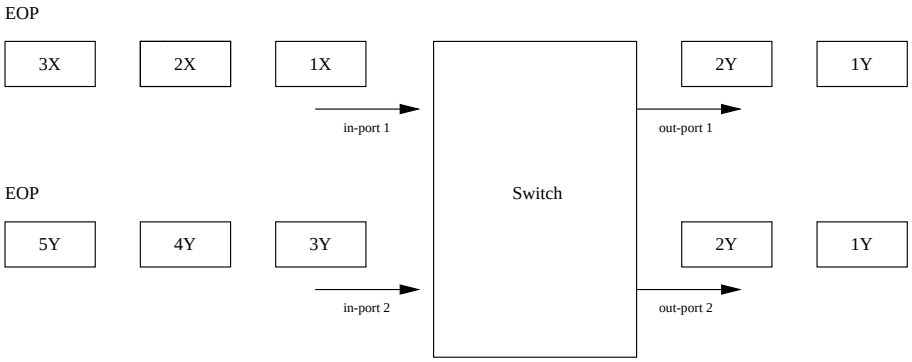

Abbildung 5: Cut-Through

Wie im Unicast-Fall resultiert der Verlust einer EOP-Zelle im Verlust des ganzen Pakets, da es an der Zielstation nicht erfolgreich reassembliert werden kann. Der bedeutende Unterschied hier ist jedoch, daß der Verlust einer EOP-Zelle dazu führt, daß die Zellen an den anderen Eingangsports weiterhin in die Warteschlangen eingereiht werden und keine Chance haben, ihre Zellen gleich zu senden. Erst nach dem Weiterleiten eines weiteren ganzen Pakets auf dem durchgeschalteten Port bekommen diese Zellen ihre Chance. Falls überhaupt keine Zelle oder keine EOP-Zelle mehr ankommt, besteht noch die Möglichkeit, auf den entsprechenden Eingangsport einen Timeout zu setzen.

Bei einem Switch mit verschieden schnellen Eingangsports kann ein langsamer Port zu hohen Verzögerungen für die Zellen führen, die auf den schnellen Ports ankommen. In diesem Fall kann man den Switch so konfigurieren, daß bei den langsamen Ports das Cut-Through ausgeschaltet wird und stattdessen die Zellen auf diesen Ports am Eingang gepuffert werden, bis das ganze Paket eingetroffen ist. Anschließend konkurrieren solche Ports genau wie beim Cut-Through mit den anderen Ports darum, die Zellen des Pakets weiterleiten zu dürfen. Dieses Konzept erlaubt den schnelleren Ports, ein Paket durchzuschicken, während der Switch auf den langsameren Links ein Paket empfängt. Dadurch wird die Nutzung des Ausgabelinks wesentlich erhöht.

Durch die Nutzung des Cut-Through-Verfahrens werden geringfügig größere Pufferspeicher in den Switches benötigt als beim reinen Zellen-Switching. Die unterschiedliche Pufferkapazität hängt dabei von Faktoren wie der Zwischenankunftszeit der Pakete und der Paketgröße ab.

4 Migration von UNI 4.0 zu SEAM

Durch den Cut-Through-Mechanismus benötigt man bei SEAM spezielle Switches. Daher ist es wichtig, für die Interoperabilität zwischen SEAM-Switches und Switches ohne diese Funktionalität zu sorgen. In Abbildung 6 wird eine Architektur für nicht-SEAM Inseln (hier die Switches S2 und S3) mit einer SEAM-Umgebung vorgestellt.

Dabei wird zum einen die Frage aufgeworfen, wie man SEAM-Switches am Rand durch die Insel hindurch verbinden kann, und zum anderen, wie man Sender und Empfänger innerhalb der Insel zu der SEAM-Gruppe verbindet. Die Lösung, die hier entworfen wurde, kann wie folgt zusammengefaßt werden:

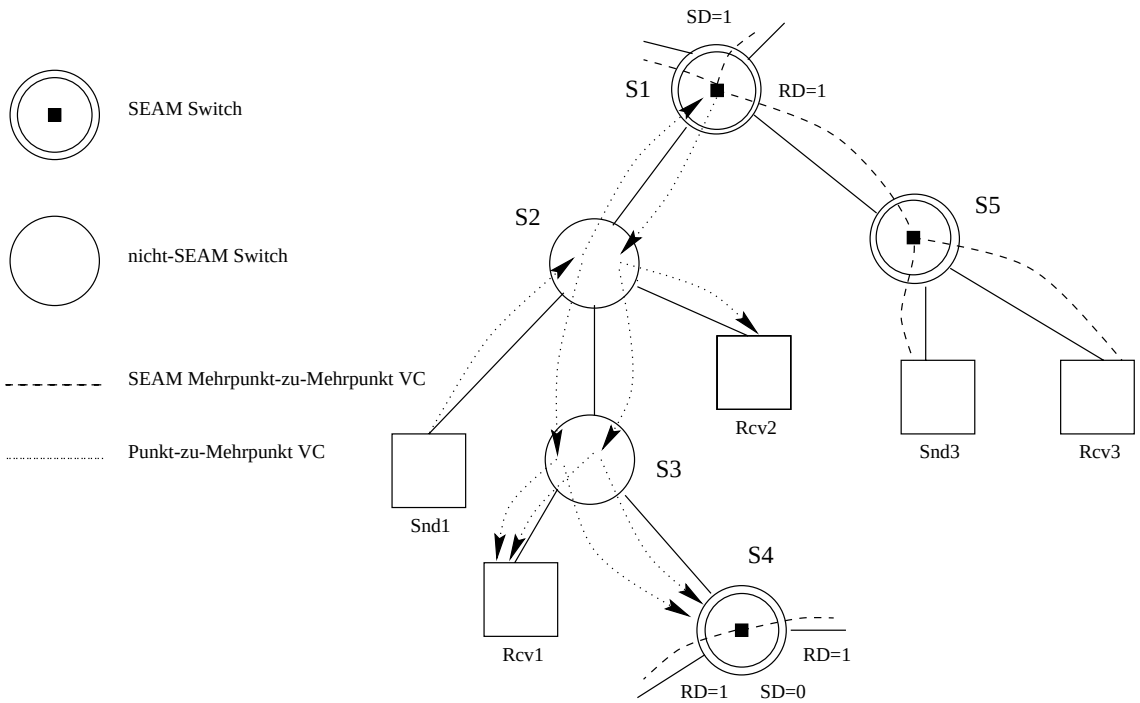

Abbildung 6: Interoperabilität zwischen nicht-SEAM und SEAM Switches

1. Jeder Sender (Snd 1) innerhalb der Insel setzt Punkt-zu-Mehrpunkt-Verbindungen zu den Empfängern innerhalb der Insel (Rcv1, Rcv2) auf, sowie zu den SEAM-Switches am Rand, die das RD-Bit an mindestens einem anderem als dem Port zur Insel gesetzt haben.

2. Jeder SEAM-Switch am Rand, der das SD-Bit an mindestens einem anderen Port als dem Port zur Insel (z.B. S4) gesetzt hat, setzt Punkt-zu-Mehrpunkt-Verbindungen zu allen Empfänger innerhalb der Insel (Rcv1, Rcv2) auf und zu allen SEAM-Switches am Rand, die das RD-Bit an mindestens einem anderem als dem Port zur Insel gesetzt haben (z.B. S4). Wenn das SD-Bit nicht gesetzt ist (wie bei S4), ist keine Punkt-zu-Mehrpunktverbindung in die Insel nötig.

3. SEAM-Switches am Rand bilden alle Punkt-zu-Mehrpunkt-Verbindungen von der Insel in Mehrpunkt-zu-Mehrpunkt-Verbindungen auf den anderen Ports ab.

Diese Lösung setzt auf verschiedenen Überlegungen auf. SEAM-Switches, die nicht zur Insel verbunden sind (S5), brauchen nichts über die Insel zu wissen. SEAM-Switches, die statt dessen direkt am Rand liegen (S1, S4), müssen alle Sender innerhalb der Insel kennen (Snd 1), um einen LIJ zu machen, der bei diesen Sendern seinen Ursprung hat. Dasselbe gilt auch für die Empfänger innerhalb der Insel (Rcv1, Rcv2). Innerhalb der Insel herrscht demgegenüber die Sicht, als ob es keine SEAM Mehrpunkt-zu-Mehrpunkt-Verbindungen gäbe.

Das Schema ist bei wachsender Senderpopulation innerhalb der Insel nicht weiter skalierbar, da die Anzahl der Punkt-zu-Mehrpunkt-Verbindungen hier stark anwächst. Es ist jedoch skalierbar mit der Anzahl der Inseln, da jede Insel nichts von der Existenz der anderen Inseln wissen muß. Ein sinnvoller Ansatz zum Umstieg auf SEAM wäre es, erst diejenigen Switches zu ersetzen, bei denen ein hoher Anstieg der Senderpopulation erwartet wird.

5 Zusammenfassung

Anhand einiger grundlegender Prinzipien wurden in den letzten Abschnitten verschiedene Formen der Gruppenkommunikation in ATM-Netzen gezeigt. Als Erweiterung der Punkt-zu-Mehrpunkt-Kommunikation wurde der LIJ von UNI 4.0 vorgestellt, der den selbständigen Beitritt von Empfängern regelt. Dieser hat den Vorteil, daß der Beitritt ohne Mitwirkung des Senders erfolgen kann und diesen somit wesentlich entlastet. Zudem ermöglicht er die Einsparung von Bandbreite und einen schnelleren Beitritt der Empfänger. Als echte Mehrpunkt-zu-Mehrpunkt-Kommunikation in ATM-Netzen wurde der SEAM-Ansatz dargestellt. Dieser ist auch bei einer großen Anzahl von Sendern und Empfängern noch effizient, da er einen einzigen Baum für alle Mitglieder der Gruppe aufspannt. Der Schlüsselmechanismus Cut-Through wurde wegen des ATM-spezifischen Problems entwickelt, daß Pakete bei mehreren eingehenden virtuellen Kanälen in Switches zerstört werden. Durch die Benutzung des EOP-Bits aus dem Zellenkopf kann hier eine effektive Mehrpunkt-zu-Mehrpunkt-Kommunikation realisiert werden, ohne in den Nutzanteil der Zelle schauen zu müssen. Letztendlich wurde auch gezeigt, daß die Migration zu SEAM möglich ist, indem man Regeln aufstellt, wie SEAM-Switches im Randbereich mit anderen Switches im Inneren operieren müssen.

Literatur

[Foru96] The ATM Forum. ATM User Network Interface Signalling Specification 4.0. Technischer Bericht af-dig-0061.0000, Juli 1996.

[GrRa97a] Matthias Grossglauser und K. K. Ramakrishnan (Hrsg.). *SEAM: An Architecture for Scalable and Efficient ATM Multicast*, Kobe, Japan, April 1997. In Proc. IEEE INFOCOM '97.

[GrRa97b] Matthias Grossglauser und K. K. Ramakrishnan (Hrsg.). *SEAM: An Architecture for Scalable and Efficient ATM Multipoint-to-Multipoint Communication*. 15th International Telegraffic Congress (ITC), Juni 1997.

[ITU-96a] ITU-T. Q.2931 – B-ISDN DSS2 UNI Layer 3 Specification for Point-to-Multipoint Call/Connection Control. Technischer Bericht, 1996.

[ITU-96b] ITU-T. Q.2971 – B-ISDN DSS2 User-Network-Interface (UNI) Layer 3 Specification for Basis Call/Connection Control. Technischer Bericht, 1996.

[Sieg97] Gerd Siegmund. *ATM - Die Technik: Grundlagen, Netze, Schnittstellen, Protokolle*. Hüthig, Heidelberg. 3. Auflage, 1997.

Parallele Kommunikationspfade in ATM-Netzen

Mohamed Moujahed

Kurzfassung

Ein Verfahren zur Vermeidung von Engpässen bei der Übertragung von Daten liegt in der Verwendung mehrerer Kommunikationsverbindungen. Das sogenannte Striping dient dabei nicht nur zu einer möglichen Erhöhung des Durchsatzes (wie es durch Kanalbündelung erreicht werden kann), sondern kann auch der Lastverteilung sowie weiteren Zielgrößen dienen.
Ziel der Ausarbeitung ist es, zuerst das Konzept des Striping vorzustellen, verbunden mit den erhofften Vorteilen und möglichen Realisierungsformen. Darauf aufbauend wird die Leistungsfähigkeit des Verfahrens untersucht.

1 Motivation

ATM (Asynchronous Transfer Mode) ist ein Übertragungsverfahren, welches auf asynchronem Zeitmultiplexing unter der Verwendung von Datenpaketen fester Länge basiert. Diese Datenpakete werden Zellen genannt und haben eine Länge von 53 Byte. Fünf der 53 Bytes sind für den Zellen-Header reserviert, der unter anderem eine Kanal- und Pfad-Adressierung enthält. Alle Netzknoten sind über eine oder mehrere sogenannte ATM-Vermittlungsknoten miteinander verbunden, welche die Zellen an ihren jeweiligen Bestimmungsort vermitteln.

ATM arbeitet verbindungsorientiert. Es werden zwei unterschiedliche Verbindungstypen definiert:

- *Virtueller Kanal*: unidirektionale virtuelle Verbindung.
- *Virtueller Pfad*: Bündel von virtuellen Kanälen mit gleichen Endpunkten (Endgeräte, Vermittlungsknoten).

Für die parallele Kommunikation wird die Information von einer Quelle in m Ströme partitioniert und in k>=m Ströme kodiert. Der Fall k=m repräsentiert eine parallele Kommunikation ohne Kodierung, und der Fall k=m=1 eine konventionelle Kommunikation ohne Aufspaltung der Information der Quelle.
Abbildung 1 zeigt die zwei Fälle: Im ersten Bild ist eine Vermittlungsstelle eines konventionellen ATM-Netzwerks (nicht parallele Pfade) dargestellt, im zweiten eine Vermittlungsstelle, die einen Datenstrom in mehreren Strömen kodiert, was den Fall paralleler Pfade in einem ATM-Netzwerk darstellt.

Wie in Abbildung 2 dargestellt ist, verteilt die erste Vermittlungsstelle, die unmittelbar mit der Quelle (Coder) verbunden ist, die k Ströme in k voneinander unabhängige Pfade. Beim Empfangsgerät wird die urspüngliche Information von dem Dekoder rekonstruiert.

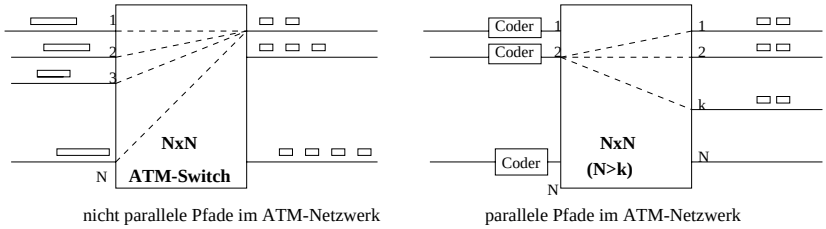

Abbildung 1: Parallele und nicht parallele Pfade

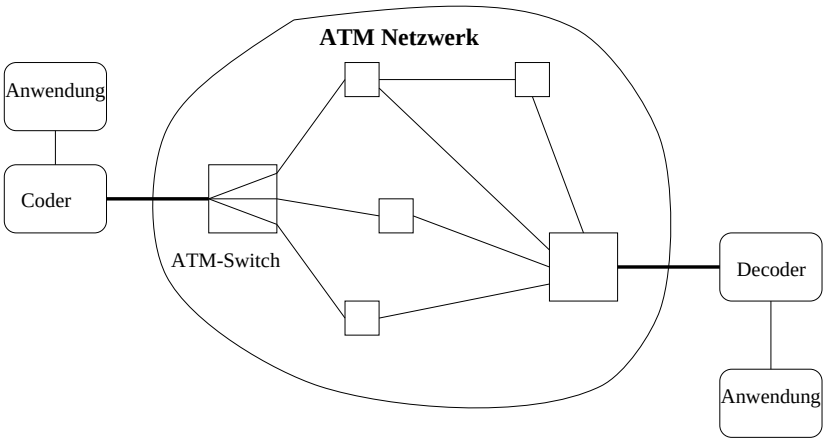

Abbildung 2: Parallele Kommunikation in ATM Netzwerken

2 Striping

2.1 Einleitung

In Hochleistungsnetzen ist die Optimierung einzelner Datenpfade zwischen einer Anwendung auf einem Host und dem Netzwerk typischerweise der effektivste Ansatz, an eine erhöhte Leistung zu kommen.

Als Beispiele für Verfahren, die zu einer Optimierung führen, sind die Minimierung der Daten-Kopien und eine effektive Verwendung der I/O-Ressourcen zu nennen. Eine Optimierung ist normalerweise in einem einzelnen Datenpfad durch die oben genannten Verfahren unmöglich. Eine Zusammenfassung von Datenpfaden (parallele Wege) kann in den beiden Fällen eine Lösung darstellen, um eine weitere Verbesserung in der Gesamtleistung des Netzes zu erzielen.

2.2 Striping und Multiplexing

2.2.1 Striping

Striping ist eine Technik, die für die Anwendung innerhalb von Platten-Subsystemen entwickelt wurde. Sie ist auch eine gute Lösung für die Leistungserhöhung von Netzwerk-Subsystemen durch die Zusammenfassung paralleler Datenpfade.

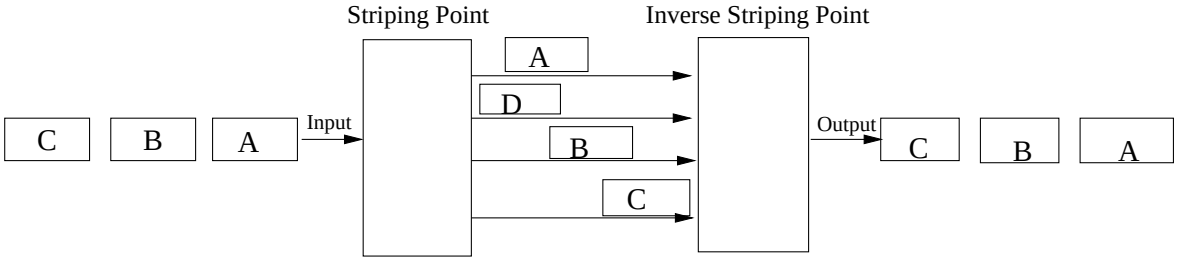

Abbildung 3: Ein gestriptes Kommunikationssystem

Wie Abbildung 3 zeigt, sind in einem gestripten Kommunikationssystem drei Haupkomponenten zu unterscheiden:

- Striping Point: Striping erfordert, daß mehrere physikalische Ressourcen zusammengefaßt werden, es ist also ein Demultiplexer erforderlich. Physikalisch stellt ein Striping Point diesen Demultiplexer dar.
- Inverse Striping Point: Dieses System ist zur Rekonstruktion des Datenstroms, der durch den Striping Point in mehrere Stripes demultiplext wurde, zuständig.
- Stripe: Ein Stripe ist eine Instanz, die einen Ausgang des Striping Point mit einem Eingang des Inversen Striping Point verbindet.

Um die Eigenschaften von Striping gut zu verstehen, ist es sinnvoll, zuerst die übliche Technik des sogenannten Multiplexing zu verstehen.

2.2.2 Multiplexing

Durch das Prinzip des Multiplexing ist es möglich, mehrere voneinander unabhängige Datenströme über ein und dasselbe physikalische Medium zu übertragen. Heute gibt es eine Vielzahl von Multiplexverfahren wie Wellenlängen-, Code-, Frequenz- oder Zeitmultiplexing. In modernen optischen Übertragungsnetzwerken haben das Wellenlängen- und Zeitmultiplexverfahren die größte Bedeutung erlangt.

Zeitmultiplexverfahren (TDM, Time Division Multiplexing) werden in synchrones (Übertragungsrahmen mit einer bestimmten Anzahl von Zeitschlitzen, wobei es eine genaue Zuordnung zwischen Kanälen und Zeitschlitzen gibt; z.B: S0-Rahmen in ISDN) und asynchrones Multiplexing (keine Zuordnung zwischen Kanälen und Zeitschlitzen; z.B Plazierung von ATM-Zellen in einem SONET-Rahmen) unterteilt.

2.2.3 Zusammenhang zwischen Striping und Multiplexing

Striping und Multiplexing scheinen synonym zu sein, was in Wirklichkeit nicht der Fall ist.

Striping ist ein physikalisches Multiplexing, und somit ein Spezialfall des Multiplexing, wobei die Ausführung des Multiplex-Algorithmus für höhere Schichten transparent ist. Die Gewährleistung dieser Transparenz kann schwierig sein, wenn Eigenschaften wie die Folge der Daten eingehalten werden müssen. Multiplexing dagegen garantiert diese Transparenz nicht.

Ein gestriptes System ist synchronisiert, wenn die Reihenfolge der Datenströme eingehalten wird. Andernfalls, also wenn das System unsynchronisiert ist, muß zur Wiedergewinnung der Synchronisation eine Prozedur im Striping Point und im Inverse Striping Point existieren, welche die Reihenfolge der Daten mittels eines Speichers rekonstruieren kann.

2.3 Anwendung von Striping im Netzwerk-Subsystem

Striping, eine Technik die zur Zusammenfassung von Ressourcen führt, kann in unterschiedlichen Situationen innerhalb des Nezwerkes angewandt wird:

1. **Network/Application Bandwidth Mismatch:** Die maximale Bandbreite wird durch Standards und die Infrastruktur festgelegt. Diese erhöht sich in der Regel stufenweise. Die Anwendungen, die an den Endpunkten des Netzwerks laufen, vergrößern ihren Bedarf an Bandbreite jedoch nicht auf diese festgelegte Weise. Netzwerk Striping liefert die Flexibilität, welche notwendig ist, um die Netzwerk-Bandbreite an die Bandbreite, die über das Netzwerk läuft, anzupassen.

2. **Vermeidung von Flaschenhälsen:** Striping kann auch benötigt werden, um hardwarenahe Probleme wie das Flaschenhals-Problem zu lösen. Wenn zum Beispiel in einem Netzwerk-I/O-Subsystem ein einziges Netzwerk-Interface, ein I/O-Bus (siehe Abbildung 4) oder irgendeine Hardware-Komponente in dem System unfähig ist, die gewünschte Bandbreite zu unterstützen, dann kann die Zusammenfassung von mehreren Instanzen dieser Komponenten diese unterstützen.

3. **Fördernde Funktion für neue Netzwerk-Technologien:** Die letzte Feststellung ist, daß Striping ein effektiver Ansatz ist, um technologische Verbesserungen in kommerziell erhältlichen Netzwerken zu motivieren. Ein Schritt nach vorn in angewandter Technologie findet statt, wenn die Technologie ausreichend ausgereift ist, und wenn gezeigt werden kann, daß es Applikationen gibt, welche von dieser Technologie profitieren.

2.4 Synchronisation

Die Techniken, die jetzt beschrieben werden, verwenden spezielle Marker-Pakete, die der Empfänger von normalen Datenpaketen unterscheiden kann. Wir nehmen an, daß — wenn entweder der Sender oder der Empfänger runter- oder hochfährt — der Kanal

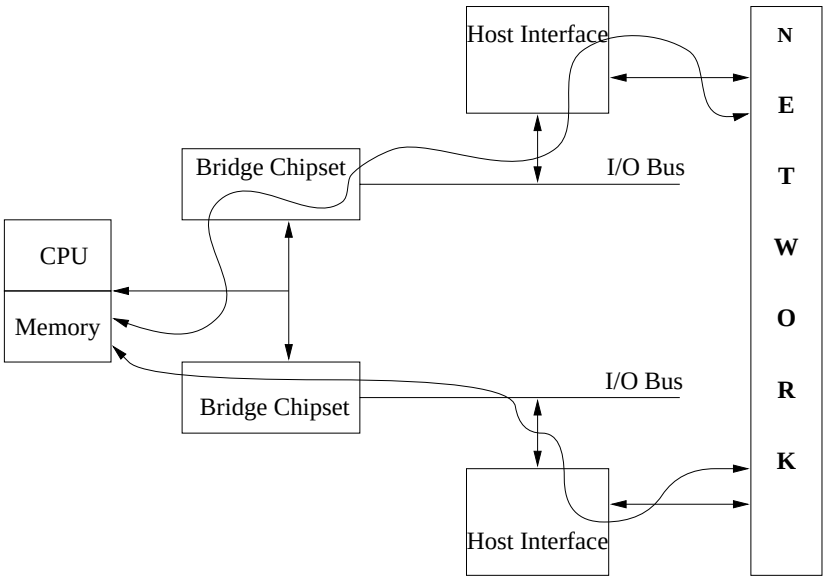

Abbildung 4: Flaschenhals-Vermeidung durch Striping

reinitialisiert und die Synchronisation wiederhergestellt wird [TrSm95].
Das Versenden von Marker-Paketen verlangt keine Modifikation der Datenpakete. Es wird einen (De)Multiplexing-Punkt benötigt, um die Unterscheidung der Marker-Paketen von normalen Paketen zu gewährleisten.
Wir beschreiben jetzt ein Marker-Synchronisationsschema unter Verwendung eines Striping-Algorithmus (z.B SRR, Surplus Round Robin). In Abbildung 5.a) ist die Konfi-

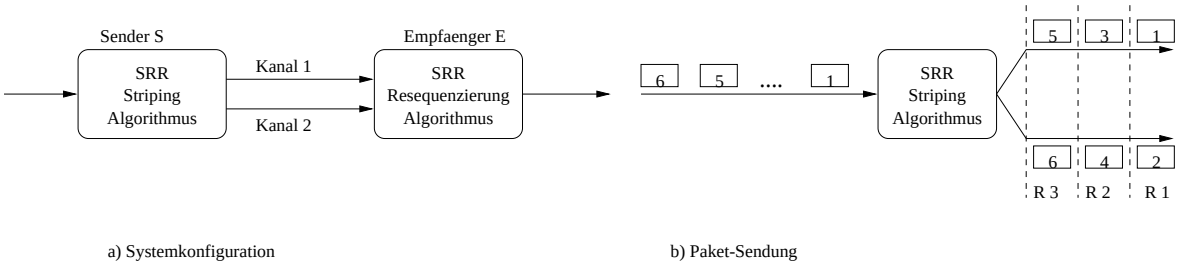

Abbildung 5: Senden von Paketen in einem konfigurierten System

guration eines System mit Striping Point und Inverse Striping Point gezeigt. Im Teil b) kommen die Pakete beim Sender an und werden mittels eines Striping-Algorithmus in zwei Kanäle gestripet. In der ersten Runde (R 1) werden dabei die Pakete 1 und 2 auf die Stripes verteilt, in der Runde 2 die Pakete 3 und 4, etc.

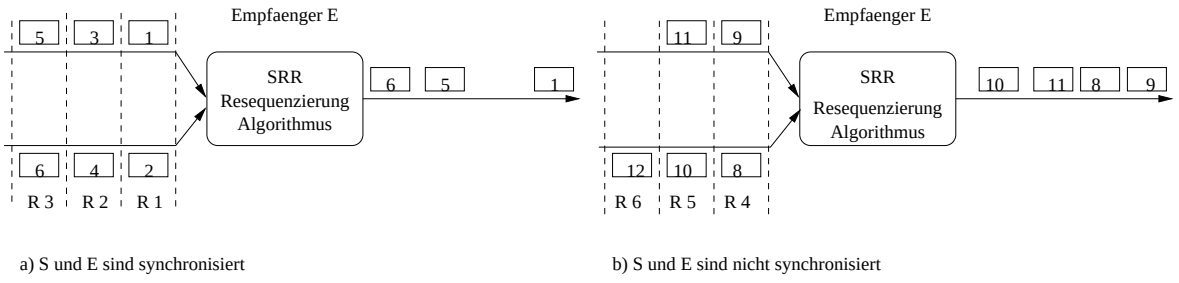

Abbildung 6: Empfang von Daten

Abbildung 6 zeigt den Empfang der Daten in den beiden Zuständen eines Systems, also in den Fällen, in denen Sender und Empfänger synchronisiert oder nicht synchronisiert sind.
Der Teil b) der Abbildung zeigt den Empfang der Daten, wenn das Paket 7 verlorengegangen ist. Der Empfänger erwartet in Runde 4 das Paket 7 auf Kanal 1, aber er nimmt stattdessen Paket 9, da Paket 7 verlorengegangen ist und 9 dann das nächste auf Kanal 1 ist. Dies führt dazu, daß der Sender und der Empfänger nicht mehr synchronisiert sind, und dazu, daß die Pakete nicht in der richtigen Ordnung ankommen.

2.5 Auswertung von Striping innerhalb des Netzwerks

Dieser Abschnitt widmet sich einer systematischen Erforschung der Netzwerk Striping Optionen, die in einem Breitband-Protokollbündel möglich sind.

Diese Optionen werden vergleichend bewertet, um die Charakteristika des Striping auf verschiedenen Ebenen zu zeigen. Diese Charakteristika liefern das Verständnis des Entwicklungskompromisses, der bei der Plazierung des Striping Point in einer Schicht des Netzwerk-Subsystems auftritt.

2.5.1 Kriterien

Für Netzwerkverkehr sind Bandbreite und Latenz die primären Charakteristika des Striping Point. Andere wichtige Kriterien (aus Sicht des Systems) sind die Skalierbarkeit und Komplexität der Implementierung der Kontroll-Algorithmen, der Bedarf nach Puffer und die Fähigkeit, vom Netzwerk induzierten Skew (Versatz) zwischen den Stripes zu tolerieren [AdPV95].

1. *Latenz und Pufferung:* Pufferung und Latenz sind, obwohl sie unterschiedlich scheinen, typische, eng verwandte Charakteristika des Striping Point. Wenn der Verkehr, der an einem Striping Point ankommt, stoßartig (bursty) ist und die zusammengenommene Bandbreite des Striping Point nicht gleich der maximalen Bandbreite des Burst ist, dann ist Pufferung nötig, um den Verlust der Daten zu vermeiden.

   Latenz ist ein Maß für die Zeit, die benötigt wird, um die Daten des Striping Point vom Eingang zu den Ausgängen in Richtung der Stripes zu schicken. Es gibt zwei Komponenten der Latenz: die erste ist die Zeit, die beim Warten in der Eingabewarteschlange verstreicht, die zweite ist die Zeit, die tatsächlich gebraucht wird, um die Daten zu senden. Die zweite Komponente ist von der Striping Technik, die verwendet wird, sowie von der unterstützten Bandbreite der Stripes abhängig.

2. *Tolerieren von Skew*: Für ein gestriptes System, bei dem die Ordnung erhalten werden muß, um Transparenz zu gewährleisten, können Skews innerhalb der Stripes eine Komplikation verursachen. Es gibt zwei Arten von Skews. Statische Skews sind Skews zwischen den Stripes, die für die Dauer der Ausführung des Striping Point festgelegt sind. Dynamisches Skew ist die Komponente des Skew, die sich mit der Zeit verändert.

3. *Skalierbarkeit und Komplexität*: Skalierbarkeit im Sinne des Striping ist ein Charakteristikum, das beschreibt, in welchem Maß es möglich ist, die Anzahl der Stripes, die in einem gestripten System benutzt werden, beliebig zu erweitern. Manche Striping-Algorithmen skalieren gut nur bis zu einer geringen Anzahl von Stripes. Bei Striping-Systemen, die in Hardware implementiert sind, können Verbindungsdichte, Timing und Pufferung die Skalierbarkeit begrenzen, während bei Software-Lösungen die Skalierbarkeit des Systems durch den Rechenbedarf des Striping-Algorithmus limitiert wird.

4. *Bandbreite*: Die maximale Bandbreite, die ein Striping-System unterstützen kann, ist von zwei Faktoren abhängig: der Bandbreite der einzelnen Stripes und der Anzahl der Stripes, die verfügbar sind. Die maximale Bandbreite, die von einem Striping-System unterstützt werden kann, ist gleich der Summe der Bandbreiten der einzelnen Stripes.

2.5.2 Striping auf unterschiedlichen Ebenen

Die Striping-Punkte, die in diesem Abschnitt studiert werden, sind: TCP Striping auf der Anwendungsschicht, IP Paket-Striping auf der TCP-Ebene, ATM-Zellen-Striping auf der ATM Adaptionsschicht und Byte-Striping auf der ATM-Schicht (siehe Abbildung 7).

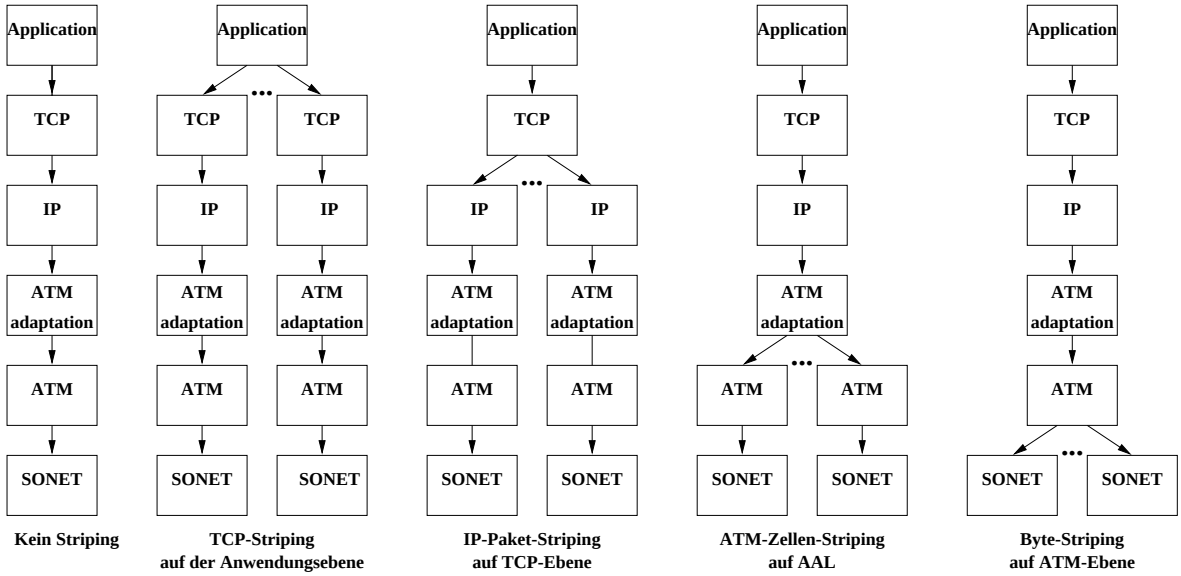

Abbildung 7: Striping auf verschiedenen Schichten

1. *Byte-Striping auf ATM-Schicht*: ATM-Schicht-Striping bedeutet das Striping von ATM-Zellen über mehrere Instanzen der physikalischen Schicht (SONET, SDH) auf der Basis einzelner Bytes. Byte Striping muß die Reihenfolge der Daten beibehalten.
   Am Anfang müssen die ATM-Zellen in den SONET-Payloads plaziert werden. Wenn die Anzahl der Stripes nicht ein Vielfaches von 53 (Anzahl der Bytes innerhalb einer Zelle) ist, wird das erste Byte jeder Zelle über die verfügbaren Stripes rotiert, wenn die Zellen übermittelt werden (siehe Abbildung 5). Diese

Rotation macht Striping Point und Inverse Striping Point komplexer, da man sich merken muß, bei welcher Stripe das erste Byte der nächsten Zelle ankommt. Es gibt zwei Techniken, das erste Byte zu identifizieren:

- Zeiger-Technik: Verwendung eines Zeigers im SONET-Rahmen, der den Anfang der ersten kompletten ATM-Zelle beschreibt.
- Padding-Technik: Padding kann zu jeder Zelle hinzugefügt werden um sicherzustellen, daß alle Zellen auf dem gleichen Stripe beginnen. Diese Technik ist in der Abbildung 7 dargestellt.

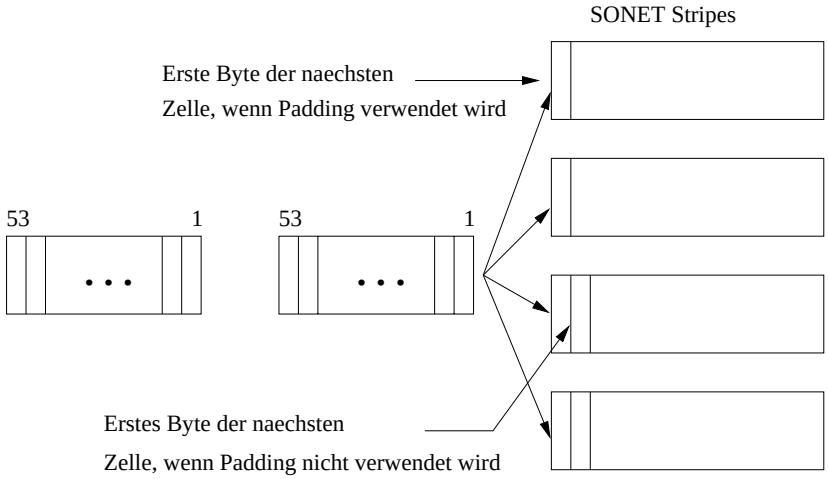

Abbildung 8: Byte-Striping auf ATM-Schicht

Obwohl die Möglichkeiten des Byte-Striping limitiert sind, ist es einfach in Hardware zu implementieren, und es hat ein hohes Potential, gut zu skalieren. Byte-Striping-Implementierungen, welche einen Durchsatz im Bereich von Gigabit pro Sekunde liefern, sind in Software, wegen der extrem hohen Geschwindigkeit, in welcher ein Striping-Algorithmus laufen müßte, unmöglich.

2. *ATM-Zellen Striping auf AAL-Ebene*: ATM-Zellen Striping auf AAL-Ebene benutzt AAL CS-PDUs (Convergence Sublayer Protocol Data Unit) als Eingabe zum Striping Point und gibt ganze Zellen über die Stripes weiter. Da die Dateneinheiten, die über die Stripes transferiert werden, die gleichen wie bei einem nicht gestripten System sind, sind keine Modifikationen an Rahmen der physikalischen Ebene nötig. Die Reihenfolge der ATM-Zellen muß über das gestripte System beibehalten werden um sicherzugehen, daß das Striping auch für Funktionen der höheren Ebene transparent ist.

3. *IP Paket-Striping auf TCP-Schicht*: Striping auf TCP-Ebene bedeutet Striping der TCP-Pakete, die von der TCP-Ebene über mehrere tieferliegende Ebenen verteilt werden. Jedes IP-Paket durchläuft einen anderen Stripe (siehe Abbildung 9). Skew unter den Stripes wird nur durch die Fehlordnung von IP-Paketen verursacht.
   IP-Striping ist für eine Software-Implementierung ideal, da die Dateneinheiten verhältnismäßig groß sind, vor allem im Vergleich zu Bytes oder ATM Zellen. Host-Software ist typischerweise die einzige Möglichkeit der Implementierung, da der Protokoll-Stack ab der AAL aufwärts fast immer auf einem Host in Form von

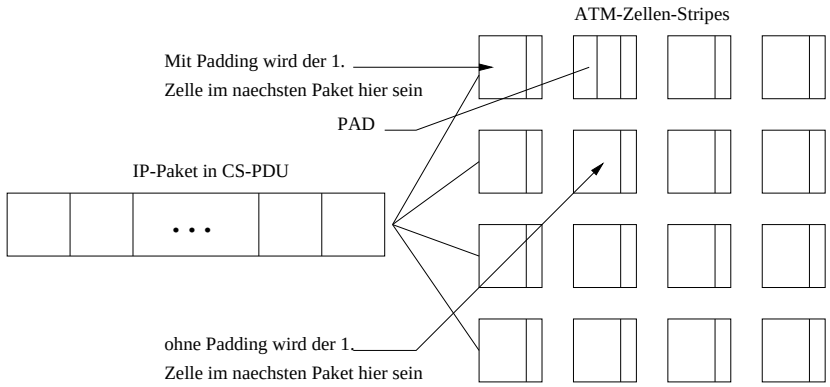

Abbildung 9: ATM-Zellen Striping auf AAL-Ebene

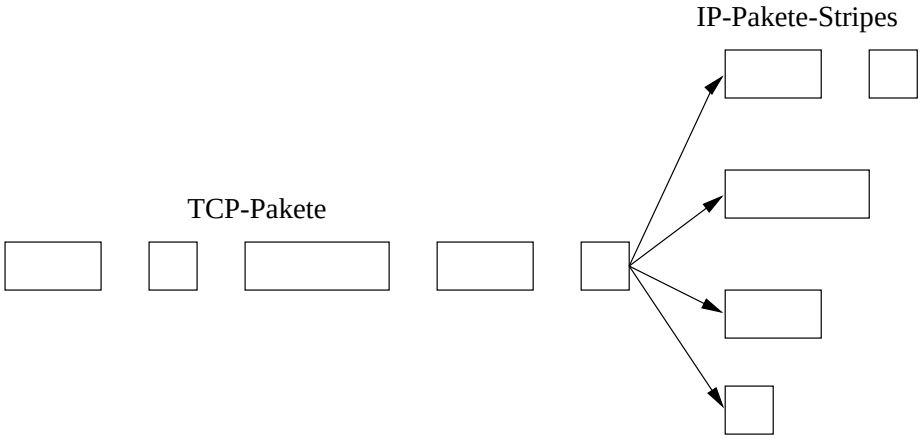

Abbildung 10: IP-Paket-Striping auf TCP Schicht

Software implementiert ist.
Ein Problem, das bei Striping auf niedrigen Ebenen auftreten kann und durch IP Striping verbessert werden kann, ist das Head-of-Line-Blockierungs-Problem. Dieses Problem tritt auf, wenn die Dateneinheiten, die gestripet werden, unterschiedliche Größen haben und wenn die kleinere Dateneinheiten auf die vollständige Übermittlung einer großen Dateneinheit warten müssen. Eine Erhöhung der Anzahl der Stripes kann die Auswirkungen der Head-of-Line-Blockierung reduzieren.

4. *TCP-Striping auf der Anwendungsschicht*: Striping kann auch auf der Anwendungsschicht über mehrere TCP-Stripes verwendet werden. Da TCP über jeden Stripe läuft, bieten die Stripes für die Anwendungsschicht-PDUs einen verläßlichen Transportmechanismus. Daten, die fehlerbehaftet sind (verlorengegangene Daten oder umgeordnete Daten), werden, bevor sie an dem Inversen Striping Point ankommen, korrigiert. Dieser Protokollmechanismus wird trivialerweise in der Transportschicht implementiert, um nicht durch die Vermittlungsschicht behobene Fehler zu behandeln. Das Hauptergebnis von Verlust und Umordnung der Daten wären erhöhte Skews zwischen den Stripes.
Da sich die Dienstgüte (Quality of Service, QoS) auf den einzelnen Stripes stark unterscheiden kann, resultiert ein einfacher RR (Round Robin) Style Striping Algorithmus in einer schwachen Leistung, weil die Gesamtleistung des gestripten Systems durch einen einzelnen schlecht arbeitenden Stripe degradiert wird. Höher-

entwickelte Striping-Algorithmen können die unterschiedliche QoS zwischen den Stripes kompensieren, um die Gesamtleistung zu optimieren.

2.5.3 Zusammenfassung

Dieser Abschnitt hat einen Grundstein für den Vergleich des Stripings auf verschiedenen Ebenen in einem typischen Hochleistungsprotokoll-Modell gelegt. Eine genaue Terminologie wurde für die Beschreibung und Unterscheidung des Stripings von Multiplexing geschaffen.
Die Auswertung des Stripings zeigt verschiedene Erkenntnisse, um die meist günstigste Ebene zu entdecken, um ein Striping Netzwerk für eine gegebene Umgebung und Arbeitsaufgabe zu rüsten.

- Striping auf höheren Schichten führt zu weniger Head-of-Line-Blockierung.
- Übertragungsraten im Bereich von Gigabit pro Sekunde können mit der jetzigen Technologie –auf niedrigen Schichten– nur mit Hardware implementiert werden, da die Granulierung der Striping-Einheiten zu fein ist, um von Software kontrollieren zu werden.
- Striping auf höheren Schichten ist typischerweise in Software implementiert, da es in schon existierende Software-Systeme eingebettet werden muß.

3 Analyse und Simulation der Leistung eines ATM-Netzwerkes

In diesem Abschnitt wird ein Vergleich zwischen parallelen und nicht parallelen Kommunikationssystemen durchgeführt.
Zur Vereinfachung des Vergleichs zwischen parallelem und nicht parallelem Schema nehmen wir an, daß das Switch Loading ausgeglichen ist und daß ein NxN Switch mit der Quelle verbunden ist.
Die Gesamtleistung eines paketvermittelten Netzwerkes hängt stark von der Leistung seiner Übermittlungsabschnitte und von der Paket-Vermittlung ab. Die Bestimmung der Größe der Puffer und der Größe der Daten, die man auf dem Übermittlungsabschnitt senden will für eine spezifische Leistung, sind Hauptpunkte bei der Entwicklung jedes paketvermittelten Netzwerks.
Die Analyse der Wahrscheinlichkeit von Zellenverlust ist kompliziert im Falle einer endlichen Puffer-Größe. Deshalb wurden die Wahrscheinlichkeiten von Zellenverlust bei der parallelen und nicht parallelen Kommunikation durch Simulationen bestimmt.
Die Quelle alterniert zwischen einem on-Zustand, wobei Zellen mit der maximalen Bitrate der Quelle übermittelt werden, und einem off-Zustand, bei dem die Quelle im Stillstand ist. Die Dauer der beiden Zustände ist jeweils exponential verteilt.
Für jede der N Quellen ist die maximale Bitrate 155,52 Mb/s. Man ändern die Wahrscheinlichkeit des on-Zustands und die durchschnittliche Anzahl der Zellen während einer on-Periode, um verschiedene Sätze von experimentell ermittelten Ergebnissen zu erhalten. In Abbildung 11 [DiLi95] sieht man, daß die Verlustwahrscheinlichkeit von paralleler Kommunikation mit k=m (d.h. Aufspaltung ohne Kodierung) viel geringer ist als die von nicht paralleler Kommunikation.

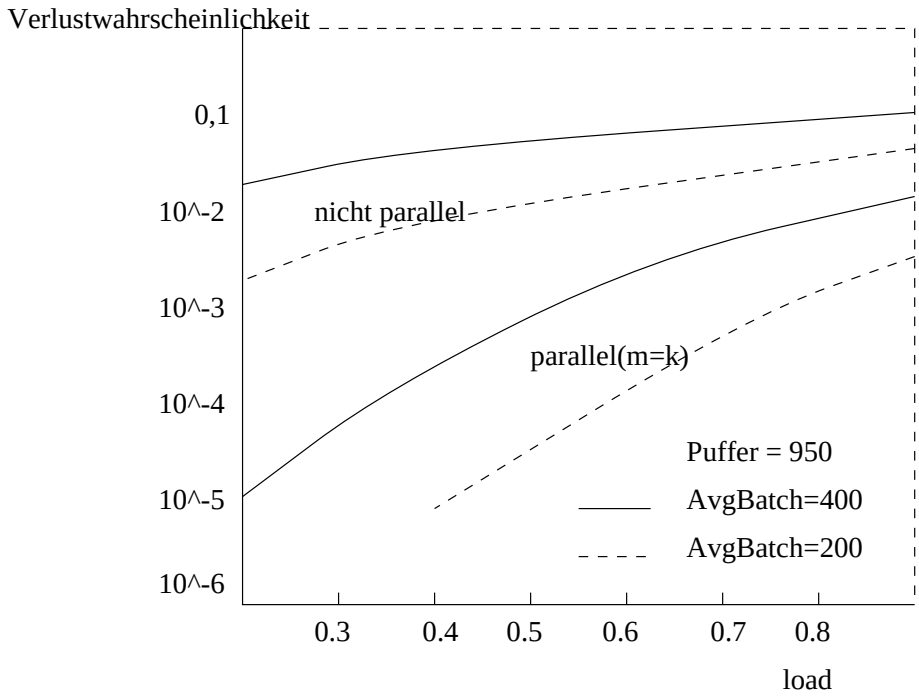

Abbildung 11: Verlustwahrscheinlichkeit als Funktion des offered load

4 Zusammenfassung

In dieser Seminararbeit wurde das Striping-Verfahren dargestellt. Striping ist eine Technik, die eine gute Lösung für die Erhöhung der Leistung eines Netzwerks darstellt. Diese Leistungserhöhung wird durch die Zusammenfassung paralleler Datenpfade gewonnen, was eine Eigenschaft der Multiplexing-Technik ist. Ein Punkt, bei dem beide Techniken sich unterscheiden, ist die Gewährleistung der Transparenz bei der Ausführung des Algorithmus für die höheren Schichten. Striping unterstützt dieser Transparenz, Multiplexing dagegen nicht.
Es wurde danach das Striping auf verschiedenen Schichten eines Hochleistungsprotokoll-Modell verglichen. Als Ergebnisse wurde zum Beispiel gezeigt, daß Striping auf höheren Schichten zu weniger Head-of-Line-Blockierung führt und daß Striping auf diesen Schichten in Software implementiert ist.

Literatur

[AdPV95] Hari Adiseshu, Guru Parulkar und George Varghese. A Reliable and Scalable Striping Protocol. Technischer Bericht, Department of Computer Science, Washington University, August 1995.

[DiLi95] Quan-Long Ding und Soung C. Liew. A Performance Analysis of a Parallel Communications Scheme for ATM Networks. Technischer Bericht, The Chinese University of Hong Kong, H.K., November 1995.

[TrSm95] C. Brendan S. Traw und Jonathan M. Smith. Striping Within the Network Subsystem. *IEEE Network*, Juli/August 1995, S. 22–29.

Ein Ansatz zur benutzergerechten Nutzung der Dienstgüte in ATM-Netzen

Marcus Schmidt

Kurzfassung

Moderne Anwendungen, beispielsweise im Bereich Multimedia, stellen hohe Anforderungen an die Qualität der Kommunikation und an die Geschwindigkeit der Netzwerke. Diese Anforderungen können erfüllt werden, wenn Kommunikationsressourcen effizient reserviert und den Datenströmen zugeteilt werden. In diesem Kontext kommt dem Management der Dienstgüte (Quality-of-Service, QoS) eine besondere Rolle zu. Im Rahmen dieser Arbeit werden zwei Szenarien vorgestellt, die am Center for Telecommunications Research (CTR) an der Columbia Universität in den letzten Jahren entwickelt wurden. Dies sind das Integrierte Referenzmodell (IRM) und seine Erweiterung das Erweiterte Referenzmodell (XRM). Weiterhin wird ein Modell zur objektorientierten Modellierung, das Binding-Modell, präsentiert.

1 Einleitung

In Verbindung mit modernen Anwendungen, die Hochleistungsanprüche an die Netzwerke stellen, spielt das Management der Dienstgüte eine besondere Rolle. Sie muß dem Benutzer direkt verfügbar gemacht und entsprechend verwaltet werden. Die Umsetzung einer solchen Forderung wäre in den 60'ern, als die UNI-und NNI-Konzepte eingeführt wurden, nicht denkbar gewesen, da die Leistungsfähigkeit der Arbeitsplatzrechner im Vergleich zu der der Netzwerkkomponenten gering war [LaLM95]. Dieses Verhältnis hat sich geändert, so daß heute die „Intelligenz" aus dem Netzwerk heraus zu den Benutzern verlagert werden kann. Die Möglichkeit, Dienste wie virtuelles Netzwerk oder Multicast als Objekte zu modellieren, verlangt nach einer Hochsprache, die Sprachkonstrukte zur Manipulation dieser Objekte zu Verfügung stellt. In Zukunft sollte der Entwurf eines Netzwerkdienstes so leicht sein, wie der Entwurf einer Applikation auf einem Rechner, der keinerlei Verbindung zum Netz hat. Weiterhin ist das Management der Dienstgüte eine Ende-zu-Ende Aufgabe, d.h. Hilfsmittel zur Beschreibung der Netzwerkressourcen und der Ressourcen in den Endsystemen sind notwendig.

Am CTR der Columbia Universität wurde zunächst das Integrierte Referenzmodell (IRM), ein Rahmenwerk für die Architektur von Breitbandnetzwerken, entwickelt. Das IRM und die zugrunde liegenden Prinzipien werden in Abschnitt 2 erläutert. Die Erweiterung des IRM, nämlich das Erweiterte Referenzmodell (XRM), wird in Abschnitt

3 vorgestellt. In Abschnitt 4 wird dann das Binding-Modell beschrieben. Das Binding-Modell ist ein objektorientierter Ansatz zur Modellierung von Netzwerkmechanismen. In diesem Zusammenhang werden auch Möglichkeiten zur Modellierung der Ressourcen vorgestellt (siehe Abschnitt 4.3).

2 Das integrierte Referenzmodell (IRM)

Das integrierte Referenzmodell ist das Modell einer Netzwerkarchitektur für Breitbandnetzwerke, das auf drei fundamentalen Prinzipien, nämlich dem Separationsprinzip, dem Schichtenprinzip und dem Prinzip des asynchronen Ressourcenmanagements, basiert. Diese drei Prinzipien ermöglichen ein logisches Verständnis von Verhaltens-, Entwurfs- und Implementierungsfragen eines Breitbandnetzwerkes. Ziel des Integrierten Referenzmodells ist es, ein Modell zur Organisation von Instanzen, die Daten transportieren, von Netzwerkinstanzen und Operatoren auf solchen Instanzen bereitzustellen. Dies erfordert die Ausführung der folgenden Aufgaben: Netzwerkmanagement/Kontrolle, Ressourcenmanagement/Kontrolle, Ressourcenmonitoring/Management, Verbindungsmanagement/Kontrolle sowie den Transport von Benutzerdaten.

Netzwerkmanagement/Kontrolle unterstützen den Netzwerkoperator bei Konfigurations-, Leistungs-, Fehler-, Abrechnungs- und Sicherheitsoperationen, die der Umgang mit einem Netz erfordert. Ressourcenmanagement/Kontrolle unterstützt die Realzeitmechanismen zur gemeinsamen Nutzung von Ressourcen sowohl im Netzwerk als auch in den Endsystemen. Aufgabe von Ressourcenmonitoring/Management ist es, Informationen über den Realzeitzustand des Netzwerks zu Verfügung zu stellen. Verbindungsmanagement/Kontrolle transportiert Signalisierungsinformation zu den Benutzern, d.h. Ende-zu-Ende, und weiterhin ist es für den Transport von Kontrollinformationen ins und im Netz zuständig. Der Transport der Benutzerdaten wird von Netzwerkinstanzen unterstützt.

Das IRM integriert Primitive für Monitoring, Kontrolle, Management und Datentransport in der Management-Architektur (MA), der Verkehrskontroll-Architektur (VKA) und der Datentransport-Architektur (DTA). Die damit verbundenen Aufgaben arbeiten auf zwei unterschiedlichen Zeitskalen: einer schnellen Zeitskala mit einer Reaktionszeit in der Größenordnung von Millisekunden und einer langsamen Zeitskala mit einer Reaktionszeit im Sekundenbereich. In dem vorgestellten Modell arbeiten die VKA und die DTA auf der schnellen Zeitskala und die MA auf einer langsamen.

In Abschnitt 2.1 werden grundsätzliche Prinzipien zur Definition einer Netzwerkarchitektur vorgestellt. Das Integrierte Referenzmodell wird in Abschnitt 2.2 erläutert.

2.1 Drei fundamentale Prinzipien

Im folgenden werden die drei schon erwähnten Prinzipien zur Organisation der Netzwerk- und Datentransportinstanzen und deren Verhältnis in einer Netzwerkarchitektur vorgestellt.

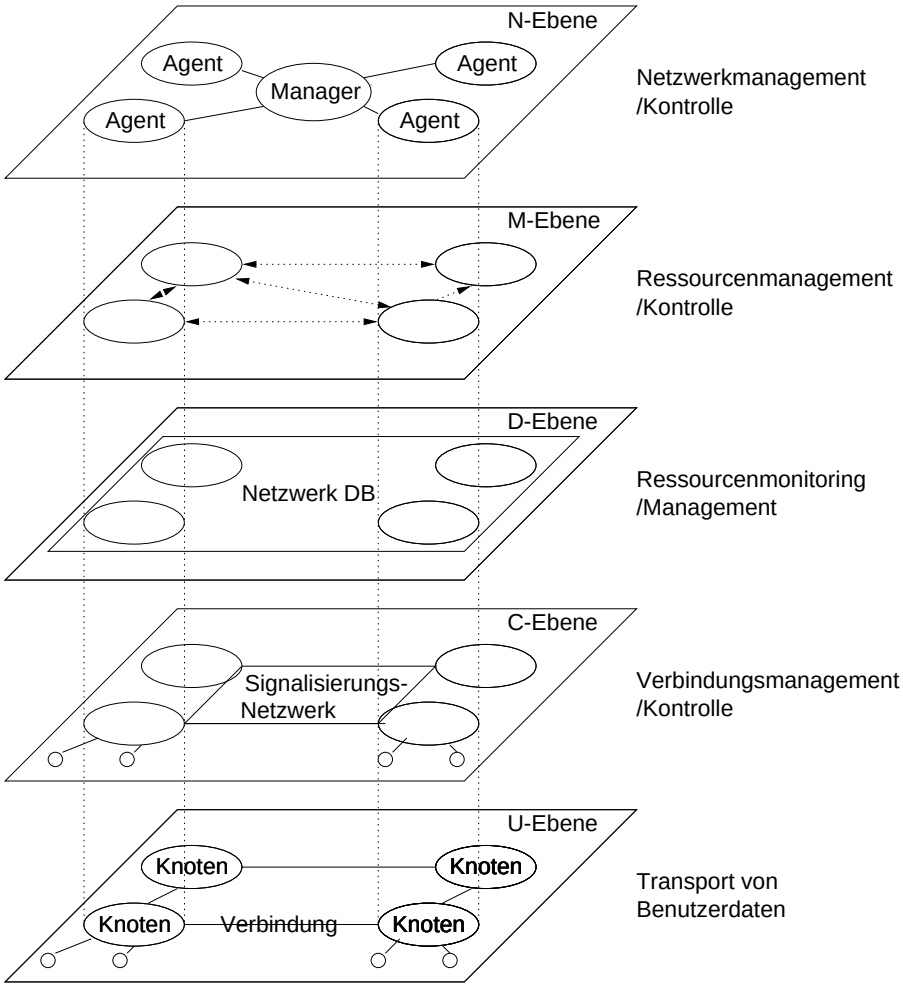

Abbildung 1: Anwendung des Separationsprinzips auf das IRM

2.1.1 Das Separationsprinzip

Das Separationsprinzip ist in zwei Teile gegliedert. Der erste Teil legt fest, daß die Kommunikationsaufgaben und die Kontrollaufgaben logisch in zwei Ebenen getrennt sind. Die Kontrollebene kontrolliert die Aktivitäten der Kommunikationsebene über eine Menge von Kontrollstrategien, die auf einer Menge von Messungen basieren. Eine Ebene wird durch eine Menge von Instanzen und ihre Beziehungen untereinander beschrieben. Der zweite Teil des Separationsprinzips beinhaltet die Teilung der Kontrollebene in vier Ebenen, wobei jeder Ebene eine unterschiedliche Aufgabe zugeordnet ist. Die Ebenen sind: Netzwerkmanagement/Kontrolle, Ressourcenmanagement/Kontrolle, Ressourcenmonitoring/Management und Verbindungsmanagement/Kontrolle (siehe Abbildung 1).
Netwerkmanagement/Kontrolle stellt Funktionen für Konfigurations-, Fehler-, Leistungs-, Abrechnungs- und Sicherheitsmanagement zu Verfügung. Trotz teilweiser Automatisierung ist die Einflußname eines Netzwerkmanagers nötig.
Die gemeinsame Nutzung von Ressourcen gehört in den Bereich von *Ressourcenmanagement/ Kontrolle*. Dies wird durch Algorithmen, die die Ressourcen in Echtzeit allokieren, erreicht. Die verwendeten Algorithmen fallen in fünf Klassen: Konfigurationskontrolle, Scheduling und Puffermanagement, Routing, Flußkontrolle und Zugangskontrolle.

Ressourcenmonitoring/Management ist für die Informationen über die Netzwerkobjekte zuständig. Dies beinhaltet auch das Vorbereiten von Sensoren für Monitoringzwecke und die Verarbeitung der anfallenden Daten. Die dabei entstehenden abstrahierten Daten werden für die Netzwerkkontrolle zu Verfügung gestellt.
In den Aufgabenbereich von *Verbindungsmanagement/Kontrolle* fällt der Aufbau von neuen Verbindungen, die neue Aushandlung von bestehenden Verbindungen und das Beenden von Verbindungen. Das Erkennen von (nicht) nutzbaren Netzwerkressourcen gehört auch zum Aufgabenbereich. Es wird deutlich, daß die Aufgaben eng mit denen aus dem Bereich Ressourcenmanagement/Kontrolle zusammenhängen.

2.1.2 Das Schichtenprinzip

Üblicherweise wird der Datentransport von einer Quelle zu einem Ziel auf verschiedenen Schichten betrachtet wie zum Beispiel Zwischensystem-zu-Zwischensystem, Endsystem-zu-Endsytem oder Prozeß-zu-Prozeß. Die Aufgabe der Datenübertragung erfordert nun eine fehlerfreie Wiederherstellung der Daten auf all diesen Schichten. Somit ist es möglich, den Prozeß der Datenübertragung zwischen Partnerinstanzen auf die Schichten aufzuteilen, zum Beispiel durch die Aufteilung in Subprozesse. D.h. es ist ein Vorgehen im Sinne des ISO/OSI-Schichtenmodells sinnvoll.

2.1.3 Das Prinzip des asynchronen Ressourcenmangements

Das Prinzip des asynchronen Ressourcenmangements bezieht sich auf die Modellierung der Arbeitsweise der Mechanismen aus der Management-Architektur (MA) und der Verkehrskontroll-Architektur (VKA), die den Prozeß der Datenübertragung unterstützen. Es spiegelt direkt die zeitlichen Anforderungen wider, die zwischen Objekten in einem verteilten System bestehen. Um dieses Prinzip zu verdeutlichen, nehme man an, die Mechanismen der MA seien zeitinvariant. Es bleiben also fünf Algorithmenklassen, deren Ausführung die Effizienz der Datenübertragung beeinflußt. Das sind Konfigurationskontrolle, Scheduling und Puffermanagement, Routing, Flußkontrolle und Zugangskontrolle. Das Prinzip des asynchronen Ressourcenmangements erklärt, daß jede der Variablen zur Netzwerkkontrolle, nämlich die oben genannten, an einer verteilten Kontrolle teilnehmen muß. Der Arbeitspunkt des Netzwerkes wird durch einen asynchronen Algorithmus zwischen diesen fünf Algorithmenklassen erreicht. Das Prinzip des asynchronen Ressourcenmangements erfordert nicht die Implementierung einer speziellen Klasse von asynchronen Algorithmen. Es erfordert, wie auch immer realisiert, den periodischen Austausch von Informationen über die Kontrollvariablen.

2.2 Die Netzwerkarchitektur

Die drei im vorigen Abschnitt beschriebenen Prinzipien stellen die Basis zur Entwicklung eines Referenzmodells für Breitbandnetzwerke, dem Integrierten Referenzmodell (IRM), dar. Die Struktur des IRM zeigt Abbildung 2. Das IRM modelliert die Kontroll- und Kommunikationsprimitive. Sie sind in der MA, VKA und DTA enthalten, wobei die Unterteilung zwischen MA, VKA einerseits und der DTA andererseits auf dem Separationsprinzip zwischen Kommunikation und Kontrolle basiert. Die Separation zwischen

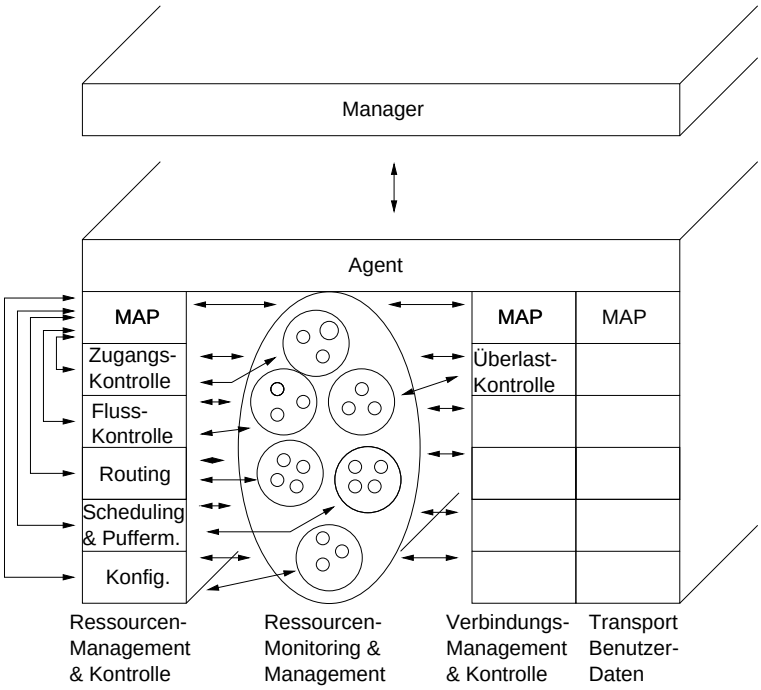

Abbildung 2: Das Integrierte Referenzmodell (IRM)

MA und VKA ist auf die unterschiedlichen Zeitskalen zurückzuführen, auf denen die Architekturen operieren.

Die Management-Architektur besteht aus der Netzwerkmanagement/Kontroll- oder N-Ebene. Die Verkehrskontroll-Architektur besteht aus der Ressourcenmanagement/ Kontroll- oder M-Ebene, der Ressourcenmonitoring/Management- oder D-Ebene und der Verbindungsmanagement/Kontroll- oder C-Ebene. Diese Aufteilung basiert auf dem zweiten Teil des Separationsprinzips. Die Datentransport-Architektur besteht aus der Ebene, die für den Transport der Benutzerdaten verantwortlich ist. Sie wird auch U-Ebene genannt. Die U- und C-Ebene sind horizontal geschichtet. Diese horizontale Unterteilung basiert auf dem Schichtenprinzip aus Abschnitt 2.1.2. Die D- und M-Ebene bestehen aus einer Menge von Objekten oder Modulen. Die N-Ebene entspricht von ihrem Entwurf her dem Manager/Agent-Prinzip und die Interaktion zwischen den Modulen in der N-, M- und C-Ebene basiert auf dem Prinzip des asynchronen Ressourcenmanagements. Die VKA und die TA interagieren, um die QoS-Garantien zu erfüllen, die beim Verbindungsaufbau ausgehandelt wurden.

3 Das Erweiterte Referenzmodell (XRM)

Das erweiterte Referenzmodell (XRM) ist die Erweiterung des Integrierten Referenzmodells zur Modellierung der Kommunikationsarchitektur von Netzwerken und multimedialen Arbeitsplatzrechnern (IRM ist ein Modell für Breitbandnetzwerke). Das XRM integriert Monitoring, Realzeitkontrolle, Management, Kommunikation und Primitive zur Abstraktion in fünf Ebenen: die Netzwerk/Systemmanagement- oder N-Ebene, die Ressourcenkontroll- oder M-Ebene, die Datenabstraktion/Management- oder D-Ebene, die Verbindungsmanagement- oder C-Ebene und die Datenübertragungs- oder U-Ebene. Die Unterteilung des XRM in die einzelnen Ebenen erfolgt nach denselben

Prinzipien wie die Unterteilung beim IRM. An dieser Stelle sei noch angemerkt, daß die Einschränkung des XRM auf multimediale Arbeitsplatzrechner dem IRM funktional gleicht.

3.1 RGB - Modell

Abbildung 3: Beziehung zwischen R-, G- und B-Modell

Um die Struktur des XRM-Modells transparenter zu gestalten, identifiziert man drei Modelle im XRM-Modell. Die Funktionalität, die im allgemeinen mit Breitbandnetzwerken und Multimediaprozessoren, Multimedianetzwerken, sowie Applikationen und Service-Netzwerken verbunden wird, definiert die einzelnen Modelle. Sie erhalten im weiteren den Namen R-, G- und B-Modell. So wie die drei Farben Rot, Grün und Blau das Spektrum des Lichtes ausmachen, besteht das XRM-Modell aus diesen drei Modellen. Die Schnittstellen zwischen den Modellen werden durch Abstraktionen oder Dienste beschrieben. QoS-Abstraktionen stellen Breitbandnetzwerk und Multimediaprozessoren dem Multimedianetzwerk zu Verfügung und Service Abstraktionen stellt das Multimedianetzwerk dem Applikationen- und Servicenetzwerk zur Verfügung (siehe Abbildung 3).

3.1.1 Die Funktionalität von Breitbandnetzwerk und Multimediaprozessoren (R-Modell)

Das R-Modell (siehe Abbildung 4) repräsentiert die Funktionalität des Breitbandnetzwerkes und der Multimediaprozessoren, welche dem Multimedianetzwerk einen Dienst in Form von QoS-Abstraktionen zu Verfügung stellen. Das R-Modell hat eine Struktur, die Management-, Kontroll- und Transportfunktionalität enthält. Eine kurze Beschreibung der Ebenen und der QoS-Abstraktionen sieht folgendermaßen aus:

- Die Funktionalität der N-Ebene entspricht der des Netzwerk- und Systemmanagements und besteht aus dem Monitoring und der Kontrolle von individuellen Zuständen, die zum Beispiel den Status einer Verbindung oder die Temperatur einer Netzwerkkarte widerspiegeln und als Objekte in der Management Information Base (MIB) realisiert sind. Eine Client/Server-Verbindung ist die Basis des Manager/Agenten-Modells zur Kontrolle der Netzwerkelemente.
- Die M-Ebene modelliert die Aufgaben der Ressourcenkontolle, wie zum Beispiel auf der Switch- oder Multiplexerebene Puffermanagement und Scheduling von Verbindungen. Flußkontrolle auf Zellebene ist ein weiterer wichtiger Ressourcenkontrollmechanismus.
- Die D-Ebene abstrahiert die Hauptnetzwerkkomponenten, wie Switches, Multiplexer und Medienprozessoren, in Form eines globalen verteilten Speichers. Insbesondere werden Kommunikationsverbindungen als FIFO Speicher modelliert. Diese Abstraktionen sind als Instanzen in der MIB enthalten.

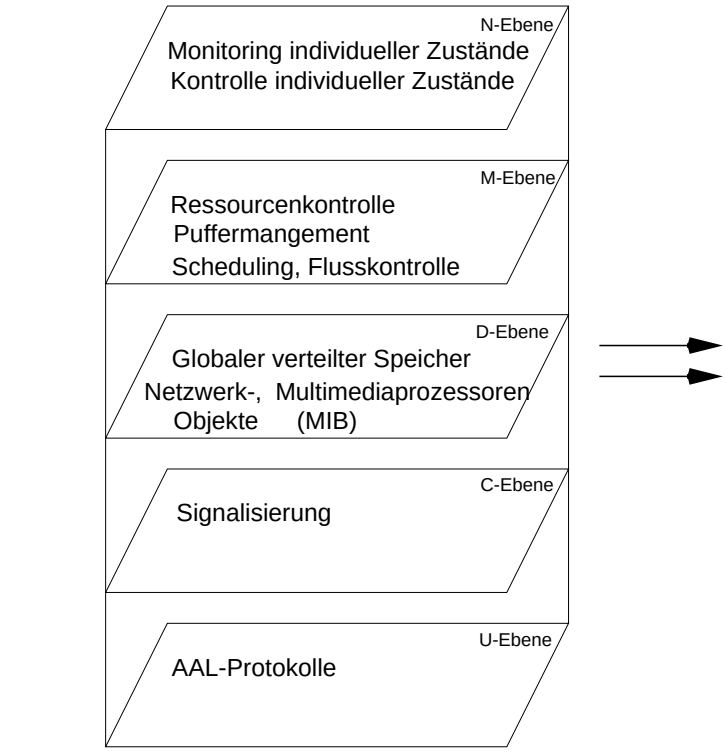

Abbildung 4: Funktionalität des Breitbandnetzwerkes und der Multimediaprozessoren (R-Modell)

- Die C-Ebene unterstützt den Austausch von Statusinformationen zwischen verteiltem Puffermanagement und Scheduling-Instanzen auf Verbindungsebene. Diese Mechanismen ermöglichen eine größere Netzwerkkapazität unter QoS-Anforderungen.

- Die U-Ebene definiert Adaptionsprotokolle auf der Zellebene für Segmentierung und Reassemblierung. Dies schließt die Funktionalität der ATM-Schicht und der ATM-Adaptionsschicht mit ein.

3.1.2 Die Funktionalität des Multimedianetzwerkes (G-Modell)

Die Funktionalität ist wieder in die fünf Ebenen des XRM-Modells unterteilt. Das Multimedianetzwerk stellt dem Service- und Applikationsnetzwerk eine Menge von Serviceabstraktionen zur Verfügung. Es nutzt zu diesem Zweck die QoS-Abstraktionen, die von der Ebene des Breitbandnetzwerks und den Medienprozessoren angeboten werden. An dieser Stelle fällt die Ähnlichkeit zum OSI-Modell auf. Einer Schicht des OSI-Modells entspricht in diesem Fall das Multimedianetzwerk. Sie bietet der nächst höheren Schicht (das Service- und Applikationsnetzwerk) Dienste an, indem sie die Dienste der unterliegenden Schicht (das Breitbandnetzwerk und die Medienprozessoren) nutzt. Dies ist aber die einzige Gemeinsamkeit zwischen den beiden Modellen. Zu den Serviceabstraktionen gehören unter anderem virtueller Kanal, virtueller Pfad, virtuelles Netzwerk und Multicast. Eine kurze Beschreibung der Funktionalität des Multimedianetzwerkes (siehe Abbildung 5) folgt:

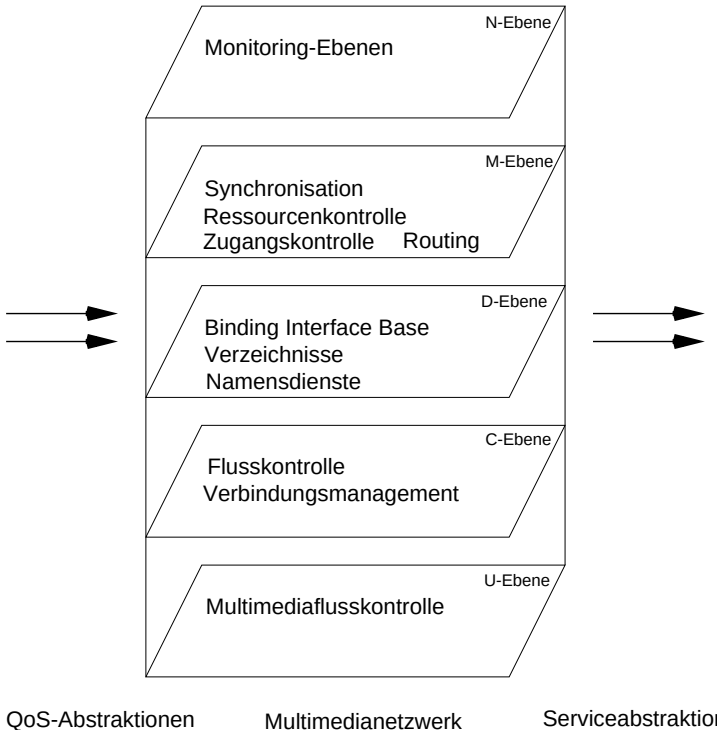

Abbildung 5: Funktionalität des Multimedianetzwerkes (G-Modell)

- Das Monitoring des Verhaltens von verteilten Systemen ist die Aufgabe der N-Ebene, wie zum Beispiel das Monitoring der verteilten objektorientierten Systeme und ihrer Interaktionen.
- Die M-Ebene bietet Mechanismen zur Ressourcenallokation, wobei die Schlüsselalgorithmen Routing und Zugangskontrolle sind.
- Die D-Ebene besteht aus einer Binding Interface Base (BIB), ein verteiltes Lager, welches Informationen über Instanzen enthält, die am Binding Prozeß (siehe Abschnitt 4) teilnehmen. Dienste werden als eine Menge von miteinander verbundenen Objekten definiert. Dies können Verzeichnis-, Agenten- oder Namensdienste sein.
- Die C-Ebene unterstützt Flußkontrolle ebenso wie Verbindungsmanagement und folglich werden damit verteilte Algorithmen unterstützt. In die C-Ebene gehören desweiteren Unicast- und Multicast-Algorithmen, die dem Verbindungsmanagement dienen.
- Die Funktionalität der U-Ebene schließt die Unterstützung einer Vielzahl von Multimediaprotokollen, wie Nutzung der direkten ATM-Dienste und anderer Realzeitprotokoll, ein. Die Protokolle können neben weit verbreiteten Transportprotokollen, die einen bestmöglichen Dienst anbieten (z.B. TCP), bestehen.

3.1.3 Die Funktionalität des Applikations- und Servicenetzwerkes (B-Modell)

An dieser Stelle erfolgt nur eine kurze Beschreibung der Funktionalität des Applikationen- und Servicenetzwerkes (siehe Abbildung 6).

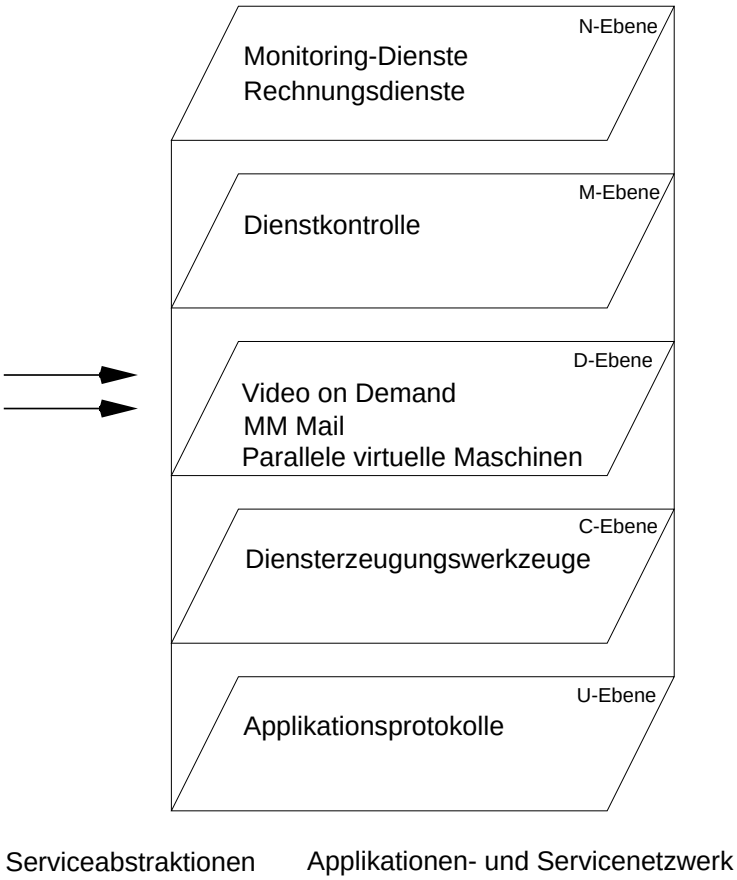

Abbildung 6: Funktionalität des Applikationen- und Servicenetzwerks (B-Modell)

Servicemanagement ist in der N-Ebene zu finden, d.h. Managementunterstützung für Sicherheits-, Zugangs-, Konfigurations-, Kosten- und Rechnungsdienste, um nur einige zu nennen. Die Zugangskontrolle ist sowohl in der N-Ebene, als auch in der M-Ebene definiert. Die Hauptaufgabe der M-Ebene ist aber die Dienstkontrolle und die Aushandlung von Netz- und CPU-Ressourcen. Die D-Ebene enthält eine Ansammlung von Objekten, die Dienste wie Multimedia-Mail, rechnergestützte kooperative Gruppenarbeit, Video-on-Demand, parallele virtuelle Maschinen, etc. repräsentiert. Navigations- und Diensterzeugungswerkzeuge sind in der C-Ebene angesiedelt und die Funktionalität der U-Ebene beinhaltet Applikationsprotokolle.

4 Binding-Modell

Das Binding-Modell ist ein Rahmenwerk für Entwurf und Implementierung von Netzwerkmechanismen, die der Kontrolle und dem Management von Multimediadiensten ATM-basierter Breitbandnetzwerke mit QoS-Garantien dienen. Binding bezieht sich dabei auf die Erstellung eines gewünschten Multimediadienstes mit QoS-Garantien. Dabei wird die Verbindung von einer Menge von Netzwerk- und Systemressourcen mit einem Multimediatransportprotokoll und dem Dienstmanagementsystem erstellt. Das Binding-Modell entspricht einem Diensterzeugungsmodell, welches die Vorgehensweise im XRM G-Modell beschreibt. Unter Diensterzeugung soll an dieser Stelle ein Prozeß verstanden sein, der Objekte zusammenstellt, wobei die Objekte von Algorithmen

aus den verschiedenen XRM G-Ebenen modifiziert werden. Das Binding-Modell beschreibt, ausgehend von den angebotenen Diensten an der Schnittstelle zwischen R- und G-Modell, wie Ressourcen zusammengestellt werden müssen.

Das Binding-Modell besteht aus zwei Blöcken: einer Menge von Zuständen, die sich in einer Binding Interface Base (BIB) befinden, und einer Menge von Algorithmen, die diese Schnittstellen modifizieren. Die Struktur mit den beiden Blöcken ist durch das Separationsprinzip gerechtfertigt. Dieses Prinzip verdeutlicht auch, was im XRM G-Modell standardisiert werden sollte und was nicht. In [LaLM95] wird vorgeschlagen, den Zugang zur Binding Interface Base zu standardisieren, aber die Algorithmen und die Diensterzeugung unberührt zu lassen.

4.1 Dienste an den Schnittstellen

4.1.1 Dienste an der Schnittstelle zwischen R- und G-Modell

Neben weiteren Diensten werden an der R-G-Schnittstelle der Physikalische-Verbindungsgraph und der Multimedianetzwerk-Kapazitätsgraph zu Verfügung gestellt. Der Physikalische-Verbindungsgraph enthält die topologischen Informationen des Netzwerks und der Ressourcen. Der Multimedianetzwerk-Kapazitätsgraph besteht aus einer Menge von Objekten, die die Qualität der Dienstabstraktion repräsentieren. Auf die Konzepte zur Bestimmung von Netzwerkkapazität und Ressourcen wird in Kapitel 4.3 eingegangen.

4.1.2 Dienste an der Schnittstelle zwischen G- und B-Modell

Neben anderen Diensten bietet die G-B-Schnittstelle Dienste wie virtueller Pfad, virtuelles Netzwerk und Multicast an. Genauer wird im Rahmen dieser Arbeit nicht auf die Dienste an dieser Schnittstelle eingegangen.

4.2 Das Diensterzeugungsmodell

4.2.1 Die Netzwerksicht

Die Binding Interface Base enthält eine Sammlung von Objekten, die Ressourcen wie Switches, Verbindungen und Multimediageräte modellieren. In der BIB ist auch der Physikalische-Verbindungsgraph und der Netzwerkkapazitätsgraph realisiert. Im XRM ist die BIB in der Telebase enthalten (die D-Ebene des G-Modells). Die Aufgabe des G-Modells ist die Umsetzung der Dienste, die das R-Modell anbietet, auf die Dienste, die dem B-Modell angeboten werden. Die Netzwerksicht der Diensterzeugung beschreibt nun, wie die verteilten Algorithmen, die am Prozeß der Diensterzeugung teilnehmen, im G-Modell interagieren. Der Diensterzeugungsprozeß generiert aus einer Menge von Objekten der BIB eine andere Menge von Zuständen. Der Prozeß wird von einer Diensterzeugungsinstanz durchgeführt, wobei viele dieser Instanzen parallel und verteilt arbeiten. Das Binding-Modell definiert nun die übergreifende Organisation der zur Diensterzeugung gehörenden verteilten Operationen. Alle Operationen befinden sich in den N-, M-, C- und U-Ebenen. Binding-Algorithmen entstehen beim

Verbindungsaufbau in Breitbandnetzwerken, bei verteilten Systemen, die Protokolle zur Synchronisation implementieren, sowie bei Protokollen, die Ressourcen allokieren. Neue Binding-Applikationen können hinzu genommen werden, ohne zugrunde liegende zu ändern. Des weiteren können mehrere eigenständige Binding-Algorithmen zur gleichen Zeit operieren.

4.2.2 Die Dienstsicht

Die Dienstsicht verdeutlicht die Schritte, die zu einer Diensterzeugung führen. Der Prozeß besteht aus fünf Schritten:

- Erzeuge ein Dienstskelett für Applikationen wie virtueller Kanal, virtueller Pfad, virtuelles Netzwerk oder Multicast. Die Struktur des Skeletts für einen virtuellen Kanal besteht zum Beispiel aus einem Graphen vom Quell- zum Zielknoten.

- Bilde das Skelett in den Namens- und Ressourcenraum ab und erzeuge dabei eine Netzwerkapplikation

- Verbinde diese Applikation mit einem Transportprotokoll und erzeuge dabei eine Transportapplikation

- Verbinde die Transportapplikation mit den Ressourcen und erzeuge dabei einen Netzwerkdienst

- Verbinde das Netzwerkmanagementsystem mit dem Netzwerkdienst und erzeuge dabei einen verwalteten Dienst. Wenn nun ein Dienst benötigt wird, stellt der Prozeß eine Dienstanfrage an die Instanz, die für die Bereitstellung des Dienstes zuständig ist. Diese führt dann die passenden Binding-Algorithmen zur Erzeugung des Dienstes durch.

4.3 Ressourcenmodellierung im Binding-Modell

Dieser Abschnitt befaßt sich mit der Modellierung der Ressourcen, auf denen die BIB aufbaut. Dazu werden die Ressourcen betrachtet, die im XRM G-Modell berücksichtigt werden. Das sind einerseits die Ressourcen im Breitbandnetzwerk und andererseits die Ressourcen in den Endsystemen, d.h. die Multimediaprozessoren. Der Grundgedanke basiert auf einem beobachtenden Ansatz zur QoS-Garantie in Breitbandnetzwerken, der von der Kapazität eines statistischen Multiplexers abstrahiert, wobei von einer Menge von vordefinierten Verkehrsklassen ausgegangen wird. Diese Abstraktion wird Schedulable Region genannt und repräsentiert die mehrdimensionale Kapazität des Multiplexers. Die Mehrdimensionalität dieses Netzwerkkapazitätsbereiches spiegelt die Anzahl der Verkehrsklassen wieder. Durch dieses Vorgehen ist es möglich, einen Bereich zu bestimmen, in dem der Multiplexer den QoS-Vorgaben genügen kann. Das genannte Konzept des Kapazitätsbereiches wurde auf die Endsysteme ausgedehnt. Die korrespondierende Instanz trägt den Namen Multimediakapazitätsbereich.

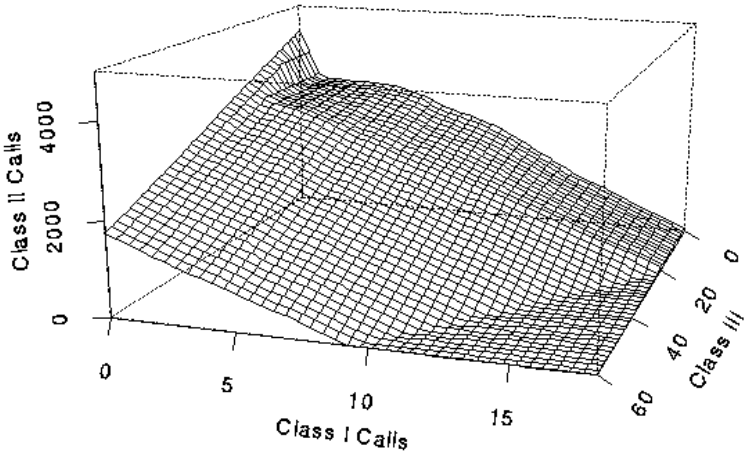

Abbildung 7: Schedulable Region eines Multiplexers mit drei Verkehrsklassen

4.3.1 Modellierung des Breitbandnetzwerkes

Schedulable Region Im XRM R-Modell wird das Netzwerk als eine Menge von Switches und Hochgeschwindigkeitsverbindungen verstanden, in dem Zellen verschiedener Verkehrsklassen von einer Quelle zu einem Ziel fließen. Der Switch nimmt einen Zellenstrom von ankommenden Verbindungen auf und leitet jede ankommende Zelle durch ein nichtblockierendes Verfahren weiter zu einer Verbindungskontrolle, wo die Zellen zur Übertragung über die Ausgangsverbindung gepuffert werden. Diese Verbindungskontrolleinheit ist im wesentlichen ein Multiplexer. Er besteht aus einer Menge von Puffern, einem Puffermanager und einem Scheduler. Der Multiplexer vermittelt zwischen den Zellen verschiedener Eingangsverbindungen einerseits und denen verschiedener Verkehrsklassen andererseits. Die Kapazität der Verbindung, die Verkehrsstatistiken, die Größe des Puffers und der Scheduling- sowie der Managementalgorithmus bestimmen, wieviele Verbindungen unter Einhaltung der QoS-Garantien unterstützt werden können. Die Menge der Punkte im mehrdimensionalen Raum, die Verbindungen mit QoS-Garantien auf Zellebene widerspiegeln, wird Schedulable Region genannt. Sie ist eine Ressourcenabstraktion, die an jeder Ausgangsverbindung eines jeden Switches im Netzwerk existiert.

Die Schedulable Region eines Multiplexers, der eine Kapazität von 100Mbit/s hat, ist in Abbildung 7 dargestellt. Auf jeder der drei Achsen ist die Anzahl der Verbindungen einer speziellen Verkehrsklasse aufgetragen. Die hervorgehobene Fläche begrenzt den Kapazitätsbereich des Multiplexers, d.h. für jede Kombination der Zellenanzahl (bzgl. der Verkehrsklassen) unter dieser Fläche sind die QoS-Garantien auf Zellebene erfüllt.

4.3.2 Modellierung der Multimedia-Arbeitsplatzrechner

Wenn das Netzwerkkapazitätsmodell auf die Endsysteme ausgedehnt wird, werden die Wünsche des Endsystembenutzers durch Dienstklassen mit QoS-Einschränkungen modelliert. Die Anzahl der Dienstklassen auf Benutzerebene wird wesentlich größer sein als die Anzahl der Verkehrsklassen auf Netzwerkebene, aber es ist nicht nötig, eine 1:1 Abbildung vorzunehmen, sondern es reicht vollkommen aus, eine Menge von Dienstklassen auf ein oder zwei Verkehrsklassen abzubilden. Auf diese Abbildung wird im Rahmen der vorliegenden Arbeit nicht weiter eingegangen.

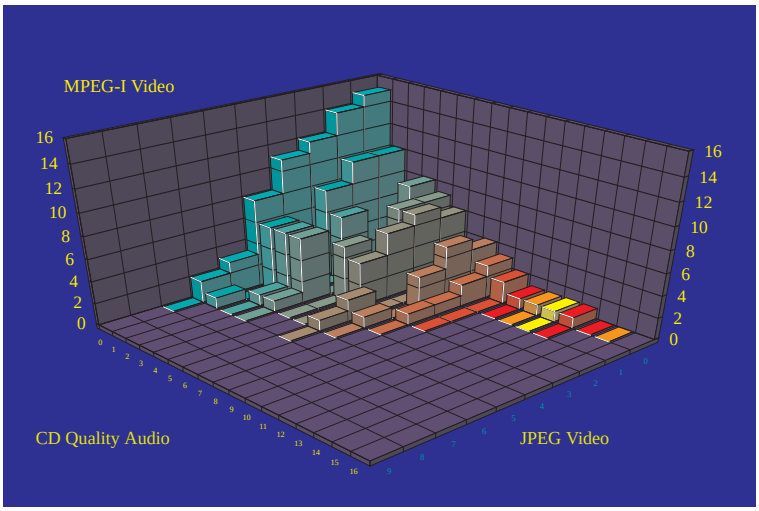

Abbildung 8: Multimediakapazitätsbereich einer Audio/Video-Einheit

Der Multimediakapazitätsbereich Der Multimediakapazitätsbereich ist die Erweiterung des Konzepts der Schedulable Region auf Multimediageräte mit QoS-Garantien in den Endsystemen. In einem Endsystem benutzen verschiedene Applikationen auch Daten von verschiedener Granularität, wie zum Beispiel eine Videokonferenzapplikation Video-Bilder, Best-Effort-Daten Seiten und Daten, die auf einen Speicher geschrieben werden, die Übertragung von Segmenten benötigen. Somit erscheint es schwierig, die QoS-Garantien für alle Anforderungen auf einem Prozessor zu erfüllen. Also liegt der Übergang zu einer Mehrprozessorarchitektur für Multimediaarbeitsplatzrechner nahe, einer Architektur, in der jeder einzelne Prozessor einen eigenen Datenstrom bearbeitet. Eine Unterteilung in eine Audio-Video-Einheit für Audio-Videodaten, eine Hauptrecheneinheit für Best-Effort-Daten und eine Speichereinheit ist möglich. Diese Unterteilung erlaubt nun QoS-Garantien. Jeder Datenstrom ist in Dienstklassen unterteilbar, wobei jede Klasse eine Dimension eines mehrdimensionalen Raumes widerspiegelt. Der Multimediakapazitätsbereich ist der Bereich, in dem die QoS-Garantien gewährleistet werden können, also ähnlich wie die Schedulable Region. Der Multimediakapazitätsbereich einer Audio/Video-Einheit ist in Abbildung 8 dargestellt.

Dienstklassen Eine Dienstklasse repräsentiert ein statistisches Modell für einen Datenstrom. Die Datengranularität eines Datenstroms wird durch die Größe der Daten bestimmt, die die Applikation benutzt. Um die multimediale Kapazität einer Audio-Video-Einheit zu bestimmen, wurden vier Dienstklassen in Erwägung gezogen, nämlich JPEG, MPEG-I, MPEG-II Videoströme und Audio in CD Qualität. Diese Dienstklassen können auf Grund ihrer verschiedenen statistischen Charakteristika eingeordnet werden:

- Klasse I: JPEG Videostrom, Peak Bit Rate 2.5 Mbit/s, Rahmengröße 320x240 Pixel mit 30 Rahmen/s
- Klasse II: MPEG-I Videostrom mit variabler Bitrate, PBR 1.5 Mbit/s, Rahmengröße 352x240 Pixel mit 30 Rahmen/s
- Klasse III: MPEG-II Videostrom mit konstanter Bitrate von 10Mbit/s, Rahmengröße 640x480 Pixel mit 30 Rahmen/s

- Klasse IV: Audio in CD-Qualität mit konstanter Bitrate von 1.411 Mbit/s

Quality of Service Die QoS-Einschränkungen für jede Dienstklasse werden durch eine Menge von Rahmenverlusten und Verlusteinschränkungen spezifiziert. Sie können für die oben genannten Klassen zum Beispiel folgendermaßen aussehen:

- Klasse I: JPEG Videostrom, maximale Verzögerung 55ms, maximale Rahmenverlustrate 2-5%
- Klasse II: MPEG-I Videostrom mit variabler Bitrate, maximale Verzögerung 55ms, maximale Rahmenverlustrate 2-5%
- Klasse III: MPEG-II Videostrom mit konstanter Bitrate, maximale Verzögerung 55ms, maximale Rahmenverlustrate 2-5%
- Klasse IV: Audio in CD-Qualität mit konstanter Bitrate von 1.411 Mbit/s, maximale Verzögerung 55ms und keine Verluste

5 Zusammenfassung

Um den hohen Anforderungen an die Qualität der Kommunikation und an die Geschwindigkeit gerecht zu werden, ist eine Modellierung der Netzwerke notwendig. Das IRM modelliert die Netzwerkarchitektur eines Breitbandnetzwerkes, wobei der Modellierung das Separaionsprinzip, das Schichtenprinzip und das Prinzip des asynchronen Ressourcenmanagements zugrunde liegen. Das XRM basiert auf den gleichen Grundüberlegungen wie das IRM. Es ist die Erweiterung zur Modellierung von Netzwerken und multimedialen Arbeitsplatzrechnern und es baut auf drei einzelnen Modellen, dem R-, G- und B-Modell auf. Durch die drei Modelle wird das XRM sehr komplex und mächtig, aber es erlaubt eine transparentere Sicht auf die Modellierung. In dieses RGB-Modell kann nun ein objektorientierter Ansatz zur Modellierung und Erstellung von Multimediadiensten eingebettet werden. Dieser Ansatz wird durch das Binding-Modell beschrieben. In diesem Zusammenhang ist eine Modellierung des Breibandnetzwerkes und der multimedialen Arbeitsplatzrechner erstrebenswert, was durch die Prinzipien der Schedulable Region und den Multimediakapazitätsbereich erreicht wird.

Als eine Umsetzung des IRM wurde am CTR der Columbia Universität ein „Intelligenter Mutiplexer“ modelliert. Die Modellierung befindet sich in [Laza92] und eine Beschreibung einer Implementierung findet man in [LaTG90]. Die vorgestellten Möglichkeiten der Ressourcenmodellierung (Schedulable Region und Multimediakapaziätsbereich) wurden zusammen mit dem Binding-Ansatz in [HILY96] für den Entwurf einer Architektur mit Ende-zu-Ende QoS verwendet.

Literatur

[HILY96] J.-F. Huard, I. Inoue, A.A. Lazar und H. Yamanaka. Meeting QOS Guarantees by End-to-End QOS Monitoring and Adaptation. Technischer Bericht, Department of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York, NY 10027-6699, 1996.

[LaLM95] Aurel A. Lazar, Koon Seng Lim und Franco Marconcini. Binding Model: Motivation and Description. Technischer Bericht, COMET Group Department of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York, NY 10027-6699, 1995.

[LaTG90] A.A. Lazar, A. Temple und R. Gidron. MAGNET II: A Metropolitan Area Network Based on Asynchronous Time Sharing. *IEEE Journal on Selected Areas in Communications* SAC-8(8), 1990, S. 1582–1594.

[Laza92] Aurel A. Lazar. A Real-Time Management, Control and Information Transport Architecture for Broadband Networks. Technischer Bericht, Department of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York, NY 10027-6699, 1992.

IP-Switching anstelle von Routing

Robert Gröver

Kurzfassung

Im Rahmen dieses Beitrags soll die Funktionsweise des IP-Switching erläutert und mit dem klassischen Routing verglichen werden. Dabei liegt der Schwerpunkt auf den neuen ATM-basierten Netzwerken. Ferner soll ein neues Protokoll, das General Switch Management Protokoll GSMP, welches beim IP-Switching in ATM-Netzen eingesetzt wird, genauer vorgestellt werden. Hierzu zählt auch qGSMP, eine Erweiterung in Richtung Dienstgüteunterstützung, die vor kurzem für GSMP vorgeschlagen wurde.

1 Einleitung

Das Internet wächst mit rasanter Geschwindigkeit. Die Zahl der Rechner, die an das Internet angeschlossen sind, hat sich schätzungsweise alle 56 Wochen seit 1989 verdoppelt. Für die Zahl der Web-Server ist dies zumindest in den letzten drei Jahren der Fall. Während dieses rasante Wachstum anhält und die Zugriffsgeschwindigkeiten bei den Internetprovidern sich ständig erhöhen, wird von den IP-Routern im Netz eine Verarbeitungsleistung von mehreren Gigabit pro Sekunde (Gbps) abverlangt. Firmennetze haben sich im Rahmen der Globalisierung von Unternehmen von kleinen LANs zu weltumspannenden Netzwerken entwickelt, und neue Anwendungen im Multimediabereich, die mit hochwertiger Video- und Audioübertragungen die Grenzen der Verarbeitungsleistung sprengen, sind weitere Beispiele für die gestiegenen Anforderungen.
Die in Netzen existierenden Router, basierend auf einem Bus und zentralen Processor, sind in der Lage, eine maximale Last von einigen hunderttausend Paketen pro Sekunde (kpps) und einem Gbps zu verarbeiten. Um diese Grenze zu überschreiten, sind alternative Architekturen notwendig. Mit IP-Switching stellt die Firma Ipsilon eine alternative Architektur für IP-Router vor.
Der folgende Beitrag soll diese vorstellen. Dabei soll sowohl der Aspekt der Integration in bestehende Netze wie auch die Unterstützung von Dienstqualitäten betrachtet werden.

2 Begriffe und Grundlagen

2.1 Routing und Switching

Die Begriffe Routing und Switching werden derzeit sehr ungenau verwendet. Hersteller von Routern sprechen von Layer-3-Switching und Hersteller von Switches von Routing-

Funktionalität. Bei der Einführung des Ethernets hat man mit Repeatern, Hubs und Bridges Geräte geschaffen, mit denen sich das Netzwerk gliedern läßt, denn mit der Vergrößerung von Netzen und der Anzahl der vorhandenen Stationen am Netz verringert sich die Brandbreite aufgrund vermehrter Kollsionen (Ethernet).

Eine Lösung ist, das Gesamtnetz über ein Gerät aufzuteilen, das Datenverkehr nur dann zwischen den Teilnetzen weiterleitet, wenn er explizit für eine Station in einem anderen Teilnetz als dem Ausgangsnetz bestimmt ist. Dadurch lassen sich das Datenaufkommen und die Kollisionen in den einzelen Teilnetzen reduzieren. Diese als Brücken (Bridges) bezeichneten Geräte verbinden zwei oder mehr Subnetze, die wiederum aus mehreren Segmenten bestehen können. Um den Datenverkehr zwischen den Subnetzen trennen zu können, werten Brücken die Hardwareadressen (MAC-Adresse) der angeschlossenen Geräte aus. Mit Brücken ist es hingegen nicht möglich, Kommunikation zwischen Netzen zu ermöglichen, die nicht direkt an die Brücke angeschlossen sind, sondern durch ein oder mehrere andere Netze getrennt sind. Dies ist die Aufgabe von Routern, die über IP-Adressen und entsprechende Routing-Protokolle den Weg für die Daten auch in Netzen feststellen können, die lokal nicht erreichbar sind.

Ein Switch stellt prinzipiell nichts anderes dar als eine Bridge mit mehreren Ports, die in der Lage ist, mehrere Subnetze zu verbinden. Im Unterschied zu einer Bridge ermöglicht ein Switch aber den parallelen Datentransfer zwischen angeschlossenen Geräten. Damit erreicht ein Switch einen höheren Durchsatz an Daten, muß aber auch eine entsprechend höhere interne Verarbeitungsgeschwindigkeit bzw. -leistung aufweisen. Sowohl Brücken wie Switches arbeiten auf Schicht zwei des OSI-Referenzmodelles. Die durch Hersteller verwendeten Begriffe Multi-Layer-Switching, IP-Switching, Layer-3-Switching suggerieren, daß Switching auf OSI-Schicht 3 stattfindet. Folgt man den Definitionen des OSI-Referenzmodells, so kann nur auf Schicht 2, der Sicherungsschicht, geswitcht werden, auf Schicht 3 wird geroutet. Layer-3-Switching gibt es im eigentlichen Sinne gar nicht. Tatsächlich bedeutet die neuen Techologien die Kombination von Schicht-2-Switching und Schicht-3-Routing.

2.2 Asynchroner Transfer Mode (ATM)

Mit dem Asynchronous Transfer Mode (ATM) wurde bereits eine neue Architektur eingeführt, die eine entsprechende Verarbeitungsleistung aufweist. ATM hat sehr viel Aufmerksamkeit auf sich gezogen aufgrund seiner hohen Kapazität, seiner Bandbreiten-Skalierbarkeit und seiner Fähigkeit, unterschiedlichste Transportdienste unterstützen zu können. ATM basiert auf der schnellen Vermittlung von Zellen. Zellen sind Pakete von 53 Byte Länge. Um einen Datentransfer durchzuführen, wird bei ATM zunächst eine virtuelle Verbindung mit den entsprechenden Dienstgüteparametern aufgebaut. Jede Verbindung erhält eine kurze Pfad- und Kanalkennung (Virtual Path Identifier VPI, Virtual Cancel Identifier VCI). Jede Zelle trägt diese Kennung und wird mit ihrer Hilfe in den einzelnen Vermittlungsknoten, ATM Switches, weitervermittelt. Aufgrund dieser kurzen Kennung und des verbindungsorientierten Konzepts ist es möglich, daß Weiterleiten (Switching) hardware-technisch zu realisieren und Verarbeitungsleistungen von mehreren hundert Gigabit pro Sekunde zu ermöglichen.

Jedoch ist ATM verbindungsorientiert, wohingegen die Mehrheit der modernen Netzwerkprotokolle verbindungslos arbeiten. Diese Unstimmigkeit hat dazu geführt, daß

weitere Protokolle entwickelt wurden, um diese Unstimmigkeit zu umgehen. Besondere Anstrengungen wurde gemacht, um ATM und IP zu integrieren. Mit IP-over-ATM, LAN-Emulation und Multiprotocol-over-ATM (MPOA) wurden Möglichkeiten geschaffen, das Internet Protokoll (IP) über ATM zu betreiben. Jedoch beinhalten alle diese Versuche eine erhöhte Komplexität, geringere Effektivität und zum Teil Verdopplungen von Funktionalität im Protokollablauf. Sowohl IP als auch ATM benötigen ein eigenes Routing-Protokoll. Dies bedeutet nicht nur die Dublizierung von Routing-Protokollen, sondern auch von Wartungs- und Managementaufgaben. Zudem müssen weitere Funktionen hinzugefügt werden, um zwischen diesen Managementebenen zu kommunizieren.

3 IP-Switching

3.1 Struktur eines IP-Switches

IP-Switching von Ipsilon versucht die Vorteile von ATM, die schnelle Switching Hardware und die Verarbeitungsleistung, mit einem Router zu koppeln. Ipsilon setzt hier bewußt ATM-Hardwarekomponenten ein, da diese standardisiert sind und kostengünstig zur Verfügung stehen (im Vergleich mit ähnlich leistungsfähigen Routern). Existierende Router sind teuer und begrenzt in ihrem Durchsatz im Vergleich zu Switches.

Ein IP-Switch besteht aus einem IP-Switch-Controller und einem ATM-Switch (Abbildung 1). Der ATM-Switch wird ohne irgendwelche Modifikationen eingesetzt. Die komplette Software, die innerhalb des Kontrollprozessors oberhalb der AAL-5-Schicht von ATM angesiedelt ist, wird entfernt. Somit wurden sämtliche Signalisierungs- und Routingfunktionalitäten, Adreßauflösungsprotokolle und Funktionen zur LAN-Emulation der ATM-Software entfernt. An die Stelle der ATM-Software tritt das einfache Kontrollprotokoll GSMP [NEHH+96a], welches ebenfalls von Ipsilon entwickelt wurde. GSMP ermöglicht den Zugriff auf die Switch-Hardware. Der ATM-Switch wird mit dem Controller über einen seiner Ports mit der Switch-Controller Schnittstelle verbunden. Über diese Verbindung wir das GSMP eingesetzt.

Der IP-Switch-Controller ist ein leistungsstarker Prozessor, auf dem eine Standard IP-Routing Software mit einigen Erweiterungen läuft, die es ermöglicht, die Switching-Hardware zu nutzen. Die Erweiterungen beinhalten das Ipsilon Flow Management Protokoll IFMP [NEHH+96b], um IP-Flows virtuellen ATM-Verbindungen zuzuordnen, eine Klassifikation von IP-Flows, welche entscheidet, ob ein IP-Flow geswitcht wird, und das GSMP, um die ATM-Switch Hardware zu kontrollieren. In einem IP-Switch Netzwerk werden zwischen den einzelnen, miteinander verbundenen Switches als erstes sogenannte Default-ATM-Verbindungen über ausgezeichnete Kanäle und Pfade aufgebaut (VCI=15 VPI=0), über die zunächst der gesamte Datenverkehr abgewickelt wird. IP-Pakete, die nicht einem geswitchten Flow angehören, werden über diese weitergeleitet.

3.2 IP-Flows

Die grundlegende Idee bei IP-Switching ist, den Datenstrom von IP-Paketen zu analysieren und sogenannte IP-Flows zu erkennen. IP-Flows werden anhand der Felder

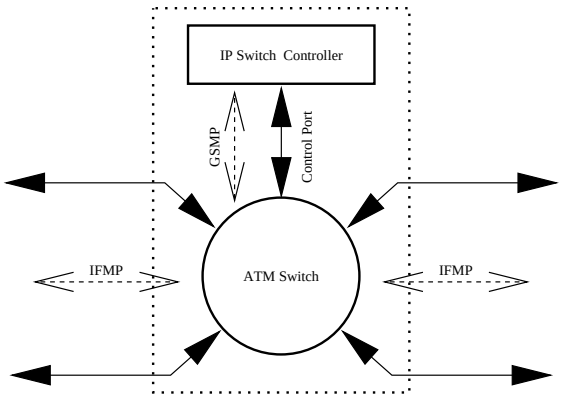

Abbildung 1: Struktur eines IP-Switches

des TCP/IP(UDP/IP)-Paketes, die auch für das Routing maßgeblich sind, bestimmt: Type-of-Service, Protokoll, IP-Quelladresse, IP-Zieladresse, Quell-Port und Ziel-Port. Zwei Pakete gehören zu einem IP-Flow, wenn die Werte der einzelnen Felder identisch sind.

Es besteht aber auch die Möglichkeit, weitere Typen von Flows aus einer anderen Kombination dieser Felder zu definieren. Derzeit sind zwei Arten von Flows definiert: Port-Pair-Flow (Flow Typ 1) und Host-Pair-Flow (Flow Typ 2).
Ein Typ 1 Flow besteht zwischen zwei Port-Adressen eines Quell- und eines Zielrechners und den zugehörigen IP-Adressen. Zusätzlich müssen noch die Felder Type-of-Service, Protokoll und Time-to-Live in den Paketen übereinstimmen. Dieser Typ erlaubt es, Flows mit unterschiedlicher Dienstgüte zwischen den Anwendungen (Ports) eines Quell- und eines Zielrechners zu unterscheiden.
Ein Host-Pair-Flow besteht zwischen einer Quell- und einer Ziel-IP-Adresse, wobei alle Pakete dieselbe Time-to-Live haben.

Wenn ein Paket auf dem Default-Kanal eintrifft, wird es zusammengesetzt (die Übertragung geschieht in ATM-Zellen) und an den Switch-Controller für die Weiterleitung übergeben. Dieser routet das Paket in der gewohnten Weise (Store-and-Forward), gleichzeitig führt er aber auch eine Flow-Klassifikation für das Paket aus. Diese Klassifikation dient dazu festzustellen, ob weitere Pakete durch die ATM-Hardware geswitcht werden sollen oder ob weiterhin eine Punkt-zu-Punkt Weiterleitung durch den IP-Switch erfolgen soll. Die Flow-Klassifikation geschieht in jedem IP-Switch lokal. Die Felder der Pakete werden entsprechend der Definition des Flows gelesen und ausgewertet, um eine Entscheidung nach der lokalen Verarbeitungsstrategie zu treffen. Beispielsweise kann auf ausgezeichnete Quell- und Ziel-Port-Nummer geachtet werden, um bestimmte Anwendungen zu ermitteln. Flows, die zu FTP Verbindungen (Port 20) gehören, werden beispielsweise geswitcht, wohingegen kurze Anfragen an DNS-Server als Datagramme durch den IP-Switch-Controller verarbeiten werden.
Eine weitere Möglichkeit der Flow-Klassifikation ist es, das Aufkommen von Paketen eines Flows zu bewerten. Wird eine festgelegte Grenze von Paketen pro Zeitintervall überschritten, wird der Flow geswitcht. Dies ist sinnvoll, da nicht jede Anwendung eine ausgezeichnete Port-Nummer besitzt. Eine Bewertung dieser Flow-Klassifikatoren erfolgt in Abschnitt 6.1. Ziel ist es, langandauerende Flows mit viel Datenverkehr zu lokalisieren. Heutige Multimedia-Anwendungen mit Sprach-, Bild- und Videoübertragung sind ein gutes Beispiel.

3.3 Ipsilon Flow Management Protokoll (IFMP)

Wenn der Flow-Klassifikator entscheidet, daß ein Flow geswitcht werden soll, wählt er unter den freien Kanalnummern des Eingangsports (Port i), auf dem er den Flow empfängt, eine Nummer aus (VCI=x). Ebenso wählt der Switch-Controller eine freie Kanalnummer auf dem Kontrollport c aus (VCI=x'). Der Kontrollport verbindet den ATM-Switch mit dem IP-Switch-Controller. Diese Verbindung kann sowohl eine explizit physikalische als auch eine virtuelle sein. Der ATM-Switch ist nun angewiesen, alle eingehenden Zellen auf Kanal x auf Kanal x' des Kontrollports weiterzuleiten (Abbildung 2 (1)). Nachdem dieser Eintrag innerhalb der Umsetzungstabelle des Eingangsports i mittels GSMP vorgenommen wurde, sendet der IP-Switch-Controller eine IFMP-Nachricht (redirection message) an den stromaufwärts liegenden IP-Switch. Diese Nachricht beinhaltet die ausgewählte Kanalnummer bzw. Marke x, die Flow-Beschreibung (IP-Adresse, Port-Adressen, ...) und die Lebensdauer (Lifetime). Mit dieser Nachricht wird der stormaufwärts liegende IP-Switch aufgefordert, alle Pakete, die der Flow Beschreibung entsprechen, mit der virtuellen Kanalkennung x (VCI=x) weiterzuleiten. Durch die Lebensdauer wird spezifiziert, wie lange die Umleitung bestand hat bzw. wann sie gelöscht werden kann.

Ab jetzt treffen alle Pakete, die diesem Flow mit der Kennung x angehören, beim Switch-Controller auf Port c mit der Kanalkennung x' ein. Nach wie vor verarbeitet dieser die Pakete, nachdem er sie aus den ATM Zellen zusammengesetzt hat, nach dem Store-and-Forward Prinzip. Ein Beschleunigung der Verarbeitung ist jetzt bereits erreicht, da die Routing-Entscheidung für diesen Flow mit der Kennung x' bereits im Cache zwischengespeichert ist und nicht für jedes Paket erneut getroffen werden muß.

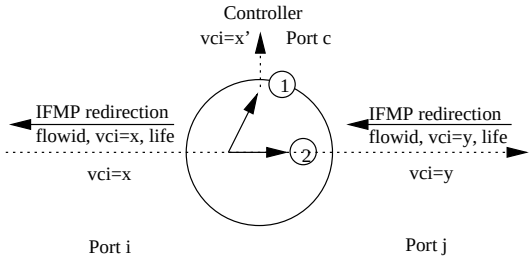

Abbildung 2: Aufbau eines geswitchen Flows

Die eigentliche Leistungssteigerung entsteht, wenn der stromabwärts liegende IP-Switch ebenfalls diesen Flow erkennt, umleitet und ihm eine Kennung VCI=y zuweist. Um die Wahrscheinlichkeit für das Zustandekommen dieses Zustandes möglichst hoch zu halten, ist es wegen der lokal unabhängigen Entscheidung eines jeden Switches notwendig, eine einheitliche Flow-Klassifizierung für das zu administrierende Netzwerk zu haben. Erhält der IP-Switch vom stromabwärtsliegenden IP-Switch eine Redirection-Nachricht auf dem Port j mit der Kanalkennung VCI=y, kann sämtlicher Verkehr, der diesem Flow zugeordnet ist, in der ATM-Hardware direkt geswitcht werden. Dazu wird die Umsetzungstabelle des Switches so beschrieben, daß ab sofort alle Zellen auf Port i mit Kennung x auf den Port j mit der Kennung y umgesetzt werden. Als Folge wird nun der Datenstrom nicht mehr vom IP-Switch-Controller verarbeitet bzw. geroutet, sondern durch die ATM-Hardware direkt auf den entsprechenden Ausgangsport geswitcht (Abbildung 2 (2)).

Erfolgt das Umschalten zwischen Routing und Switching, kann es zu Reihenfolgevertauschungen kommen. Das letzte Paket, das per Routing den IP-Switch verläßt, kann nach dem ersten zu switchenden Paket ausgeliefert werden. Da Pakete aus Zellen bestehen, kann es passieren, das innerhalb eines Paketes umgeschaltet wird. Dies hat den Verlust des Paketes zur Folge. Diese Fehlerursache kann vermieden werden, wenn man den Pfad zwischen Quelle und Ziel rückwärts aufbaut. Dieser Effekt kann dadurch erreicht werden, daß Endsysteme früher als IP-Switches im Netzinneren einen Flow erkennen und switchen. Auf diese Weise wir der geswitchte Pfad Stück für Stück aus Richtung des Empfängers aufgebaut, während der Datenverkehr weiterhin über den Store-and-Forward Pfad fließt. Das Umschalten zwischen Switching und Routing kann durch den IP-Switch bei dieser Vorgehensweise an der Grenze eines IP-Paketes durchgeführt werden.
Wenn ein IP-Switch eine Redirection-Nachricht annimmt, ändert sich auch das Paketformat mit dem er die Pakete eines Flows verschickt. Bei der Übertragung von Paketen über die Default-Verbindung wird eine LLC/SNAP Einkapselung über der AAL-5 Adaptionsschicht von ATM vorgenommen.
Im Gegensatz hierzu werden bei allen IP-Paketen eines geswitchten Flows sämtliche Felder aus dem IP-Kopf entfernt, durch die der Flow definiert ist. Das IP-Paket mit dem verkleinerten Kopf wird nun in einen Rahmen der AAL-5-Schicht verpackt und über den definierten virtuellen ATM-Kanal verschickt. Die entfernten Felder werden gespeichert und mit der virtuellen ATM-Kanalnummer assoziiert. Das ursprüngliche Paket kann zurückgewonnen werden, indem über die Kennung auf die entsprechenden Paketkopf-Felder zugegriffen wird.

Dieser Ansatz wurde aus Sicherheitsgründen gewählt. Es erlaubt dem IP-Switch ähnlich einer Firewall die Flows zu überwachen, ohne daß er den Inhalt jedes einzelnen Paketes überprüfen muß. So kann verhindert werden, daß ein Benutzer einen geswitchten Flow zu einer zulässigen Adresse oder einem Dienst hinter einer Firewall aufbaut und anschließend die Pakete mit einem anderen Kopf versieht, um Zugriff auf eine geschützte Adresse bzw. Dienst zu erlangen. Der Switch kann mit der Rekonstruktion der Pakete solche Änderungen feststellen und unterbinden.

3.4 IP vs. IP-Switching

Eine grundlegende Voraussetzung von IP ist, daß die Lebenszeit (Time-to-Live, TTL) eines IP-Paketes bei jedem Knotenpunkt verringert wird. Wenn die TTL auf Null ist, muß ein Paket verworfen werden. Aus diesem Grund beinhaltet jeder Flow-Klassifikator das TTL-Feld, um sicherzustellen, daß der Wert des TTL-Feldes am Ende des geswitchten Pfades dem eines von Punkt-zu-Punkt weitergeleiteten Paketes entspricht. Dementsprechend wird ein Paket mit TTL=0 nicht weitergeleitet. Das-TTL Feld wird bei einem geswitchten Flow nicht mitübertragen, wird aber durch die im Ziel-Switch gespeicherten Informationen wiederhergestellt.
Eine Folge aus der Aufnahme des TTL-Feldes in die Flow-Klassifikation ist eine vermehrte Anzahl von Flows aufgrund unterschiedlicher TTL-Werte. Unterschiedliche TTL-Werte erzwingen zwei separate Flows.

Ein weiteres wichtiges Feld im IP-Paket stellt die Kopfprüfsumme (Header Checksum) dar. Um die Korrektheit dieser Prüfsumme, die auch das TTL-Feld mit einschließt, zu bewahren, wird beim Übergang von Routing auf Switching die Prüfsumme entfernt

und im Controller gespeichert. Am Ende des geswitchten Flows wird unter anderem die dort gespeicherte Prüfsumme und das TTL-Feld wieder hinzugefügt.
Bei dieser Vorgehensweise wird davon ausgegangen, daß auf dem geswitchten Übertragungsweg keine Fehler im Paket entstehen. Entlang des geswitchten Übertragungsweges findet keine Fehlerüberprüfung statt. Falls Fehler innerhalb des IP-Kopfes auftreten, werden diese durch eine falsche Prüfsumme erkannt, da das Entfernen und Hinzufügen des TTL-Feldes so geschieht, daß die Prüfsumme der eines gerouteten Paketes entspricht. Das Paket wird verworfen, wenn die Prüfsumme das nächste Mal überprüft wird.

3.5 Stabilität des Verfahrens

Es ist wichtig, daß das Protokoll sich stabil gegenüber Fehlern im Netz, wie dem Ausfall eines Knoten oder eines Übertragungsweges, verhält. Solche Fehler bewirken Inkonsitenzen, die es gilt zu vermeiden bzw. deren Existenz möglichst schnell zu erkennen. Eine Fehlermöglichkeit dieses Entwurfes besteht darin, daß ein stormaufwärtsliegender Switch einen Flow A mit einer für den stromabwärtsliegenen Switch unbekannten Kennung x sendet. Es gibt hier zwei Möglichkeiten für das weitere Vorgehen:
Der Switch geht davon aus, daß ein weiterer Flow übertragen wird und versucht diesen weiterzuleiten. Die Daten werden fehlgeleitet und später verworfen.
Der Switch ignoriert den Flow mit der unbekannten Kennung und verwirft alle Pakete, die mit dieser Kennung eintreffen. Um diese Fehler zu unterbinden, enthält GSMP ein Protokoll (adjacency protocol), das es erlaubt, Nachbarknoten zu identifizieren, Verbindungszustände zu synchronisieren und den Wechsel eines Nachbarknotens festzustellen.

Bevor IFMP-Nachrichten verschickt werden können, werden durch dieses Protokoll zunächst die Nachbarn bestimmt. Hierzu speichert der Switch-Controller eine eindeutige Identifikationsnummer und die IP-Adresse des Nachbarn ab. Jeder Knoten schickt periodisch Nachrichten mit seiner Identifikationsnummer. Wechselt ein Nachbar oder wird die Verbindung zu diesem unterbrochen, werden alle Flows, die über diese Verbindung geswitcht wurden, gelöscht.
Nach wie vor kann es zu Inkonsistenzen bei den Verbindungszuständen kommen. Um die Existenz solcher Inkonsistenzen zu vermeiden bzw. zu verkürzen, beinhalten alle Redirect-Nachrichten eine Gültigkeitsdauer. Diese beträgt üblicherweise eine bis zwei Minuten, was ungefähr einem Datenfluß im Internet entspricht. Nach Ablauf dieser Zeit unterläßt es der Switch, den Flow auf die entsprechende Kennung zu switchen. Der IP-Switch-Controller überwacht jeden Flow. Werden während einer Periode Daten empfangen, sendet der Controller eine weitere Redirekt-Nachricht, um die Gültigkeit des Flows zu verlängern. Andernfalls schickt der Controller eine Reclaim-Nachricht an den stromaufwärtsliegenden Switch, um die verwendete Kennung zurückzufordern. Der Flow wird erst gelöscht, wenn diese Nachricht bestätigt wurde. Für Flows, die eine Kennung haben, aber nicht geswitcht werden, kann der IP-Controller die Gültigkeit bzw. Ankunft von Daten auf einem Flow selbst prüfen. Für geswitchte Flows muß er eine Anfrage an die ATM-Hardware stellen, ob die Flows kürzlich aktiv waren.

Bei dem Ausfall eines Routers oder einer Verbindung wird über das dynamische Routing ein neuer Weg berechnet. Alle Flows, die über die alte Verbindung liefen, werden ungültig definiert und gelöscht. Der betroffene Datenverkehr wird wieder an den

IP-Switch-Controller weitergeleitet und über die Default-Verbindungen mittels Store-and-Forward geroutet. Ab der Fehlerstelle werden neue virtuelle Verbindungen über den neu berechneten Weg aufgebaut. Der Fehler wird lokal behoben. Es ist nicht notwendig, alle virtuellen Verbindungen abzubauen, wie dies bei verbindungsorientieren Netzwerken der Fall ist. Diese Fehlerbehebung ist wesentlich effizienter.

Während die Routingprotokolle die Routingtabellen aufstellen, kann es zeitweilig zur Schleifenbildung kommen. Ein geswitchter Flow, der sich aufbaut, während eine solche Schleife innerhalb des Routings besteht, wird einen Switch mehrmals passieren und jeweils eine weitere Kennung (VCI) belegen. Dabei wird das TTL-Feld kontinuierlich heruntergesetzt. Erreicht dieses den Wert Null, wird der Flow nicht weiter geswitcht. Wenn das Routing einen konsistenten Zustand hergestellt hat, lösen sich auch die geswitchten Schleifen auf.
Eine Möglichkeit für den Switch, Schleifen zu erkennen, existiert nicht, da hierzu ein Switch nicht nur lokale sondern globale Informationen über eine geswitchte Verbindung haben müßte. Es ist auch nicht möglich, anhand des TTL-Feldes eine Schleife zu dedektieren, da durchaus Situationen denkbar sind, in denen sich zwei Flows nur durch das TTL-Feld unterscheiden.

4 General Switch Management Protokoll

In den vorherigen Abschnitten wurde die Arbeitsweise des IP-Switchings vorgestellt. Im wesentlichen wird dieses durch das IFMP-Protokoll festgelegt. Im nun folgenden Abschitt soll näher auf die Steuerung des Switches durch den Switch-Controller eingegangen werden. Hierzu wurde von Ipsilon das GSMP-Protokoll [NEHH+96a] entworfen. Ziel beim Entwurf dieses Protokolls war es, eine möglichst hardwarenahe Möglichkeit zu schaffen, alle wesentlichen Typen von ATM-Switches über ein Protokoll steuern zu können. GSMP erlaubt es dem Controller, Verbindungen zu benachbarten Switches auf- und abzubauen, Äste zu Punkt-zu-Mehrpunkt Verbindungen hinzuzufügen und zu entfernen, die Ports des Switches zu konfigurieren, Konfigurationsinformationen zu erfragen und Statistiken vom Switch anzufordern. Weiterhin erlaubt es dem Switch, den Controller von außergewöhnlichen Ereignissen wie dem Zusammenbrechen einer Verbindung zu informieren.

GSMP ist ein einfaches Master/Slave-Protokoll. Der Master (Switch-Controller) sendet Anfragen bzw. Befehle an den Switch und erwartet eine positive oder negative Antwort. Alle GSMP Nachrichten werden bestätigt.
Die Verbindung zwischen Switch und Controller ist aus Einfachheits- und Geschwindigkeitsgründen nicht gesichert (Man kann davon ausgehen, daß diese Verbindung entweder sehr zuverlässig oder defekt ist, so daß eine Sicherung nicht sinnvoll ist.). Das GSMP-Protokoll läuft auf einem einzigen ausgezeichneten virtuellen Kanal des Switches. Alle Nachrichten benutzen die Einkapselung der AAL-5-Schicht von ATM. Die meisten Narchrichten sind so klein, daß sie innerhalb einer einzelnen Zelle übertragen werden können (siehe Abbildung 3).
Ein einzelner Controller kann mehrere Switches kontrollieren, indem er mehrere Protokollinstanzen benutzt und jeweils eine separate Kontroll-Verbindung zu diesen aufbaut. Die LCC/SNAP Kapselung wird verwendet, um auch anderen Protokollen neben GSMP zu ermöglichen, Daten mit einem anderen SNAP-Typ auf das Netz zu multiplexen. Während GSMP zum Beispiel nur sehr einfache Netzwerkmanagement-Funktionen

beinhaltet, wird das Simple Network Management Protokoll (SNMP) zwischen Switch und Controller benötigt, um volle Netzwerkmanagement-Funktionalität zu haben.

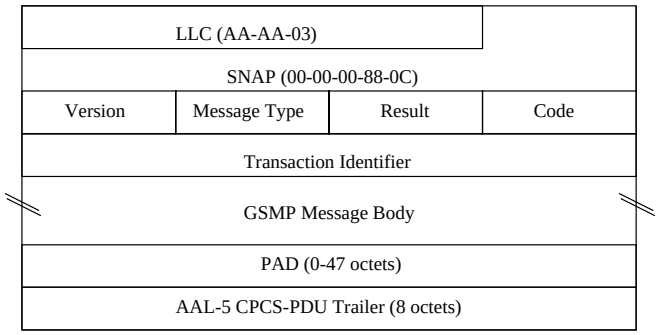

Abbildung 3: GSMP-Paketformat

GSMP definiert vier Klassen von Nachrichten bzw. Anforderungen: Verbindungsmanagement, Port-Management, Statistik und Konfiguration. Alle GSMP Nachrichten verwenden das in Abbildung 3 dargestellte Format, wobei sich der Messagebody je nach Klasse ändert.

5 Unterstützung von Dienstqualitäten - qGSMP

Mit der Definition von Dienstqualitäten wird eine bewußte Bevorzugung von Flows, daß heißt eine schnellere Verarbeitung in den Switches, ermöglicht. Einige Pakete werden schneller weitergeleitet entsprechend der vereinbarten Regeln. Mit RSVP wurde bereits ein Protokoll für die Unterstützung von Dienstqualtiäten für IP entwickelt. Es ist ein reines Signalisierungsprotokoll und setzt voraus, daß der Benutzer seine Anforderungen an Ressourcen genau spezifiziert. Der Transport der Daten erfolgt über einfache IP-Pakete. Das RSVP kontrolliert den Zuguang zu den Netzressourcen und reserviert Bandbreite für die Übertragung.
Die IP-Switchtes stellen den höheren Schichten, unabhängig von der Realisierung des Switches, die Ethernet Standards IEEE 802.2 und IEEE 802.3 zur Verfügung, so daß der Einsatz von RSVP über den IP-Switches möglich ist. RSVP bietet keine garantierte Dienste. IP-Switching läßt den Einsatz von RSVP zu, unterstützt dieses aber nicht direkt.

Im GSMP ist zur Unterstützung von Dienstqualtitäten ein Prioritätsfeld definiert. Es erlaubt, für jeden geswitchten Flow eine Priorität festzulegen. Diese Priorität wird am Ausgangsport ausgewertet, wenn auf diesem mehrere Flows den Switch verlassen. Die Zellen eines Flows mit einer höheren Priorität sollten den Switch früher verlassen als die Zellen eines Flows mit einer niedrigeren Priorität. Qualitätsparameter wie max. Durchsatz bzw. max. Zellrate werden nicht unterstützt. Zu diesem Zweck wurde eine Erweiterung des GSMP vorgeschlagen, das qGSMP-Protokoll [AdLN97].

Das qGSMP erweitert das GSMP um die in Abbildung 4 dargestellten Nachrichtenklassen und setzt beim ATM-Switch drei weitere Funktionseinheiten voraus (Abbildung 4): Scheduler, Buffermanager, Schedulable Region Estimator (SRE). Mittels der qGSMP-Nachrichten kann der Scheduling-Algorithmus, die Pufferverwaltung und der Algorithmus zur Bestimmung der Schedulable Region (SR) ausgewählt werden, die

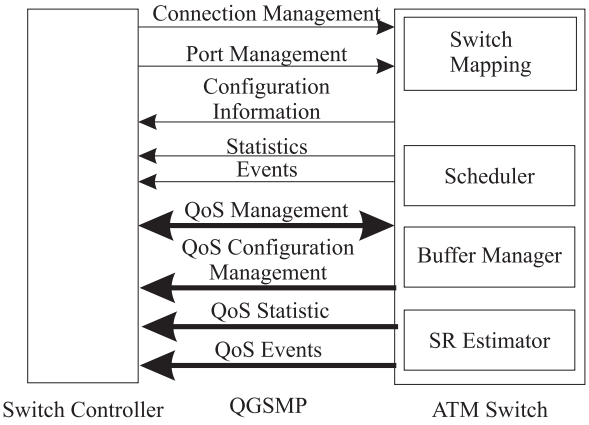

Abbildung 4: Erweiterung von GSMP - Das qGSMP-Protokoll

Verkehrsparameter und die QoS-Beschränkungen gesetzt werden und Informationen, die für das QoS-Management entscheidend sind, gesammelt werden. Dabei besitzen die Nachrichten dasselbe Format wie bei GSMP.

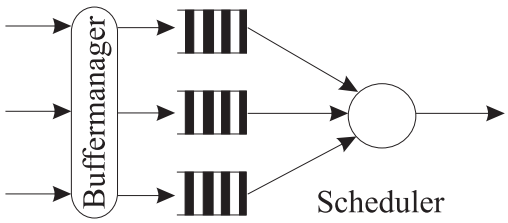

Abbildung 5: Modell eines Ausgangsports als Multiplexer

Jeder Ausgangsport eines Switches ist als ein Multiplexer aufgebaut. Abbildung 5 stellt das Modell eines Ausgangsport dar. Die Pufferverwaltung steuert den Zugriff auf die Puffer eines Ausgangsports und der Scheduler den Zugriff auf die Ausgangsverbindung. Jeder Ausgangspuffer repräsentiert eine Flowklasse. Diese Flowklassen sind innerhalb des zu administrierenden Netzwerks definiert und werden dem Benutzer zur Verfügung gestellt. Eine Flowklasse definiert ein statistisches Modell eines Informationsflußes, das die max. Zellrate und weitere charakteristische Parameter (z.B. von Video, Sprache und Daten) enthält.

Jeder IP-Switch trifft für jeden zu switchenden Flow die Entscheidung, welcher Flowklasse er angehören soll. Dabei orientiert er sich an den in der Domain vorgegebenen Regeln. Diese Entscheidung kann er anhand der Anwendung (erkennbar durch die TCP/UDP/RTP-Portnummer), Typ-of-Service Feldes, der IP-Adressen usw. treffen. Beachten muß der Switch dazu aber auch die Fähigkeiten der verwendeten ATM-Switch-Hardware und die QoS-Forderungen auf entsprechend angebotene Flows abbilden. Der SR-Estimator bestimmt für jeden Ausgangsport die Anzahl der Flows pro Klasse, die über den Port geswitcht werden können, ohne die garantierten Qualitätparameter zu verletzen. Diese Information kann dann über den Switch-Controller abgerufen werden.

Dieser Ansatz bietet die Möglichkeit, Flows verschiedener Dienstgüten zu unterstützen. Garantierte Dienste mit Dienstgüteparametern können aber nicht geleistet werden, da die Entscheidung, ob ein Flow geswitcht wird und welcher Flowklasse er angehört, lokal getroffen wird.

6 Bewertung

IP-Switching von Ipsilon ist noch kein Standard und es stehen noch weitere Verfahren zur Debatte. Allerdings hat Ipsilon mit den RFCs zu GSMP und IFMP bereits einen Schritt in Richtung Standardisierung gemacht. Alle hier vorgestellten Eigenschaften sind noch sehr theoretisch. Implementierungen, die die gesamte Funktionalität beinhalten, existieren noch nicht.

6.1 Leistung und Durchsatz

Zur Leistungsanalyse wurden ca. 10 minütige Aufzeichnungen aus dem Internet herangezogen. Die beiden bereits in Abschnitt 3.2 vorgestellten Flow-Klassifikatoren wurden auf diese Aufzeichnungen angewendet.
Ausschlaggebend bei der Bewertung ist im wesentlichen das Verhältnis von geswitchten Paketen zu der Anzahl bestehener Flows. Diese gibt an, wie lang im Durchschnitt die Flows sind, die geswitcht werden.
Je mehr Flows unterhalten werden, desto höher ist der Datenverkehr, der zum Aufbau und zur Erhaltung der Flows benötigt wird. Dieser kann bei zu vielen Flows die Leistung stark drücken. Wie aus Tabelle 1 ersichtlich ist, wird bereits bei der Klassifikation des Datenverkehrs nach Anwendungen (Ports) bekannter Protokolle (ftp, telnet, gopher, http, nntp, netbios, login, cmd, audio, AOL, X-11) ein deutlich besseres Verhältnis zwischen switchtbaren Paketen und Anzahl der Flows erreicht, als beim zweiten Flow-Klassifikator.
Weitere Messungen haben gezeigt, daß im Schnitt 4 Pakete für den Aufbau eines jeden Flows benötigt werden und 2 Pakete für die periodische Redirect-Nachricht pro Erneuerungsintervall (die Simulation verwendete 20 Sekunden). Für diese Simulation ergibt sich daraus ein zusätzlicher Management-Datenverkehr von 1918 Paketen/s. Ingesamt ergibt sich eine Verarbeitungsleistung von ca. 16.700 Paketen/s, wobei bei einem Store-and-Foward-Routing gerade einmal 4.518 Pakete/s gemessen wurden. Hieraus resultiert eine Leistungssteigerung mit dem Faktor 3,7.
Eine weitere Versuchsreihe, bei der als Flow-Klassfikator die Anzahl der Pakete pro Zeitintervall benutzt wurde (10 Pakete in 60 Sekunden), ergab, daß 86 Prozent der Pakete geswitcht wurden. Durchschnittlich wurden 118 Flows/s aufgebaut und 18.000 Verbindungen verwaltet. Damit ergeben sich annähernd dieselben Werte wie beim ersten Klassifikator.

Mit der statistischen Flow-Klassifikation ist es einfach, je nach Leistungsfähigkeit des Switches das Verhältnis zwischen den geswitchten und gerouteten Paketen zu variieren, indem die Grenze, ab der ein Flow geswitcht wird, verändert wird. Damit läßt sich die Anzahl der Verbindungen, die pro Sekunde aufgebaut werden, und die Größe der Verbindungstabelle bestimmen. Die statistische Flow-Klassifikation ist somit flexibler und wird auch Datenverkehr gerecht, der nicht einer Anwendung zugeordnet werden kann. Ingesamt ergibt sich beim IP-Switching im Durchschnitt eine Durchsatzsteigerung mit einem Faktor von 3,5.

Gegegebenheiten	1	2
Flow Klassifikator	bekannte Protokolle und Ports, alle TCP-Pakete, keine UDP-Pakete	alle Pakete
Flow Typ	Typ 2	Typ 2
Durchschnittliche Paketanzahl/s	16.700	16.700
Messungen		
Switchbare Pakete	84 Prozent	100 Prozent
Durchschnittliche Flow-Anzahl	92 Flows/s	422 Flows/s
Einträge Flow-Tabelle	15.500	42.000
Geswitchte Pakete	14.100	
Geroutete Pakete	2.600	

Tabelle 1: Leistungsanalyse IP-Switching

6.2 Skalierbarkeit

IP-Switching versucht nicht wie z.B. IP-over-ATM [AhTe94] oder LAN-Emulation [ATM-97] die Eigenschaften von LANs nachzubilden (gemeinsames Medium), sondern setzt auf ein Punkt-zu-Punkt Verbindungskonzept auf, was dem ATM-Modell entspricht. Alle Routing-Protokolle können auf dieser Grundlage arbeiten. Das Internet hat gezeigt, daß IP sehr gut skalierbar ist. Dadurch, daß die Entscheidung, ob ein Flow geswitcht oder geroutet wird, lokal getroffen wird, ist es möglich, das Switching auch auf große Netzwerke auszudehnen. Die Switch-Entscheidung für eine Verbindung hat keinen Einfluß auf das restliche Netzwerk.

6.3 Integration

Das hier vorgestellte Konzept für Switches dient für den Einsatz im Backbone-Bereich. Aufgrund der Punkt-zu-Punkt Topologie ist es nötig, für den Anschluß lokaler Netze sogenannte IP-Switch-Gateways einzusetzen. Über diese werden LANs an das IP-Switch-Netzwerk gekoppelt. Die Gateways beinhalten im wesentlichen einen IP-Protokollturm auf Seite es lokalen Netzes und den in Abbildung 6 dargestellten IP-Switch-Protokollturm. Die Integration von IP-Switch-Teilnetzen geschieht völlig transparent.

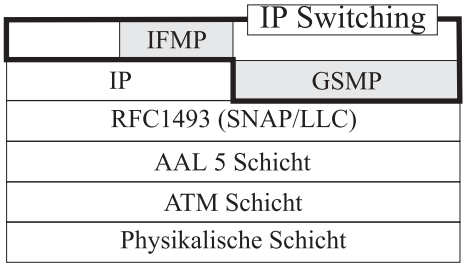

Abbildung 6: Protokollturm

6.4 Unterstützte Dienste

Ein IP-Switch unterstützt IP-Multicast, ohne das Änderungen am Internet Group Management Protokoll (IGMP) oder Multicast Routing Protokoll vorgenommen werden

müssen. Das Flow-Switching funktioniert für Mulitcast-Verbindungen genauso wie für Unicast-Verbindungen. Empfängt ein IP-Switch einen Multicast-Flow, leitet er die Pakete auf die entsprechenden Ausgangsports. Dies funktoniert zunächst über das IP-Multicast-Protokoll im Controller. Wenn dieser Flow eine Kennung zugewiesen bekommen hat, kann dieser Flow mit Hilfe der Multicastfähigkeit der ATM Hardware geswitcht werden, wenn die stromabwärtsliegenden Empfänger den Flow mittels einer Redirection-Nachricht umleiten. Gleichzeitig kann der Switch eine Kopie des Multicast-Flows an den Switch-Controller schicken, falls Knoten existieren, die den Flow nicht umleiten. Diese empfangen die Multicast Pakete dann über die Default-Verbindung.

7 Zusammenfassung

Klassische Router bieten eine ganze Reihe wichtiger Funktionen: Schutz durch Firewalls, Broadcast-Kontrolle, Kontrolle der Netzwerkauslastung, Einrichtung von besonders sichere Arbeitsgruppen. Entsprechend ergeben sich auch eine Reihe von Nachteilen wie z.B die hohen Kosten von Routern, große Latenzzeiten, relativ schwieriger Umbau von Netzen und komplexe Topologien. IP-Switches bieten hier eine leistungsstarke und preisgünstige Alternative. Dennoch ist gerade aufgrund der oben beschriebenen Funktionalität von Routern nicht davon auszugehen, daß diese vollständig verschwinden. Sie werden an den Rand des Netzwerkes wandern; IP-Switches werden im Backbone-Bereich von Netzen eingesetzt werden.

Literatur

[AdLN97] Constantin M. Adam, Aurel A. Lazar und Mahesan Nandikesan. QOS Extensions to GSMP. Draft, University New York, April 1997.

[AhTe94] M. Ahmed und K. Tesink. Definition of Managed Objects for ATM Management Version 8.0 using SMIv2. RFC 1695, August 1994.

[ATM-97] The ATM-Forum. LAN Emulation Over ATM Version 2 - LUNI Specification. af-lane-0084.000, Juli 1997.

[Detk98] Kai-Oliver Detken. ATM mit IP im Bunde. *Gateway*, Februar 1998, S. 92–96.

[Kuri98] Jürgen Kuri. Beziehungskiste. *c't* Band 6, 1998, S. 354–361.

[NEHH+96a] P. Newman, W. Edwards, R. Hinden, E. Hoffman, F. Ching Liaw, T. Lyon und G. Minshall. General Switch Management Protokoll. Rfc 1987, Ipsilon Networks, August 1996.

[NEHH+96b] P. Newman, W. Edwards, R. Hinden, E. Hoffman, F. Ching Liaw, T. Lyon und G. Minshall. Ipsilons Flow Management Protkoll. Rfc 1953, Ipsilon Networks, Mai 1996.

[NeML97] Peter Newmann, Greg Minshall und Tom Lyon. IP Switching: ATM Under IP. Draft, Ipsilon Networks, 1997.

[Neuk97] Thomas Neukam. Layer 3 Switching mit MPOA - Zwei in einem. *Gateway*, Dezember 1997, S. 122–125.

[NMLH97] Peter Newmann, Greg Minshall, Tom Lyon und Larry Huston. IP Switching and Gigabit Routers. Technischer Bericht, Ipsilon Networks Inc., 1997.

[Rede97a] Bernd Reder. Layer 3 Switching - Zwitterwesen. *Gateway*, Dezember 1997, S. 74–77.

[Rede97b] Bernd Reder. Routing and IP-Switching. *Gateway*, März 1997, S. 124–128.

[Rede97c] Bernd Reder. Routing und IP Switching - Kampf um das richtige Konzept. *Gateway*, März 1997, S. 149–151.

Vergleich aktueller Key-Management-Protokolle für die IP-Sicherheitsarchitektur

Andreas Wachowski

Kurzfassung

Seit Erscheinen der IP-Sicherheitsarchitektur Mitte der neunziger Jahre ist die Grundlage für eine sichere Datenübertragung im Internet gelegt. Die Realisierung des Schlüsselmanagements wurde jedoch noch nicht geklärt. SKIP, Photuris und IKE sind aktuelle Entwürfe für Schlüsselmanagement-Protokolle, die für die IPSec-Sicherheitsarchitektur spezifiziert werden. Der Artikel gibt einen Überblick über diese Protokolle und vergleicht sie anhand der Eigenschaften Schlüsselerstellung, Authentifizierung, Perfect Forward Secrecy und Nachrichtenanzahl.

1 Einleitung

Je mehr das Internet an Verbreitung gewinnt, desto dringender wird der Bedarf an effektiven Sicherheitsmechanismen, die einen vertraulichen Datenaustausch über die IP-Schicht ermöglichen. Eine Untergruppe der IETF (Internet Engineering Task Force), die IPSEC (IP Security Working Group) beschäftigt sich damit, entsprechende Standards zu entwerfen.

Der erste Schritt in diese Richtung wurde 1995 mit der Veröffentlichung einer allgemeinen IP-Sicherheitsarchitektur gemacht (RFC 1825-27 [Atki95c] [Atki95a] [Atki95b]). Die jeweiligen Dokumente beschreiben Sicherheitsmechanismen sowohl für IPv4 als auch IPv6. Die IP-Sicherheitsarchitektur, oder IPSEC, sieht allerdings kein spezifisches Schlüsselmanagement vor. Statt dessen werden grobe Rahmenbedingungen vorgegeben, die ein IPSEC-konformes Schlüsselmanagement erfüllen soll.

Daraufhin wurden mehrere Vorschläge für Schlüsselmanagement-Protokolle (Key Management Protocols) entwickelt. Auf den folgenden Seiten werden die bestehenden Ansätze dargestellt und verglichen. Im einzelnen geht es dabei um das "Simple Key-Management Protocol for Internet Protocols" (SKIP), "Photuris" und den "Internet Key Exchange" (IKE) auf Basis des "Internet Security Association and Key Management Protocol" (ISAKMP).

Bevor die Protokolle im einzelnen vorgestellt werden, soll noch ein kurzer Überblick zu der IPSec-Architektur und eine Einführung ins Schlüsselmanagement gegeben werden.

1.1 IPSec

Grundlage von IPSec sind zwei verschiedene Header (1, 2). Der Authentication Header (AH) garantiert die Echtheit und Integrität des IP-Datagramms, der Encapsulating Security Payload (ESP) enthält die verschlüsselten Daten inklusive der zur Verschlüsselung notwendigen Information.

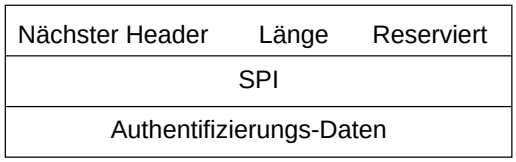

Abbildung 1: Der IPSec Authentication Header.

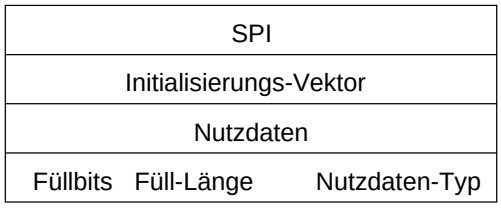

Abbildung 2: Der IPSec Encapsulating Security Payload.

Der *Security Parameters Index* (SPI) verweist auf eine *Security Association* (SA). Die SA wird zwischen zwei kommunizierenden Instanzen vereinbart und enthält u. a. die verwendeten kryptographischen Schlüssel und Angaben zum Chiffrieralgorithmus. Die SAs selbst existieren in einer lokalen Datenbank beim Empfänger bzw. Sender.

Beim Empfang eines Pakets wird mit dem AH-Header die Echtheit des Senders überprüft, danach entschlüsselt man die im ESP enthaltene Nutzlast. Eine Frage des Schlüsselmanagements ist es nun, wie die für AH und ESP benötigten Schlüssel vereinbart werden.

1.2 Schlüsselmanagement

Um das Wesen des Schlüsselmanagements zu erfassen, ist es günstig, sich den Lebenszyklus eines Schlüssels näher anzusehen. Dieser kann in die Bereiche 1) Erzeugung, 2) Verteilung, 3) Aktivierung/Deaktivierung, 4) das Ersetzen bzw. ein Update, 5) die Rücknahme und 6) die Terminierung eines Schlüssels gegliedert werden (nach [Ford94, Abschnitt 4.5]). Schlüsselmanagement-Protokolle realisieren die Phasen 1) und 2). Im folgenden werden diese Phasen und ihre Eigenschaften detaillierter beschrieben.

Die *Schlüsselerzeugung* basiert in der Regel auf Zufallszahlen. Die Wichtigkeit der Zufallszahlenqualität wird dabei oft unterschätzt; doch die Analyse eines mangelhaften Zufallszahlengenerators kann unter Umständen dazu führen, daß ein Angreifer aus einer aufgezeichneten Schlüsselfolge die nachfolgende Sequenz konstruieren kann. Am sichersten sind echte Zufallszahlen, also solche, die nicht durch Algorithmen, sondern durch physikalische Prozesse erzeugt werden. Rauschgeneratoren lassen sich zum Beispiel zu diesem Zweck einsetzen. Für mehr Informationen zur Zufallszahlenerzeugung siehe [ErCS94].

Die *Verteilung* von Schlüsseln ist abhängig vom eingesetzten Kryptosystem. Es existieren zwei Arten von Kryptosystemen, symmetrische und asymmetrische Systeme. Symmetrische Systeme benutzen für Ver- bzw. Entschlüsselung den gleichen Schlüssel, asymmetrische Systeme dagegen verfügen über einen öffentlichen Schlüssel zur Chiffrierung und einen geheimen zur Entschlüsselung.

Für symmetrische Systeme gilt die Forderung, daß nur die beiden Kommunizierenden den Schlüssel kennen. Ein geeigneter Algorithmus zur Erstellung eines gemeinsamen Geheimnisses ist der nach Diffie-Hellman (s. Abschnitt 1.2.1). Bei asymmetrischen Systemen ist der geheime Schlüssel nur dem Besitzer bekannt, der öffentliche Schlüssel dagegen muß für alle potentiellen Partner erreichbar sein. Insbesondere muß ein öffentlicher Schlüssel eindeutig seinem Besitzer zugeordnet werden können. Letzteres wird über *Zertifikate* ermöglicht; das sind in diesem Zusammenhang von einer vertraulichen Stelle unterschriebene Schlüssel.

Beim Abschnitt Schlüsselerzeugung wurde schon eine *Sequenz von Schlüsseln* statt eines einzelnen Schlüssels angesprochen. In der Tat sollte ein Schlüssel nicht zu lange eingesetzt werden, weder zeitlich noch bzgl. der verschlüsselten Datenmenge. Je mehr Material oder je mehr Zeit ein Angreifer zum Herausfinden des Schlüssels hat, desto wahrscheinlicher wird sein Erfolg. Deshalb werden Schlüssel in regelmässigen Intervallen erneuert.

Ein Schlüsselmanagement-Protokoll kann beide Arten von Kryptosystemen einsetzen. Meistens benutzt man unsymmetrische Verfahren wegen ihres Rechenaufwands nur für die Aufstellung einer sicheren Verbindung. Dabei wird ein gemeinsames Geheimnis berrechnet und daraus ein Langzeitschlüssel abgeleitet. Bei der Verschlüsselung der Datenübertragung werden statt dessen wechselnde Schlüssel eines symmetrischen Systems benutzt. Die symmetrischen Schlüssel selbst, und nur sie, werden mithilfe des Langzeitschlüssels codiert, damit bei einem Angriff auf den Langzeitschlüssel nicht zuviel Chiffretext zur Verfügung steht. Protokolle können auch ein Update des Langzeitschlüssels vorsehen.

1.2.1 Schlüsselerzeugung nach Diffie-Hellman

Die Diffie-Hellman-Methode ist ein weit verbreiteter Algorithmus zur Erzeugung eines gemeinsamen Geheimnisses zwischen zwei oder mehr Parteien. Die beiden potentiellen Kommunikationspartner A und B einigen sich zunächst auf eine Primzahl p und einen Wert g, welcher modulo p primitiv ist (d. h., g ist ein erzeugendes Element der multiplikativen Gruppe 0,1,...,p-1). Diese Kommunikation kann ungesichert erfolgen. Anschließend wählen sowohl A und B jeweils eine große zufällige Zahl a bzw. b (die geheimen Schlüssel) und senden ihrem Gegenüber nun $g^a \, mod \, p$ bzw. $g^b \, mod \, p$ (die öffentlichen Schlüssel). Mit dem öffentlichen Schlüssel als Basis und dem geheimen als Exponenten erfolgt eine weitere Potenzierung mod p. So bekommen beide Partner das gemeinsame Geheimnis $g^{ab} \, mod \, p = g^{ba} \, mod \, p$.

Die Bestimmung von a oder b ist für einen Angreifer aufgrund der nötigen Berechnung des diskreten Logarithmus schwer, vorausgesetzt p und g sind geeignet gewählt. Die Größe von p ist dabei von besonderer Bedeutung, denn sie bestimmt die Größe des Geheimnisses. Momentan werden häufig 512-Bit-Primzahlen benutzt, empfehlenswert sind mindestens 1024 Bit.

Diffie-Hellman wird sinnvollerweise immer zusammen mit einer Authentifizierung eingesetzt. Wäre sich der Initiator einer Verbindung nicht eindeutig sicher, daß er den gewünschten Empfänger anspricht, könnte ein Angreifer selbst als Empfänger auftreten und so unbemerkt sensible Daten abfangen ("Man in the Middle Angriff"). Für Protokolle, die Diffie-Hellman benutzen, ist daher auch wichtig, einen guten Authentifizierungsmechanismus zu besitzen.

1.2.2 Einweg-Hashfunktionen

Ein weiteres wichtiges Element der Kryptographie und damit der Schlüsselmanagement-Protokolle sind Einweg-Hashfunktionen. Eine *Einweg-Hashfunktion* ist eine Funktion h, die für eine (beliebig lange) Eingabe E einen Funktionswert (Hashwert) H fester Länge erzeugt, und die zusätzlich die folgenden Eigenschaften besitzt:

a) Die Berechnung von H=h(E) ist leicht
b) Die umgekehrte Bestimmung von E mittels h und H ist schwer
c) Die Ermittlung einer Eingabe E' bei gegebenem h und E, so daß h(E')=h(E) gilt, ist schwer.

Eine *starke Einweg-Hashfunktion* erfüllt zusätzlich die Eigenschaft der Kollisionsfreiheit, d. h., es gibt keine E und E' mit h(E')=h(E). Authentifizierung ist ein Bereich, bei dem starke Einweg-Hashfunktionen gefordert sind.

Einweg-Hashfunktionen werden zur Generierung von Schlüsseln und zur Authentifizierung eingesetzt. Beispiel-Funktionen sind MD5 [Rive92] und SHA [oSTe94] bzw. – für Authentifizierung – "Keyed MD5" (RFC 1828) bzw. "Keyed SHA" (RFC 1852) (s. auch [KrBC97]).

Im folgenden werden die drei Schlüsselmanagement-Protokolle SKIP, Photuris und IKE vorgestellt.

2 SKIP

SKIP (Simple Key-Management for Internet Protocols, [AzMP96]) ist ein bereits 1995 von SUN entwickeltes Protokoll. Es wurde im August 1996 als Internet-Entwurf vorgelegt, aber diese Bemühungen wurden anscheinend nicht fortgesetzt; es gibt keinen neueren Entwurf oder einen Standard-RFC. Auf *http://skip.incog.com/* findet sich aktuelles Material, darunter auch eine Spezifikation (*http://skip.incog.com/spec/SKIP.html*).

Ein Protokoll-Überblick soll zunächst die Grundlagen vermitteln, danach werden einige Eigenschaften von SKIP beschrieben.

2.1 Protokoll-Überblick

SKIP basiert auf der Annahme, daß jeder Teilnehmer-Knoten einen authentifizierten öffentlichen Diffie-Hellman-Schlüssel besitzt. Die Authentifizierung ist über vorhandene Mechanismen wie z. B. X.509 Zertifikate möglich. Die Verteilung der öffentlichen Schlüssel geschieht mittels eines Verzeichnisdienstes oder dem "Certificate Discovery

Protocol". Damit eine Verständigung über Diffie-Hellman überhaupt möglich ist, müssen kommunizierende SKIP-Implementierungen die gleichen Basen g und Primzahlen p (bzw. Gruppen) benutzen. In der Spezifikation von SKIP sind diese Werte vorgeschrieben.

Mit den existierenden Diffie-Hellman-Schlüsseln können beliebige Teilnehmerpaare symmetrische, geheime Schlüssel $g^{ij} \, mod \, p$ erzeugen. $g^{ij} \, mod \, p$ ist das sog. Langzeitgeheimnis und hat eine Länge zwischen 512 und 1024 bits, bei Bedarf auch mehr. Daraus erstellt SKIP einen Schlüssel K_{ij}, welcher Eingabe für ein beliebiges symmetrisches Schlüssel-Kryptosystem, etwa DES, RC2 oder IDEA, ist. Die Länge von K_{ij} hängt vom konkret benutzten Kryptosystem ab und bewegt sich typischerweise zwischen 40 und 256 bits. K_{ij} entsteht aus $g^{ij} \, mod \, p$ einfach durch Kopieren einer ausreichenden Anzahl niederwertiger Bits. Mit K_{ij} als Langzeitschlüssel werden die sog. Paketschlüssel K_p chiffriert.

Die Paketschlüssel K_p *(packet key)* dienen als Träger der beiden Schlüssel zur Authentifizierung und Verschlüsselung einzelner IP-Pakete. Die Paketschlüssel werden zufällig erzeugt. Den Authentifizierungs-Schlüssel K_a und den Chiffrier-Schlüssel K_e gewinnt man aus K_p durch Berechnung einer Einweg-Hashfunktion h:

$K_a = \mathrm{h}(K_p \mid \text{Crypt Alg} \mid 02\mathrm{h}) \mid \mathrm{h}(K_p \mid \text{Crypt Alg} \mid 00\mathrm{h})$
$K_e = \mathrm{h}(K_p \mid \text{MAC Alg} \mid 03\mathrm{h}) \mid \mathrm{h}(K_p \mid \text{MAC Alg} \mid 01\mathrm{h})$

"Crypt Alg" kennzeichnet den Verschlüsselungsalgorithmus, "MAC Alg" den Authentifizierungsalgorithmus. Alle zwei Minuten oder nach 10 MByte übertragenen Daten (der jeweils zuerst eintretende Fall zählt) sollte K_p erneuert werden.

Zur Erhöhung der Sicherheit für den Fall, daß K_{ij} in die Hände eines Angreifers gerät, wird der Langzeitschlüssel in regelmäßigen Intervallen erneuert. SKIP berechnet den neuen K_{ij} mittels eines Zählers n, welcher im SKIP-Header mitgeschickt wird. Der Zähler ist fortlaufend und eine 32bit Zahl. Bei stündlicher Erhöhung wiederholt er sich alle $3,5 \cdot 10^9$ Jahre, also praktisch gar nicht. K_{ij} und n sind Eingaben für die gleiche Einweg-Hashfunktion, die auch zur Berechnung der K_a und K_e benutzt wird. Der Funktionswert ist der neue K_{ij}.

Die stündliche Inkrementierung des Zählers setzt eine – wenn auch grobe – Synchronisation der Teilnehmer voraus. Empfängt eine Partei ein Paket, welches einen um mehr als eins abweichenden Zähler n enthält, wird die Nachricht abgelehnt.

2.2 Protokoll-Eigenschaften

Der Einsatz von SKIP ist möglich für alle Protokolle, die Schlüsselmaterial benötigen. Dazu wird im "next header" Feld des SKIP-Headers auf dieses Protokoll verwiesen, und umgekehrt enthält der Header des folgenden Protokolls einen Verweis auf die SKIP-Schlüssel. Für die IPSec Header AH und ESP (Abschnitt 1.1) enthält das SPI-Feld dann den reservierten Eintrag "SKIP_SPI".

SKIP identifiziert Teilnehmer nicht anhand der IP-Adresse, sondern über sog. *Master Key IDs*.

Es kann eine Authentifizierung an einem Zwischenknoten erwünscht sein, etwa beim Einsatz einer Firewall. Dazu muß der Server die geheimen Schlüssel aller angeschlossenen bzw. an der SKIP-Kommunikation beteiligten Knoten erhalten. Empfängt er dann

ein Paket, berechnet er den K_{ij} der beiden kommunizierenden Parteien, entschlüsselt K_p und authentifiziert das Paket. Er könnte auch die Daten entschlüsseln, und damit ist es eine Vertrauensfrage, ob SKIP in dieser Art und Weise für persönlichen Datentransport eingesetzt werden soll.

SKIP ist sicher gegen Man in the Middle, Known Keys und Denial of Service Angriffe.

3 Photuris

Photuris wird seit 1994 von P. Karn und W. Simpson entwickelt. Die aktuelle Spezifikation [KaSi98] ist vom Februar 1998.

3.1 Protokoll-Überblick

Photuris kennt vier Hauptstadien zur Aufstellung einer Kommunikations-Verbindung:

1. *Cookie-Austausch:* Denial of Service Abwehr, Bekanntgabe möglicher Austausch-Schemata
2. *Werte-Austausch:* Auswahl eines Austausch-Schemas und Kommunikation zugehöriger Werte
3. *Identifikations-Austausch:* Verifikation, Authentifizierung, Berechnung des gemeinsamen Geheimnisses und der Sitzungsschlüssel
4. *Weitere Nachrichten:* Schlüssel-Updates

3.1.1 Cookie-Austausch

Der Cookie-Austausch von Photuris beinhaltet zwei Nachrichten und sieht folgendermaßen aus:

$\Rightarrow$ Cookie-Anfrage ($Cookie_I$, ...)
$\Leftarrow$ Cookie-Antwort ($Cookie_I$, $Cookie_R$, Angebotene Austausch-Schemata, ...)

Hier und in den folgenden Austauschen bezeichnet "$\Rightarrow$" eine Kommunikation vom Initiator zum Empfänger *(Responder)*, "$\Leftarrow$" die Kommunikation vom Empfänger zum Initiator.

Bei der Cookie-Anfrage teilt der Initiator der Kommunikation dem Partner seinen Cookie ($Cookie_I$) mit. Als Antwort erhält er seinen eigenen Cookie zur Verifizierung, einen entsprechenden Cookie des Kommunikationspartners ($Cookie_R$) und eine Liste von Austausch-Schemata. Die Cookies dienen der Identifikation der beiden Parteien untereinander. Im Werte-Austausch wird ein konkretes Austausch-Schema gewählt, welches dann zur Etablierung eines gemeinsamen Geheimnisses benutzt wird (s. 3.1.2). Im folgenden wird die Generierung und der Nutzen des Cookies näher beschrieben und kurz auf das Format der Austausch-Schemata-Liste eingegangen.

Der Cookie ist ein 256-Bit-Wert. Seine Erzeugung ist an drei Auflagen gebunden: a) Der Cookie muß von den kommunizierenden Parteien abhängen, um Angriffe mit einem geraubten Cookie, gesendet von anderer Stelle, abzufangen. b) Nur der Cookie-Erzeuger selbst darf in der Lage sein, von ihm verifizierbare Cookies zu generieren. Das wird durch die Benutzung lokaler, geheimer Information erreicht, etwa durch einen zufälligen Schlüssel. Wählt man diesen für jede weitere Verbindung neu, werden Replay-Attacken erschwert. c) Die Cookie-Generierung muß schnell sein, um Denial-of-Service-Angriffen entgegenzuwirken.

Die Autoren von Photuris empfehlen für die Generierung der Cookies die Benutzung einer kryptographischen Einweg-Hashfunktion h (z. B. MD5):

Initiator-Cookie = h(Sl_I, IP-Quelladr., IP-Zieladr., UDP-Quellport,UDP-Zielport)
Antwort-Cookie = h($Cookie_I$, Zähler, Sl_R, IP-Quelladr., IP-Zieladr., eigener UDP-Zielport, angebotene Austausch-Schemata)

Sl_I bzw. Sl_R sind die lokalen Geheimnisse von Initiator und Empfänger. Der Zähler im Antwort-Cookie dient der Verhinderung von Replay-Attacken.

Dem Initiator einer Kommunikation ist es insbesondere möglich, die bestehenden Verbindungen zu den Partnern zu überwachen. Dazu wird der Initiator-Cookie jedes eingehenden Datagramms mit allen Cookies verglichen, die vom Initiator für die aktuellen Verbindungen schon erzeugt wurden. (Entweder durch Neugenerierung des Cookies oder durch Vergleich mit einem zwischengespeicherten Cookie.) Ein Angreifer, der sich mittels gestohlener oder zufällig erzeugter Cookies in eine der bestehenden Verbindungen einschalten möchte, scheitert somit an der Verifizierung durch den Initiator.

Die *angebotenen Austausch-Schemata* sind eine nach Präferenz geordnete Liste der vom Antwortenden unterstützten Austausch-Schemata. Die Liste ist mindestens vier Byte lang. Jeder Eintrag ist ein 16-Bit-Zahlenwert, der für ein Austausch-Schema steht (s. 3.1.2).

3.1.2 Werte-Austausch

Der Werte-Austausch gliedert sich in eine Anforderung und eine Antwort:

⇒ Werte-Anforderung(Cookies, Schema-Wahl, Initiator-Austausch-Wert, ...)
⇐ Werte-Antwort(Cookies, Antwort-Austauschwert, ...)

Nach dem Cookie-Austausch erfolgt der *Werteaustausch*. Der Initiator wählt aus der beim Cookie-Austausch erhaltenen Liste ein Austausch-Schema (*Schema-Wahl)* und sendet dieses zusammen mit einem dazu passenden Wert zur Generierung des gemeinsamen Geheimnisses. Der Antwortende schickt dem Initiator einen analogen Wert. Die Berechnung des gemeinsamen Geheimnisses erfolgt auf beiden Seiten nach Erhalt der Austausch-Werte.

Ein Schema ist durch einen 16-bit-Wert spezifiziert. Die möglichen Werte gliedern sich wie folgt in Gruppen: 0 und 1 sind reserviert, 2 ist Standardwert (s. unten), 3 bis 255 stehen für "zukünftig wohlbekannte, veröffentlichte Schemata", 256–32767 sind für "hersteller-spezifische unveröffentlichte Schemata" vorgesehen, 32768–65535 schließlich

können für private Schemata benutzt werden. Die hersteller-spezifischen Werte müssen bei den Autoren reserviert werden, um Eindeutigkeit zu gewährleisten. Bei privaten Schemata ist eine Mehrfachnutzung von einzelnen Werten erlaubt. ([KaSi98, Abschnitt 9])

Der Standardwert steht allgemein für einen Diffie-Hellman-Austausch auf Basis einer endlichen Gruppe mit Primzahl p und Basis 2, der Schlüsselgenerierung über MD5, der Privacy-Methode "einfaches Masking" und der Validierungsmethode MD5-IPMAC. Die Implementierung dieses Werts ist zwingend.

3.1.3 Identifikations-Austausch

Beim Identifikationsaustausch werden die folgenden Nachrichten übertragen:

⇒ Identitäts-Anfrage (Cookies, SPI, SPILT, Identitätswahl, Identifikation, Verifikation, SPI-Attribute, ...)
⇐ Identitäts-Antwort (Cookies, SPI, SPILT, Identitätswahl, Identifikation, Verifikation, SPI-Attribute, ...)

Zunächst soll ein Überblick über den Nachrichtenaustausch, anschließend dann eine Erläuterung des Nachrichtenaufbaus gegeben werden.

Nachrichtenaustausch:
Anfrage und Antwort sind strukturell gleich. Die Nachrichten besitzen die gleichen Parameter. Zum Senden einer Nachricht treffen beide Seiten mehrere Vorbereitungen: Auswählen von SPI, SPILT (SPI-Lebenszeit), SPI-Attributen und eigener Identifikation, dann Berechnung der Verifikation zur Authentifizierung, schließlich Geheimhaltungsschlüsselerstellung ("privacy key generation") und Verschlüsselung der Nachricht ("masking").

Der Erhalt der Identitäts-Anfrage durch den Empfänger bzw. der Identitäts-Antwort durch den Initiator gestaltet sich so: Validierung u. a. der Cookies, der Identifikation, der Verifikation und der Attributwahl. Bei erfolgreicher Validierung Berechnung der Sitzungsschlüssel für beide Richtungen. Start der verschlüsselten bzw. authentifizierten Übertragung von Nutzdaten.

Nachrichtenaufbau:
Die *Verifikation* ist ein 128-Bit-Wert, der Ergebnis der Berechnung eines Hashwerts mit einem teilnehmerabhängigen Verifikationsschlüssel und Daten der Parteien ist. Zu den Daten gehören Cookies, Identitätswahl, Identifikation, Attribut-Werten, Austauschwerte von Initiator und Empfänger, usw.

Identitätswahl und Identifikation:
Photuris bietet zwei Optionen zur Identitätswahl, von denen der konkrete Wert für den Verifikationsschlüssel abhängt. Die erste Alternative ist *symmetrische Identifikation* und basiert auf einer seitens des Protokolladministrators vorkonfigurierten Liste der potentiellen Teilnehmer-Identitäten (Name, Site-Name, EMail etc.) samt zugehöriger symmetrischer Schlüssel. Vorgeschrieben ist eine Identitätslänge von min. 496 Bits und eine Schlüsselstärke von min. 64 Bits. Der Verifikationsschlüssel (für Identitäts- und Validitätsverifikation) ist für diesen Fall der MD5-Wert von der Konkatenation des symmetrischen Schlüssels mit dem gemeinsamen Geheimnis. Zur Berechnung der Sitzungsschlüssel wird der symmetrische Schlüssel als Generierungsschlüssel benutzt.

Die zweite Identitätswahl ist die an RFC 1828 angelehnte *Authentifizierung*. Der Verifikationsschlüssel besteht dann aus den 384 höchstwertigen Bits der mehrfach iterierten Schlüsselgenerierungsfunktion für den Sitzungsschlüssel. Dabei ist diese Funktion durch die Schemawahl spezifiziert. Als Eingabe erhält sie u. a. die Cookies und das gemeinsame Geheimnis.

Geheimhaltungsschlüssel-Generierung und Verschlüsselung:
Vor dem Senden werden sowohl Identitäts-Anfrage als auch -Antwort über die durch die Schema-Wahl festgelegte Geheimhaltungsmethode ("privacy-method") getarnt ("masked"). Beim einfachen Tarnen ("simple masking") wird die Nachricht durch eine Exklusiv-Oder-Verknüpfung mit dem Geheimhaltungsschlüssel chiffriert. Basis zur Berechnung dieses Schlüssels ist die in der Schemawahl festgelegte Schlüsselgenerierungsfunktion. Parameter für die Funktion sind u. a. die SPI-Austauschwerte, die Cookies und das gemeinsame Geheimnis.

Nach Abschluß des Identifikations-Austausches sind damit beide Parteien im Besitz von Sitzungsschlüsseln und haben sich auf eine Sicherheits-Assoziation geeinigt. Alle nachfolgend zu sendenden Datagramme können mithilfe der Sitzungsschlüssel authentifiziert bzw. verschlüsselt werden.

3.1.4 Weitere Nachrichten

Weitere Nachrichten zum periodischen Ändern der Sitzungsschlüssel oder zum Aufbau neuer oder revidierter Sicherheitsparameter werden bei Bedarf in dieser Phase ausgetauscht. Anhand der übertragenen Datenmenge oder über ein Zeitintervall werden automatische Updates der Sitzungsschlüssel durchgeführt. Diese Kommunikation ist wie der Identifikations-Austausch über den gemeinsamen Schlüssel codiert.

4 IKE

IKE (Internet Key Exchange [HaCa98]) ist mit dem letzten Entwurf vom Juni 1998 recht aktuell. Es ist eine Mischung aus zwei weiteren Schlüsselmanagement-Protokollen, Oakley und SKEME ([Orma96] und [Kraw]) und wird mit dem Ziel entworfen, ISAKMP-konform zu sein. Das von Forschern der NSA (National Security Agency) entwickelte *Internet Security Association and Key Management Protocol* (ISAKMP, [MSST98]) definiert einen Rahmen zur Verarbeitung von *Security Associations*. Die von ISAKMP definierten Nachrichtenformate ermöglichen den Transport von Schlüsseln und Authentifizierungsdaten unabhängig von der konkreten Schlüsselerstellungstechnik. Damit stellt ISAKMP einen Rahmen für die Schlüsselverwaltung dar, beschreibt aber selbst kein konkretes Verfahren. IKE ist ein entsprechendes Schlüsselverwaltungs-Verfahren und wird im folgenden erklärt.

4.1 IKE Protokoll-Überblick

IKE besitzt zwei Phasen, die unterschiedliche Zwecke erfüllen:

- Phase 1: Aufstellung eines sicheren, authentifizierten Kommunikationskanals (Aufbau einer ISAKMP Security Association).
  - Main Mode / Aggressive Mode
  - New Group Mode (optional)
- Phase 2 *(Quick Mode)*: Auffrischung des Schlüsselmaterials und Behandlung von nicht-ISAKMP-konformen Sicherheitsmechanismen.

4.2 IKE Phase 1

In Phase 1 kommt entweder der *Main Mode* oder der *Aggressive Mode* zur Ausführung. Der Main Mode entspricht dem ISAKMP "Identity Protect Exchange"; Aggressive Mode ist eine Realisierung des "Aggressive Exchange" von ISAKMP. Bei beiden werden die authentifizierten Schlüssel mittels eines Diffie-Hellman-Austausches (s. Abschnitt 1.2.1) erstellt.

Im *Main Mode* werden zuerst zwei Nachrichten zur Einigung auf das weitere Verfahren ausgetauscht. Anschließend erfolgt der Austausch von DH-Werten und weiteren für den Austausch wichtigen Attributen, und zuletzt wird die Authentifizierung der Partner untereinander vorgenommen. Die Authentifizierungsnachrichten werden mit dem vorher erstellten gemeinsamen Geheimnis verschlüsselt.

Der *Aggressive Mode* ist charakterisiert durch die gleichzeitige Schlüsselerstellung und Authentifizierung. Ein Vorteil ist die höhere Geschwindigkeit, ein Nachteil der Verzicht auf Identitätssicherung (die IDs werden unverschlüsselt übertragen, da noch kein gemeinsames Geheimnis erstellt ist). Zudem kann die DH-Gruppe nicht vereinbart werden.

Sowohl Main Mode als auch Aggressive Mode kennen vier *Authentifizierungsarten:* Digitale Signatur, zwei Authentifizierungs-Varianten über Public Key und Authentifizierung über einen separat (z. B. manuell) vereinbarten Schlüssel. Der zu benutzende Signatur-Algorithmus wird zwischen den Parteien ausgehandelt. Der Vorteil von Public Key über digitale Signatur ist die auch im Aggressive Mode mögliche Identitätssicherung. Dafür ist Public Key langsamer: Die erste Variante braucht zwei public-key-Verschlüsselungen und zwei private-key-Entschlüsselungen. Die zweite Variante wurde zu diesem Zweck geändert und kommt mit je einer dieser aufwendigen Ver- bzw. Entschlüsselungen aus.

Die Authentifizierung erfolgt über folgende Hashwerte:
$HASH_I$ = prf(SKEYID, $g^{xi} \mid g^{xr} \mid Cookie_I \mid Cookie_R \mid SAi_b \mid IDii_b$)
$HASH_R$ = prf(SKEYID, $g^{xr} \mid g^{xi} \mid Cookie_R \mid Cookie_I \mid SAi_b \mid IDir_b$)

Die prf *(pseudo random function)* ist, wenn nicht anders ausgehandelt, die HMAC Version ([KrBC97]) der verwendeten Einweg-Hashfunktion. "$_I$" bzw. "$_R$" stehen für Initiator und Empfänger. "g^{xi}" und "g^{xr}" sind die öffentlichen DH-Schlüssel vom Initiator bzw. Empfänger.

SKEYID ist ein Hashwert, dessen Berechnung von der Wahl der Authentifizierungsart abhängt:

Digitale Signatur: SKEYID = prf($Ni_b \mid Nr_b$, g^{xy})
Public Key: SKEYID = prf(hash($Ni_b \mid Nr_b$), $Cookie_I \mid Cookie_R$)
Separat vereinbarter Schlüssel: SKEYID = prf(separater Schlüssel, $Ni_b \mid Nr_b$)
(g^{xy} ist das gemeinsame DH-Geheimnis.)

Der *New Group Mode* ist zur eigentlichen Kommunikation nicht notwendig. Er ist eine Erweiterung von ISAKMP und wird im Anschluß an den Main Mode ggf. zur Definition privater Gruppen für den DH-Austausch eingesetzt. IKE kennt standardmäßig bereits vier Gruppen ("Oakley-Gruppen"), wovon zwei über einen endlichen Körper modulo einer Primzahl, die anderen zwei über elliptische Kurven definiert sind.

4.3 IKE Phase 2

Der Quick Mode ist unabhängig von ISAKMP spezifiziert. Er sorgt für die Erneuerung von Schlüsselmaterial während einer Kommunikation. Durch einen eingelagerten flüchtigen Diffie-Hellman-Austausch kann der Quick Mode *Perfect Forward Secrecy* bieten (s. Abschnitt 5.3 und [HaCa98, Abschnitt 5.5]).

5 Gegenüberstellung und Bewertung

Nach ISAKMP lassen sich Schlüsselaustauschprotokolle u. a. an den Kriterien Schlüsselerstellungsmethode, Authentifizierung, Symmetrie, Perfect Forward Secrecy und Back Traffic Propagation unterscheiden. Davon wird im folgenden auf Schlüsselerstellungsmethode, Authentifizierung und Perfect Forward Secrecy näher eingegangen und außerdem die Anzahl der auszutauschenden Nachrichten verglichen.

5.1 Schlüsselerstellung

SKIP, IKE und Photuris haben grundlegende Gemeinsamkeiten. Alle etablieren ein gemeinsames Geheimnis zwischen den Parteien, das zur Verschlüsselung der richtungsabhängigen, zufällig erzeugten Datagramm-Schlüssel eingesetzt wird. Diese Paket-Schlüssel sind Grundlage zur Authentifizierung und Chiffrierung von Nachrichten. Bei SKIP heißt das gemeinsame Geheimnis K_{ij} und der Paketschlüssel K_p. Bei Photuris entspricht der Werte-Austausch der Erstellung des gemeinsamen Geheimnisses, und der Identifikations-Austausch liefert u. a. die ersten Sitzungsschlüssel. Alle Protokolle setzen dabei primär auf Diffie-Hellman. Unterschiede gibt es beim Grad der Flexibilität bzgl. anderer Austausch-Schemata.

SKIP ist in dieser Beziehung nicht so allgemein gefaßt wie Photuris und IKE. Es setzt zur Generierung des gemeinsamen Geheimnisses einen Diffie-Hellman-Austausch mit expliziter Gruppe ein.[1] Der Nachteil dabei ist, daß eine Einigung bzgl. dieser Gruppe bereits auf Implementierungsebene getroffen werden muß. Im Entwurf werden folglich auch explizite Werte zur Implementierung empfohlen. Sollten in Zukunft aber gerade

[1] Die Spezifikation weist zwar auf die Möglichkeit der Verallgemeinerung hin. Das bedeutet aber eine Abänderung der Spezifikation.

für diese Werte mathematisch bedingte Schwächen entdeckt werden, dann steht man vor dem Problem, alle bestehenden Installationen von SKIP abändern zu müssen. Sobald nun zwei Implementierungen zwei verschiedene Gruppen oder Basen benutzen, können sie nicht mehr miteinander kommunizieren. Ein gleichzeitiges, globales Update ist aber genausowenig praktikabel. In dieser Hinsicht enthält SKIP also eine Einschränkung.

Photuris und IKE sehen allgemeiner die Einigung der Parteien auf eines von mehreren Austausch-Schemata vor. IKE bietet dabei Diffie-Hellman mit variablen Gruppen an[2]. Photuris geht den allgemeinsten Ansatz und spezifiziert Diffie-Hellman nur als eines von vielen einsetzbaren Austauschschemata. Es unterteilt dabei neben DH in a) "zukünftig wohlbekannte, veröffentlichte Schemata", b) "herstellerspezifische unveröffentlichte", und c) private Schemata.

Photuris kennt einen Cookie-Austausch, SKIP nicht. Der Cookie wird zur Initialisierung der Kommunikation eingesetzt, da sich die Partner zuerst auf ein Austausch-Schema einigen müssen.

5.2 Authentifizierung

Die Authentifizierung der den Protokollen zugrundeliegenden Diffie-Hellman-Werte ist auf sehr unterschiedliche Weise gelöst. SKIP setzt die Echtheit der öffentlichen Schlüssel voraus und verweist dabei auf bestehende Methoden zur Zertifizierung. Wie diese realisiert und an SKIP angebunden werden ist nicht Teil der Spezifikation von SKIP.

Photuris ist nicht notwendigerweise an einen DH-Austausch gebunden, sondern sieht Raum für viele Austausch-Schemata vor. Der Cookie-Austausch erreicht unabhängig vom Austausch-Schema eine Identifizierung der Kommunikationspartner für den weiteren Protokoll-Verlauf, bis der Identifikationsaustausch eine Authentifizierung der gesendeten Daten erreicht hat.

IKE bietet zur Authentifizierung des DH-Austauschs sowohl für Main Mode als auch Aggressive Mode vier Alternativen. Die erste, eine digitale Signatur, ist die von ISAKMP geforderte Authentifizierungsart. Außerdem kennt IKE zwei Möglichkeiten der public key Authentifizierung und die Alternative über einen separat vereinbarten Schlüssel.

Alle Protokolle bieten einen authentifizierten Schlüsselaustausch an. SKIP und Photuris verfolgen die in RFC 1826 [Atki95a] vorgestellte Idee: Die invarianten Teile des Datagramms werden inklusive des gemeinsamen Geheimnisses gehasht, und dieser Wert als MAC mit dem Datagramm geschickt ([AzMP96, Abschnitt 1.10] und [KaSi98, Abschnitt 1.2, 4.3]). IKE arbeitet im Quick Mode auf eine ähnliche Weise ([HaCa98, Abschnitt 5.5]).

5.3 Perfect Forward Secrecy

Der Begriff *Perfect Forward Secrecy* (PFS) "bezieht sich auf die Eigenschaft, daß die Kompromittierung genau eines Schlüssels nur Zugriff auf durch genau einen Schlüssel

[2]IKE verwirklicht damit nur einen Teil des ISAKMP-Rahmens, welcher einen allgemeinen, DH-unabhängigen "Key Exchange Payload" kennt [MSST98, Abschnitt 3.7]

gesicherte Daten erlaubt. Damit PFS möglich wird, darf der zur Datenübertragung benutzte Schlüssel nicht zur Ableitung weiterer Schlüssel benutzt werden, und wenn der zur Datenübertragung benutzte Schlüssel aus anderem Schlüsselmaterial abgeleitet wurde, dann darf dieses Material nicht benutzt werden, um weitere Schlüssel abzuleiten." [HaCa98, Abschnitt 3.3].

Die PFS-Eigenschaft ist demnach erwünscht. Falls ein Angreifer bei einem Protokoll ohne Perfect Forward Secrecy in den Besitz eines Schlüssels kommt wäre evtl. die Vorausberechnung zukünftiger Schlüssel möglich und damit die Sicherheit de facto zerstört.

Bei SKIP wird das gemeinsame Geheimnis zur Ableitung von Datenübertragungsschlüsseln K_{ij} benutzt. Die mit K_{ij} übertragenen Daten sind die Paketschlüssel K_p. Die stündlichen Aktualisierungen von K_{ij} erfolgen mittels einer Funktion, welche als Eingabe den alten K_{ij} und den aktuellen Zähler enthält. Sollte ein Angreifer in den Besitz des gemeinsamen Geheimnisses oder eines K_{ij}'s kommen, kann er unter Umständen sämtliche noch folgenden K_{ij} berechnen. SKIP ist also in der Original-Spezifikation nicht PFS-sicher. [Aziz97] ist eine Erweiterung, welche beschreibt, wie SKIP PFS-fähig gemacht werden kann.

Photuris besitzt Perfect Forward Secrecy. Nach der PFS-Definition darf aus einem zur Datenübertragung benutzten Schlüssel kein weiterer Schlüssel hergeleitet werden. Erst ab dem Identifikations-Austausch werden Nachrichten verschlüsselt übertragen, und dazu wird ein aus dem gemeinsamen Geheimnis hergeleiteter Schlüssel verwendet ("message-privacy key"). Die anderen Schlüssel, d. h. die zur Chiffrierung der Datenübertragung generierten Sitzungsschlüssel, basieren ebenfalls auf dem gemeinsamen Geheimnis und zusätzlicher weiterer Attribute. Ein *direkter* Zusammenhang zwischen diesen Schlüsseln besteht jedoch nicht, vielmehr sind beide Schlüssel vom gemeinsamen Geheimnis abhängig. Auch die Sitzungsschlüssel-Updates werden unabhängig von vorigen Schlüsseln generiert. Es gibt damit keinen direkten Zusammenhang zwischen den verwendeten Schlüsseln und somit Perfect Forward Secrecy.

IKE besitzt Perfect Forward Secrecy. Beim Update von Schlüsseln über den Quick Mode (Abschnitt 4.3) wird dies mit einem kurzlebigen Diffie-Hellman-Austausch erreicht, d. h., ein in den Quick Mode zusätzlich aufgenommener "Key Exchange Payload" (KE) dient zur Berechnung eines nur für das folgende Schlüsselupdate benutzten gemeinsamen Geheimnisses.

5.4 Nachrichtenanzahl

Bei SKIP ist kein zusätzlicher Nachrichtenaustausch zur Etablierung der Schlüssel notwendig. Statt dessen findet die Berechnung des K_{ij} ohne Kommunikation mit dem Empfänger direkt beim Initiator statt. Im ersten gesendeten Datagramm ist der K_{ij} bereits enthalten und kann vom Empfänger verifiziert werden. Letzterer verfährt auf analoge Weise. K_{ij}-Updates werden ebenfalls ohne zusätzliche Kommunikation über den Zähler gemacht. Photuris benutzt sechs Nachrichten: zwei zur Parameterverhandlung, zwei zum Diffie-Hellman-Austausch, zwei zur Authentifizierung. Die im Main Mode von IKE ausgetauschten Nachrichten sind in Anzahl und Bedeutung mit denen von Photuris vergleichbar. Der Aggressive Mode von IKE kommt mit drei Nachrichten aus. Während Photuris beim Schlüsselupdate zwei Nachrichten sendet, braucht IKE im Quick Mode drei Botschaften.

6 Zusammenfassung

SKIP, Photuris und IKE sind drei Schlüsselmanagement-Protokolle, die mit dem Ziel entworfen werden, eine sichere Datenübertragung über das Internet zu ermöglichen.

SKIP setzt zertifizierte Diffie-Hellman-Schlüssel und eine explizite Diffie-Hellman-Gruppe voraus. Über das gemeinsame Geheimnis zweier Kommunikationspartner werden Langzeitschlüssel generiert. Damit werden die per Zufall erzeugten Paketschlüssel chiffriert. Diese wiederum sind Grundlage für Authentifizierungs- und Chiffrier-Schlüssel.

Die Basis-Spezifikation von SKIP bietet keine Perfect Forward Secrecy. Die Kommunikationsinitialisierung findet implizit mit dem ersten gesendeten Datagramm statt. Schlüsselupdates – sowohl der Langzeit- als auch der Paketschlüssel – werden ebenfalls implizit vorgenommen und benötigen keinen zusätzlichen Overhead.

Photuris' drei Phasen zur Erstellung einer Sicherheits-Assoziation sind Cookie-Austausch, Werte-Austausch und Identifikations-Austausch. Der Cookie-Austausch richtet sich gegen Denial-of-Service-Angriffe und beginnt die Verhandlung über ein Austausch-Schema (nicht unbedingt Diffie-Hellman). Im Werte-Austausch erfolgt vor allem die Festlegung auf ein Schema bzw., bzgl. Diffie-Hellman, die Festlegung auf eine Gruppe. Im Identifikations-Austausch werden die Partner authentifiziert, Sitzungsschlüssel erstellt und eine Verifikation der bisherigen Kommunikation durchgeführt.

Identifikation der Parteien ist a) durch vom Administrator vorkonfigurierte Identitäten und Passwörter oder b) über IPSec-AH-basierte Authentifizierung mittels der SPI-Sitzungsschlüssel möglich. Photuris bietet Perfect Forward Secrecy. Es braucht sechs Nachrichten zum SA-Aufbau und zwei für ein Schlüsselupdate.

IKE ist ein hybrides, auf Oakley und SKEME basierendes, ISAKMP-konformes Protokoll. Die SA wird im Main Mode (6 Nachrichten) oder im Aggressive Mode (3 Nachrichten) aufgebaut. Im Quick Mode (3 Nachrichten) finden Schlüsselupdates statt. Der Main Mode besteht aus zwei Nachrichten zur Verfahrensverhandlung, zwei weiteren zum Austausch von DH-Werten und zweien zur Authentifizierung.

IKE verhandelt den Verschlüsselungs- und den Hash-Algorithmus, ggf. eine pseudozufällige Funktion und eine von vier standardmäßigen Diffie-Hellman-Gruppen. Weitere Gruppen können mit den New Group Mode im Anschluß an Main oder Aggressive Mode verhandelt werden. Authentifizierung ist möglich über digitiale Signaturverfahren, Public Key Algorithmen und separat vereinbarte Schlüssel. IKE besitzt die Möglichkeit zur Perfect Forward Secrecy.

Von allen drei Protokollen gilt IKE wegen der ISAKMP-Konformität und der großen Flexibilität als das zukunftsträchtigste.

Literatur

[Atki95a] R. Atkinson. IP Authentication Header. *Standards Track RFC-1826*, August 1995.

[Atki95b] R. Atkinson. IP Encapsulating Security Payload (ESP). *Standards Track RFC-1827*, August 1995.

[Atki95c] R. Atkinson. Security Architecture for the Internet Protocol. *Standards Track RFC-1825*, August 1995.

[Aziz97] A. Aziz. SKIP Extension for Perfect Forward Secrecy (PFS). *http://skip.incog.com/spec/EPFS.html*, April 1997.

[AzMP96] A. Aziz, T. Markson und H. Prafullchandra. Simple Key-Management Protocol for Internet Protocols (SKIP). *Internet Draft draft-ietf-ipsec-skip-07.txt, work in progress*, August 1996.

[DiHe76] W. Diffie und M. E. Hellman. New Directions in Cryptography. *IEEE Transactions on Information Theory* IT-22(6), November 1976, S. 644–654.

[ErCS94] D. Eastlake 3rd, S. Crocker und J. Schiller. Randomness Recommendations for Security. *Informational RFC-1750*, Dezember 1994.

[Ford94] W. Ford. *Computer Communications Security.* Prentice-Hall. 1994.

[HaCa98] D. Harkins und D. Carrel. The Internet Key Exchange (IKE). *Internet Draft draft-ietf-ipsec-isakmp-oakley-08.txt, work in progress*, Juni 1998.

[KaSi98] P. Karn und W. A. Simpson. Photuris: Session-Key Management Protocol. *Internet Draft draft-simpson-photuris-18.txt, work in progress*, Februar 1998.

[Kraw] H. Krawczyk. SKEME: A Versatile Secure Key Exchange Mechanism for Internet. *from IEEE Proceedings of the 1996 Symposium on Network and Distributed Systems Security.*

[KrBC97] H. Krawczyk, M. Bellare und R. Canetti. HMAC: Keyed-Hashing for Message Authentication. *Informational RFC-2104*, Februar 1997.

[MSST98] D. Maughan, M. Schertler, M. Schneider und J. Turner. Internet Security Association and Key Management Protocol (ISAKMP). *Internet Draft draft-ietf-ipsec-isakmp-09.ps, work in progress*, Marz 1998.

[Orma96] H. Orman. The Oakley Key Determination Protocol. *draft-ietf-ipsec-oakley-02.txt*, 1996.

[oSTe94] National Institute of Standards und Technology. Digital Signature Standard. *NIST FIPS PUB 186, U.S. Department of Commerce*, Mai 1994.

[Rive92] R. L. Rivest. The MD5 Message Digest Algorithm. *RFC-1321*, April 1992.

[Schn96] B. Schneier. *Angewandte Kryptographie.* Addison-Wesley. 1996.

Unterstützung integrierter Dienste im LAN-Bereich

Jürgen Blaschek

Kurzfassung

Mit zunehmender Leistungsfähigkeit der lokalen Netze und des Internets steigen auch die Anforderungen. Bei Anwendungen mit zeitabhängigem Datenverkehr, wie z.B. Internet-Telefonie oder Video-Konferenzen, soll der Datenfluß zwischen den Endsystemen dabei möglichst stetig und reibungslos ablaufen, wofür den Anwendungen im Netzwerk kurzzeitig entsprechend viele Ressourcen zur Verfügung gestellt werden müssen. Für die Reservierung der zur Verfügung stehenden Ressourcen sorgt dabei das Reservierungsprotokoll RSVP. Dafür wird ein Modell benötigt, das im LAN-Bereich für die Ressourcenreservierung sorgt und die mit RSVP eingeführten Dienstklassen auf Schicht-2-Dienste abbildet.

1 Einleitung

Die Architekturen für lokale Ethernet-Netze und der Fast Ethernet-Standard IEEE 802.12, wie auch das Internet, unterstützen traditionsgemäß nur den „best-effort"-Verkehr, d.h. sie erbringen Dienste mit bestmöglicher Qualität. Jedoch werden mit den zunehmend verbesserten Übertragungstechnologien und der Vielzahl von neu entwickelten Echtzeitanwendungen, wie z.B. Videokonferenzen oder Internet-Telefonie, auch neue Anforderungen an diese Übertragungstechnologien gestellt, d.h. es werden Mechanismen benötigt, die Echtzeitdienste über Internet bzw. Intranet ermöglichen. Mit dem RFC 1633 (Request for Comments) wurde von der Arbeitsgruppe für Integrierte Dienste der IETF (Internet Engineering Task Force) ein erstes Regelwerk geschaffen, das diese neuen Anforderungen berücksichtigt: Es wurden verschiedene Klassen von Netzwerkdiensten mit verbesserter Dienstqualität definiert, wie z.B. „Kontrollierte Last" („Controlled Load") und die Klasse „Garantierte Dienste" („Guaranteed Service").

Die Dienstklasse der Kategorie „kontrollierte Last" versucht, ein Ende-zu-Ende Verkehrsverhalten zu erreichen, das den traditionellen „best-effort"-Diensten in unbelasteten oder nur leicht belasteten Netzwerken nahekommt. Um die voraussichtliche Netzwerkbelastung abschätzen zu können, gibt die Anwendung, welche die „Kontrollierte Last"-Dienstklasse anfordert, dem Netzwerk eine Schätzung, wieviel Netzressourcen voraussichtlich von ihr benötigt werden.

Die „Garantierte Dienste"-Klasse bietet den Anwendungen die Möglichkeit, Übertragungen mit garantierter Bandbreite und Verzögerungszeiten zu nutzen. Es wird dafür lediglich die akkumulierte, variable Verzögerung im Ende-zu-Ende-Verkehrspfad beschränkt, die durch unterschiedliche Bearbeitungszeiten in den Zwischensystemen

entsteht. Die „garantierte Dienste“-Klasse kontrolliert jedoch nicht die Minimal- oder Durchschnittsverzögerung bzw. den Jitter (Abweichung zwischen minimaler und maximaler Verzögerung).

Jede dieser Dienstklassen wurde entwickelt, um verschiedene Dienstqualitäten anzubieten, wobei verschiedene Parameter eingehalten werden müssen. Eine Anwendung muß nun diejenige Dienstklasse auswählen, welche die gestellten Anforderungen am besten erfüllt.

Einen Mechanismus, um als Endsystem solche Dienste in einem IP-Netzwerk zu benutzen, bietet z.B. das Ressourcenreservierungsprotokoll RSVP. RSVP arbeitet im Gegensatz zu den meisten anderen Reservierungsprotokollen verbindungslos. Es werden die Ressourcen entlang eines durch herkömmliche Routing-Protokolle bestimmten Weges reserviert. Die IP-Datagramme folgen dann diesem Weg und nutzen die für sie „bestimmten“ Ressourcen. Es handelt sich beim RSVP um empfängerorientierte Reservierungen, d.h. der Empfänger initiiert die Reservierung, wobei je Empfänger unterschiedliche Dienstanforderungen möglich sind. Eine Sitzung beginnt dabei mit dem Initiieren der Reservierung und endet mit dem Abbau der Reservierung.

RSVP-fähige Systeme müssen verschiedene Elemente enthalten:

- Zugangskontrolle: Prüft, ob der Knoten die Anforderung erfüllen kann.
- Policy-Kontrolle: Entscheidet, ob eine Anwendung Reservierungen machen darf.
- Classifier: Teilt Pakete einer QoS-Klasse zu.
- Packet Scheduler: Teilt die Ressourcen zu, z.B. Verwaltung der Bandbreite, bestimmen des nächsten zu sendenden Paketes.

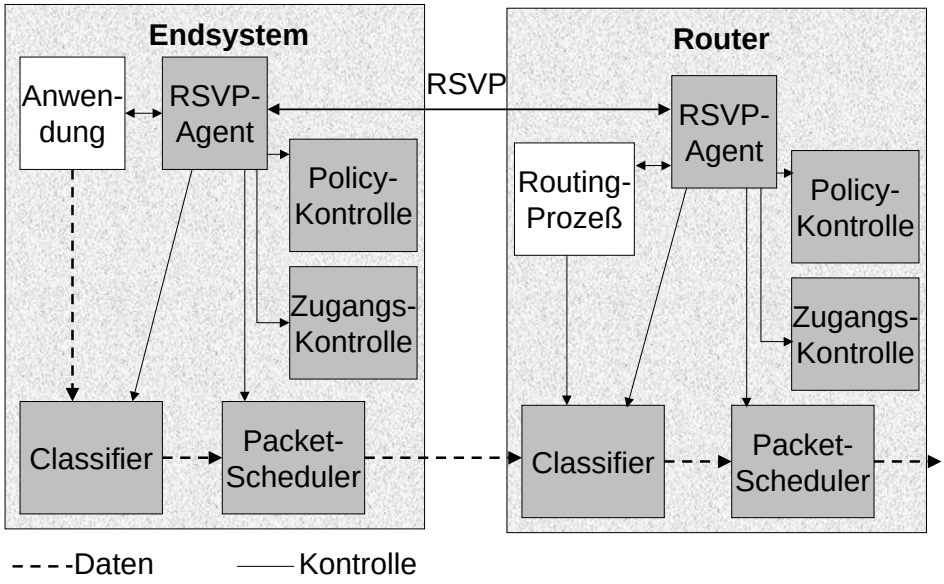

Abbildung 1: Endsystem und Router eines RSVP-fähigen Systems

Ein initialer RSVP-Protokollablauf sieht dabei wie folgt aus:

1. Ein Sender schickt periodisch Path-Nachrichten und baut damit einen Baum auf, wobei die wichtigsten Informationen in jedem auf dem Verkehrspfad liegenden Knoten gespeichert werden. Die Path-Nachrichten enthalten dabei folgende Informationen:
   - Senderbeschreibung (z.B. IP-Adresse, Port, FlowLabel)
   - Verkehrscharakteristiken des Senders
   - Aushandlungsinformationen für den Empfänger (optional)
2. Ein Empfänger tritt der Multicast-Gruppe bei
3. Der Empfänger erhält die Path-Nachricht und sendet Resv-Nachrichten mit seinen Anforderungen. Die Resv-Nachrichten durchlaufen das System auf dem gleichen Pfad zum Sender, auf dem die Path-Nachricht zum Empfänger gekommen ist
4. Der Sender erhält die Resv-Nachricht und beginnt mit Senden

Die IEEE (Institute of Electrical and Electronical Engineers) hat in ihrem 802-Projekt Standards für viele verschiedene LAN-Technologien definiert. Diese bieten den Protokollen auf höheren Schichten, wie z.B. IP, typischerweise alle den gleichen MAC-Schicht-Datagrammdienst an, obwohl die Implementierungen oft unterschiedliche dynamische Verhaltenscharakteristiken aufweisen. Die Beachtung dieser Unterschiede ist wichtig, wenn später die Fähigkeit zur Unterstützung von Realzeitdiensten betrachtet wird. IEEE definiert 802-Standards, um mehrere LAN-Segmente zusammenzuschließen, indem Geräte wie z.B. „MAC-Bridges“ oder „Switches“ benutzt werden. Zusätzlich wurden Verkehrsklassen, „Multicast Filtering“ und virtuelle LAN-Fähigkeiten für diese Geräte definiert.

Solche LAN-Technologien bilden oft den letzten Abschnitt zwischen Internet/Intranet und dem Endbenutzer, genauso wie sie für ganze Campus-Netzwerke den ersten Baustein darstellen. Aus diesem Grund ist es wichtig, standardisierte Mechanismen anzubieten, um diese Technologien für Ende-zu-Ende-Realzeitdienste zu nutzen. Deshalb muß es Mechanismen geben, die Ressourcenmanagement auf der Sicherungsschicht (Data Link Layer) unterstützen. Ressourcenmanagement beinhaltet in diesem Kontext auch die Funktionen Zugangskontrolle, Scheduling, Verkehrsüberwachung („Traffic Policing“), etc. Die ISSLL (Integrated Services over Specific Link Layers)-Arbeitsgruppe der IETF wurde mit dem Ziel ins Leben gerufen, Mechanismen zur Unterstützung integrierter Dienste zu erforschen und zu standardisieren.

In den nun folgenden Abschnitten werden zuerst die Voraussetzungen und Ziele erläutert, die mit den Integrierten Diensten erreicht werden sollen. Danach wird die Funktion des Bandbreitenmanagers am Beispiel des Subnetzbandbreitenmanagers (SBM) und seine Einordnung in das Modell der Integrierten Dienste beschrieben.

2 Die ISSLL-Architektur

In diesem Abschnitt werden die Voraussetzungen und Ziele beschrieben, die das Design einer Architektur beeinflussen sollten, um Integrierte Dienste über LAN-Technologien

unterstützen zu können. Die Voraussetzungen beziehen sich dabei auf Funktionen und Eigenschaften, die unterstützt werden müssen, während die Ziele auf Funktionen und Eigenschaften anspielen, die wünschenswert wären, aber nicht unbedingt notwendig sind.

Voraussetzungen:

- Ressourcen-Reservierung: Der Ressourcenreservierungsmechanismus muß imstande sein, Ressourcen auf einem einzelnen oder auf mehreren Segmenten zu reservieren und diese über Bridges/Switches zu verbinden.
- Zugangskontrolle: Der Zugangskontrollmechanismus muß in der Lage sein, die Anzahl der benötigten Ressourcen zu berechnen oder zu schätzen, um im Voraus eine nicht erfüllbare Anforderung ablehnen zu können.
- „Flow Separation“ und „Scheduling“: Es ist notwendig, einen Mechanismus für die Trennung der Verkehrsflüsse anzubieten, so daß den Realzeit-Flüssen eine bevorzugte Behandlung vor den „best-effort“-Flüssen gegeben werden kann.
- Überwachung/Verkehrsformung: Verkehr muß von den End- und Zwischensystemen (Workstations, Router) geformt und/oder überwacht werden, um die Übereinstimmung mit den verhandelten Verkehrsparametern zu garantieren. Geformten Verkehr zu erzeugen ist dabei das empfohlene Verhalten für Verkehrsquellen.
- „Soft State“: Die Statusinformationen müssen periodisch erneuert werden, solange die Reservierung erhalten bleibt.
- Zentrale oder verteilte Implementierung: Im Falle einer zentralen Implementierung verwaltet eine Station die Ressourcen eines ganzen Subnetzes. Dies hat den Vorteil, daß es einfacher einzusetzen ist, da Bridges und Switches nicht notwendigerweise durch zusätzliche Funktionalität erweitert werden müssen. Bei der verteilten Implementierung hat jede Station eine eigene, die Ressourcen verwaltende Einheit. Der Vorteil dieser Implementierung ist die bessere Skalierbarkeit, jedoch müssen nun alle Bridges und Switches diesen Mechanismus auch unterstützen.
- Skalierbarkeit: Die Mechanismen und Protokolle sollten niedrigen Aufwand verursachen und bis zu den größten Empfängergruppen effizient funktionieren.
- Fehlertoleranz und Wiederherstellung: Der Mechanismus muß in der Lage sein, auch während des Auftretens von Fehlern zu funktionieren.
- Interaktion mit existierenden Ressourcenmanagementkontrollen: Die Wechselwirkung mit existierenden Infrastrukturen für Ressourcenmanagement muß spezifiziert werden.

Ziele:

- Unabhängigkeit von Protokollen der höheren Schichten: Der Mechanismus sollte so weit wie möglich von Protokollen höherer Schichten wie z.B. RSVP und IP unabhängig sein. Diese Unabhängigkeit ist wünschenswert, um die Zusammenarbeit mit anderen Reservierungsprotokollen zu sichern.

- Empfängerheterogenität: Dies bezieht sich auf Multicast-Kommunikation, wo verschiedene Empfänger verschiedene Dienstqualitäten beanspruchen können.
- Unterstützung für verschiedene Filterarten: Es ist wünschenswert, für die verschiedenen von RSVP definierten Filterarten Unterstützung anzubieten.
- Pfadauswahl: In quellgerouteten LAN-Technologien wie z.B. Token Ring/IEEE 802.5 wäre es sinnvoll für den Mechanismus, die Funktion der Pfadauswahl mit einzubeziehen, um den Gebrauch der Netzwerkressourcen zu optimieren.

2.1 Das Konzept des Bandbreitenmanagers

Die gerade beschriebenen funktionalen Voraussetzungen werden von einem Element namens Bandbreitenmanager (BM - Bandwidth Manager) durchgeführt. Der BM ist dafür zuständig, daß den Anwendungen oder Protokollen der höheren Schicht die Mechanismen zur Anforderung der Dienstqualität eines Netzwerkes angeboten werden. Der BM besteht aus einem Requester Module (RM) und einem Bandwidth Allocator (BA), wobei diese nicht zusammen in einem System realisiert sein müssen.

Das RM befindet sich in jedem Endsystem eines Subnetzes. Eine seiner Funktionen ist, eine Schnittstelle zwischen Anwendungen mit Protokollen der höheren Schicht wie RSVP, ST2, SNMP, etc. und dem BM anzubieten. Eine Anwendung kann die verschiedenen Funktionen des BM aufrufen, indem sie die Schnittstellenprimitive des RM benutzt und diese mit entsprechenden Parametern versieht: Dem gewünschten Dienst (garantierter Dienst oder kontrollierte Last), die Verkehrsbeschreibung und die enthaltene Summe an Ressourcen, die reserviert werden sollen. Das RM muß danach die Vermittlungsschichtadressen in Sicherungsschichtadressen übersetzen und die Anfrage in ein für die anderen Komponenten des BM verständliches Format konvertieren. Desweiteren ist das RM für die Rückgabe der Statusinformationen der durch den BM bearbeiteten Anfragen verantwortlich.

Der Bandwidth-Allocator ist für die Ausführung der Ressourcenzuweisung, d.h. Zugangskontrolle und aktualisieren des Ressourcenzustands, im Subnetzwerk verantwortlich. Ein Endsystem kann verschiedene Dienste wie z.B. Bandbreitenreservierung, Änderung einer bestehenden Reservierung, Fragen zur Ressourcenverfügbarkeit, etc. anfordern. Diese Anfragen werden vom BA bearbeitet. Die Kommunikation zwischen Endsystem und BA wird über den RM abgewickelt.

Für die Kommunikation zwischen den verschiedenen Komponenten des BM müssen Protokolle spezifiziert werden:

- Kommunikation zwischen Protokollen höherer Schichten und RM: Der BM muß Primitive für die Anwendungen definieren, um Reservierungen zu initiieren, Anfragen an den BA wegen verfügbarer Ressourcen stellen, Reservierungen ändern oder löschen, etc. Diese Primitive können als Anwendungsschnittstelle (Application Programming Interface - API) implementiert werden.
- Kommunikation zwischen RM und BA: Ein Signalisierungsmechanismus muß für die Kommunikation zwischen RM und BA definiert werden. Dieses Protokoll spezifiziert die Nachrichten, die zwischen RM und BA ausgetauscht werden müssen, um verschiedene Anfragen durch die höhere Schicht zu bedienen.

- Kommunikation zwischen gleichwertigen BAs: Für den Fall, daß sich in einem Subnetz mehr als ein BA befindet, muß ein Mittel zur Kommunikation zwischen diesen BAs zur Verfügung gestellt werden. BAs sind in der Lage, untereinander auszumachen, wer für welches Segment, welche Brücke oder welchen Switch verantwortlich ist.

Anhand von Beispielen werden zuerst die beiden verschiedenen Implementierungsmöglichkeiten der zentralen und verteilten Implementierung vorgestellt. Beiden Implementierungen ist gemeinsam, daß das RM in jedem Endsystem, das Reservierungen vornehmen möchte, vorhanden sein muß. Der BA wiederum ist nun entweder in jeder vermittelnden Bridge/Switch vorhanden (verteilte Implementierung), oder es gibt in einem Subnetz nur einen BA der dann für die Zugangskontrollentscheidungen des ganzen Subnetzes verantwortlich ist. In diesem Fall spricht man von zentraler Implementierung.

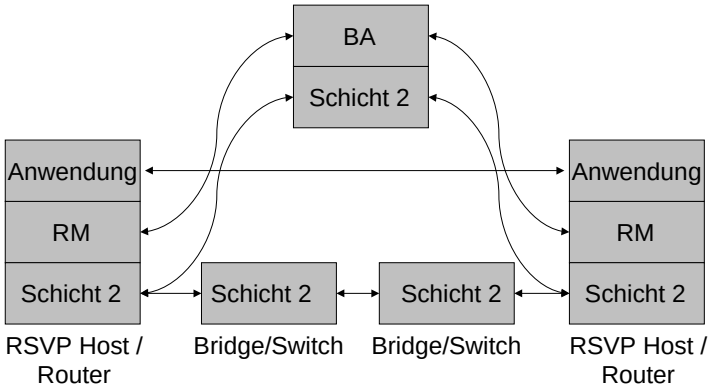

Abbildung 2: Bandbreitenmanager mit zentralem Bandwidth-Allocator

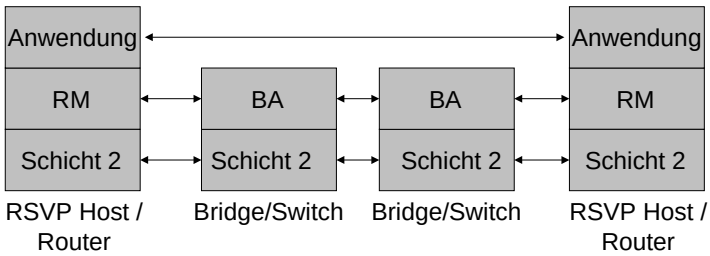

Abbildung 3: Bandbreitenmanager mit verteiltem Bandwidth-Allocator

2.2 Der Subnetzbandbreitenmanager

Das Konzept des *Subnetz-Bandbreiten-Managers (Subnet Bandwith Manager - SBM)* ist ein Vorschlag, der ein standardisiertes Signalisierungsprotokoll für LAN-basierte Zugangskontrollverfahren im Zusammenhang mit RSVP-Datenflüssen anbietet.

Das SBM-Protokoll und sein Gebrauch zur Zugangskontrolle und Bandbreitenmanagement basieren auf den folgenden architektonischen Zielen und Annahmen:

- Zur Verwaltung der Bandbreite wird ein Signalisierungsprotokoll spezifiziert, das auch die seit kurzem exisitierenden zusätzliche Funktionalitäten, wie z.B. die explizite Unterstützung verschiedener Verkehrsklassen und verschiedener Integrierte Dienste berücksichtigt.

- Durch den SBM ist lediglich eine Signalisierungsmethode und ein Signalisierungsprotokoll für LAN-basierte Zugangskontrollen definiert. Es werden keine Verkehrskontrollmechanismen für die Schicht 2 angeboten. Das Protokoll wurde entworfen, um die schon vorhandenen und durch IEEE 802 definierten Mechanismen auszunutzen.

- Fehlen Verkehrskontrolle oder prioritätengesteuerten Warteschlangenmechanismen im darunterliegenden LAN, beschränkt der SBM-basierte Zugangskontrollmechanismus nur die Gesamtmenge der Verkehrslast, die durch RSVP-Datenflüsse erzeugt wird.

Dabei strebt der SBM-Vorschlag bei Kombination von Flußkontrolle des Endsystems mit Verkehrskontrolle und Prioritäts-Warteschlangen auf der Schicht 2 eine Annäherung an die „Kontrollierte Last"- und „Garantierte Dienste"-Klasse an. In Umgebungen dieser Art ist kein Mechanismus vorhanden, der zwischen RSVP-Datenflüssen und „best-effort"-Verkehr unterscheidet. Dies stellt die Brauchbarkeit des SBM-Modells in einer LAN-Infrastruktur, die das Paketversenden mit einer IEEE 802.1p Prioritätsstufe nicht unterstützt, in Frage.

In jedem von einem SBM verwalteten Segment gibt es einen SBM, der dazu bestimmt wird, das LAN-Segment zu leiten. Dieser ausgezeichnete SBM (Designated Subnet Bandwidth Manager - DSBM) ist eine Protokolleinheit in einem Schicht-2- oder Schicht-3-Gerät und verwaltet die Ressourcen eines Schicht-2-Segmentes. Meistens existieren DSBMs für jedes Segment und sind mit Informationen über die maximal zu reservierende Bandbreite jedes Unter-Segments versehen.

Es gibt zwei Möglichkeiten, wie ein SBM zum DSBM bestimmt werden kann: Durch statische Konfiguration in SBM-fähigen Geräten oder durch dynamische Auswahl. Startet im Falle der dynamischen Auswahl ein SBM zum ersten Mal seinen Dienst, so bleibt er im Hintergrund und wartet auf eingehende DSBM-Funktionsnachrichten, bis die Wahl eines neuen DSBM nötig wird. Existiert kein DSBM, so sorgt der SBM für eine Wahl, indem er die DSBM_WILLING-Nachricht aussendet, die seine IP-Adresse und SBM-Priorität als DSBM-Kandidat enthält. Im folgenden sendet jeder SBM seine Priorität und Adresse aus. Der DSBM wird nun anhand der Priorität bestimmt. Gibt es zwei DSBM-Kandidaten mit gleicher Priorität, so entscheidet ein Wahlalgorithmus über die IP-Adresse, welcher SBM bevorzugt zu behandeln ist. Ist der DSBM einmal gewählt, so sendet er periodisch (z.B. alle 5 Sekunden) eine I_AM_SBM-Nachricht an alle SBMs. Bleibt diese Nachricht eine bestimmte Zeit lang aus, so bedeutet dies, daß der DSBM nicht mehr seine Aufgabe erledigen kann. Dies löst dann eine neue Wahl zwischen den bestehenden SBMs aus. Wenn kein DSBM zur Verfügung steht, können zur Kommunikation die Standard-RSVP-Weiterleitungsregeln benutzt werden.

Sendet ein SBM eine Pfad-Nachricht über eine an einem verwalteten Segment angeschlossene Schnittstelle oder gibt diese weiter, so schickt er die Nachricht an den DSBM statt an die RSVP-Sitzungszieladresse, wie dies in konventionellen RSVP-Prozessen gemacht wird. Als Teil eines Prozesses initiiert und verwaltet der DSBM einen Pfadstatus für die Sitzungen und vermerkt den vorherigen Knoten (Previous Hop - PHOP), der die Nachricht sendete. Möchte ein DSBM-Client eine Reservierung für eine RSVP-Sitzung machen, folgt er den Standard-RSVP-Regeln und sendet eine Resv-Nachricht an die entsprechende PHOP-Adresse, die in einer angekommenen Pfad-Nachricht festgelegt

wurden. Die DSBM-Prozesse erhalten Resv-Nachrichten mit den möglichen Bandbreiten und beantworten diese mit ResvErr-Nachrichten an den Anfragenden, falls die Anfrage nicht gewährt werden kann, bzw. leiten die Resv-Nachricht an den PHOP weiter, wenn ausreichend Ressourcen vorhanden sind und die Reservierungsanfrage gewährt werden kann. Der DSBM verschmelzt und ordnet Reservierungsanfragen in Übereinstimmung mit traditionellen RSVP-Regeln.

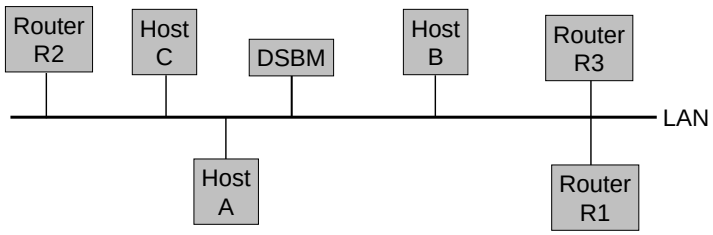

Abbildung 4: Beispiel eines verwalteten Segmentes

Abbildung 2 zeigt ein Beispiel eines verwalteten Segments in einer Schicht 2-Domäne, die einige Hosts und Router miteinander verbindet. Die grundlegende DSBM-basierte Zugangskontrollprozedur funktioniert wie folgt:

1. DSBM-Initialisierung: Als Teil einer initialen Konfiguration bekommt der DSBM Informationen wie z.B. die Grenzen der Fragmentierung von erhältlichen Ressourcen, die in jedem verwalteten Segment unter seiner Kontrolle reserviert werden können. Z.B. ist die Bandbreite eine solche Ressource.

2. DSBM Client-Initialisierung: Für jedes angehängte Element entscheidet ein DSBM-Client, ob ein DSBM an dieser Schnittstelle existiert.

3. DSBM-basierte Zugangskontrolle: Um eine Ressourcenreservierung anzufragen, folgen die DSBM-Clients den folgenden Schritten:

   (a) Wenn ein DSBM-Client eine RSVP-Pfadnachricht über eine mit dem verwalteten Segment verbundene Schnittstelle sendet oder weiterleitet, sendet er die Pfadnachricht anstatt zur RSVP-Sitzungszieladresse zum DSBM des Segments. Nach der Behandlung dieser Nachricht leitet der DSBM diese zur Zieladresse weiter. Als Teil des Prozesses initiiert und speichert der DSBM einen Pfadstatus für die Sitzungen und vermerkt den vorigen Schicht 2 oder Schicht 3-Hop, der ihm die Pfadnachricht gesendet hat.

   (b) Wenn eine Anwendung im Host A eine Reservierung für eine RSVP-Sitzung machen möchte, folgt der Host A den Standard-RSVP-Protokollnachrichtenregeln und sendet eine RSVP Resv-Nachricht zur Schicht 2/Schicht 3-Adresse des vorigen Hops.

   (c) Der DSBM behandelt die RSVP Resv-Nachricht basierend auf der verfügbaren Bandbreite und gibt eine ResvErr-Nachricht an den Anfragenden (Host A) zurück, wenn für die Anfrage keine Garantie übernommen werden kann. Wenn jedoch die benötigten Ressourcen verfügbar sind und für die Reservierungsanfrage eine Garantie übernommen werden kann, gibt der DSBM die Resv-Nachricht zum PHOP weiter.

(d) Enthält die Schicht 2-Domäne mehr als ein verwaltetes Segment, so könnten der Requester (Host A) und der Forwarder (Router R1) durch mehr als ein verwaltetes Segment getrennt sein. In diesem Fall würde sich die ursprüngliche Pfad-Nachricht durch mehrere/viele DSBMs (eines für jedes verwaltete Segment auf dem Weg von R1 nach A) fortpflanzen und in jedem DSBM den Pfadstatus verändern. Aus diesem Grund würde sich die Resv-Nachricht in umgekehrter Richtung von Knoten-zu-Knoten durch die vermittelnden DSBMs fortpflanzen und eventuell den ursprünglichen Forwarder (Router R1) in der Schicht 2-Domäne erreichen, falls die Zugangskontrolle in allen DSBMs erfolgreich verlief.

Nachdem in diesem Abschnitt die Funktionsweise des Bandbreitenmanager am Beispiel des SBMs beschrieben wurde, folgt nun die Einordnung eines Bandbreitenmanagers in ein lokales Netzwerk.

Im Endsystem-Modell ist jeder sendende Client (siehe Abbildung 5)(Client hat hier die Bedeutung eines auf Schicht 3, jedoch am Übergang zu Schicht 2 operierenden Gerätes) für die lokale Zugangskontrolle und das Paket-Scheduling seiner Verbindung gemäß der ausgehandelten Dienste selbst verantwortlich. Würde dieser Client einen RSVP-Prozeß starten, der folgendes anbietet: Eine Schnittstelle zum Einrichten einer Sitzung für Anwendungen, das Senden von Signalisierungen über das Netzwerk, das Programmieren eines Schedulers und Classifiers im Treiber und Schnittstellen zu einem Überwachungskontrollmodul. Insbesondere bildet RSVP auch eine Schnittstelle zu einem lokalen Zugangskontrollmodul, das im Mittelpunkt unserer Betrachtungen stehen soll. Sie ist innerhalb des Clients für die Abbildung von Schicht 3-„Session Establishment Requests“ in die Schicht 2-Sprache verantwortlich.

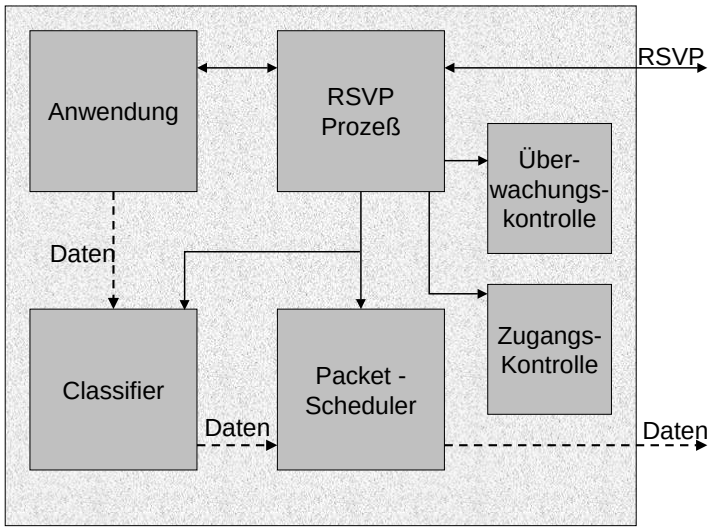

Abbildung 5: RSVP in einem sendenden Host

Die Funktionen des Requester-Moduls in einem sendenden Endsystem (Abbildung 6) können wie folgt beschrieben werden: Vom IP- bzw. vom RSVP-Protokoll kommt der Auftrag, Daten zu versenden. Die vom IP-Protokoll mitgelieferte Zieladresse wird im Adreßabbildungsmodul zu einer Schicht 2-Adresse konvertiert und dem Request-Modul übergeben. Das Request-Modul fragt bei dem lokalen BA-Modul an, ob der lokale Zugang gewährt werden kann. Desweiteren wird eine SBM-Anfrage an weitere

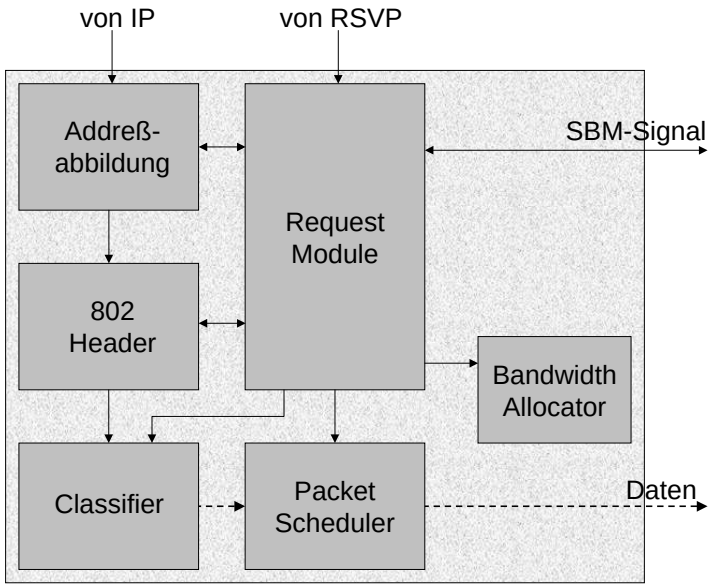

Abbildung 6: ISSLL in einem sendenden Endsystem

BA weitergeleitet. Diese Anfrage enthält neben der Schicht 2-Adresse die notwendigen RSVP-Parameter. Empfängt das Endsystem die Antwort des Netzwerks, dann wird die Zugangskontrollentscheidung inklusive der durch Aushandlung veränderten Parameterwerte an die Einheiten der höheren Schicht übergeben. Die vom Empfänger kommende user_priority wird in der „802 Header"-Tabelle gespeichert. Diese wird benutzt, wenn für die zu sendenden Datenpakete die Schicht 2-Köpfe erzeugt werden. Die BA-Komponente ist nur dann vorhanden, wenn ein verteiltes BA-Modell implementiert wurde. Wenn sie vorhanden ist, hat sie hauptsächlich die Funktion, die lokale Zugangskontrolle für die abgehenden Verbindungsbandbreiten sicherzustellen.

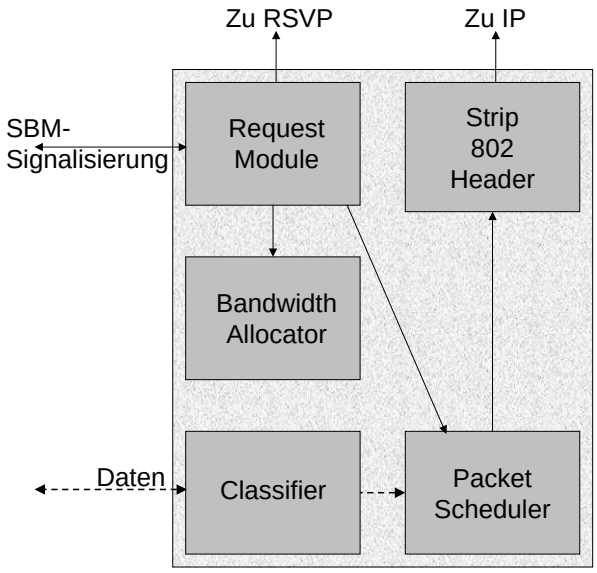

Abbildung 7: ISSLL in einem empfangenden Endsystem

Die Abläufe in einem empfangenden System (Abbildung 7) finden wie folgt statt:

Das Request Modul beantwortet schon vor der Übertragung der eigentlichen Datenpakete die Anfragen des Request Moduls des Senders. Es handelt mit jedem lokalen BA die Zugangskontrollentscheidungen aus und gibt die Anfragen an RSVP weiter, wenn

diese zugelassen wurden. Nach dem Aushandeln der Übertragungsformalitäten teilt des Request Modul dem Packet Scheduler mit, unter welchen Bedingungen die Datenpakete beim Empfänger ankommen werden. Der Classifier empfängt dann die Datenpakete, identfiziert sie, und gibt sie an den Packet Scheduler weiter. Das „Strip 802 Header"-Modul entfernt dann die Schicht 2-Datenköpfe und gibt die restlichen Daten an IP weiter.

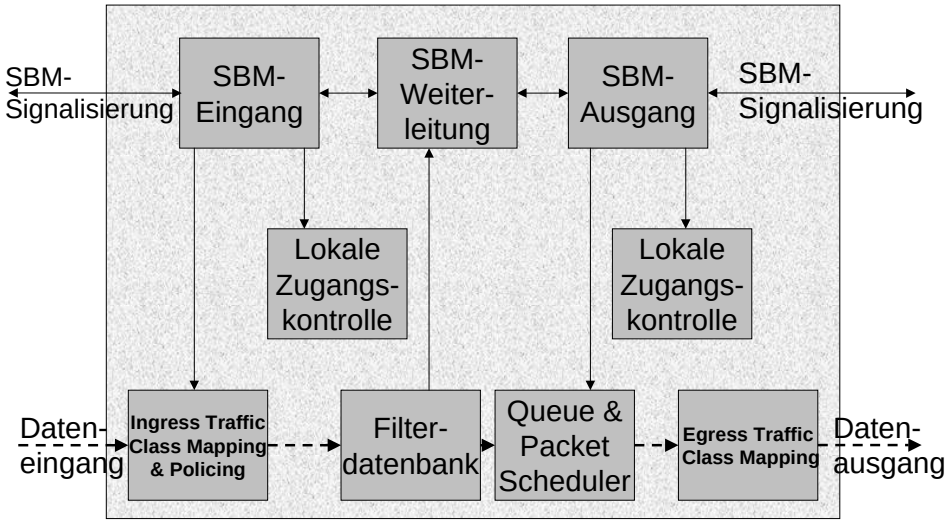

Abbildung 8: ISSLL in einem Switch

Das Switch-Modell (Abbildung 8) benutzt die Terminologie des SBM-Protokolls als Beispiel für ein Zugangskontrollprotokoll. Das Modell ist jedoch genauso anwendbar, wenn andere Mechanismen wie z.B. statische Konfiguration oder Netzwerkmanagement als Zugangskontrolle benutzt werden. Innerhalb eines Switches werden die folgenden Dinge definiert:

- Lokales Zugangskontrollmodul: Es befindet sich eines an jedem Port und zählt die an diesem Port verfügbare Verbindungsbandbreite.
- Eingangs-SBM-Modul: Eines an jedem Port führt auf der Netzwerkseite des Signalisierungsprotokolls die Verbindungen mit Clients oder anderen Switches der gleichen Schicht aus.
- SBM Weiterleitungsmodul (SBM Propagation Modul): Leitet Anfragen, welche die Zugangskontrolle am Eingangsport passiert haben, zum relevanten Ausgangsport des SBM-Moduls weiter.
- Ausgangs-SBM-Modul: Leitet Anfragen an die nächste Schicht 2- oder Schicht 3-Teilstrecke weiter.
- Classifier, Queue und Scheduler-Modul: Das Classifier-Modul identifiziert die relevanten QoS-Informationen der eingehenden Pakete und benutzt diese, um zu entscheiden, an welchem Ausgangsport und an welche Verkehrsklasse das Paket weitergegeben werden soll. Queue und Scheduler stellen die Ausgangswarteschlangen für Ports zur Verfügung und bieten den Algorithmus an, mit dem die Pakete zur Übertragung auf die Ausgangs-Verbindung gegeben werden, um damit den IntServ-Dienst anbieten zu können.

- Ingress Traffic Class Mapping und Policing Module: Dieses optional vorkommende Modul kann die Daten innerhalb von Verkehrsklassen auf die Einhaltung der verhandelten Parameter überwachen.

- Egress Traffic Class Mapping Module: Dieses optionale Modul könnte das Remapping von Verkehrsklassen auf einer Basis pro Ausgangsport vornehmen.

Wenn ein zentrales BA-Modell implementiert ist, nehmen Switches am Zugangskontrollprozeß nicht teil. Die Zugangskontrolle ist durch einen zentralen BA implementiert, z.B. einem „Subnetz Bandbreiten Manager".

2.3 Anwendung der Architektur auf verschiedene LAN-Typen

Das Ausmaß, in dem Netzwerke Dienstgarantien anbieten können, hängt zu einem Großteil von der Fähigkeit ab, die Hauptfunktionen Flußidentifikation und Scheduling im Zusammenhang mit Zugangskontrolle und Überwachung anzubieten. Für die hier betrachteten Technologien sind geteilte (shared), halbduplex-geschaltete (switched half duplex) und vollduplex-geschaltete LANs die Haupttopologien.

In der geteilten Topologie teilen sich mehrere Sender ein einziges Segment. Eine halbduplex-geschaltete Topologie ist im wesentlichen eine geteilte Topologie mit der Einschränkung, daß nur zwei Transmitter um die Ressourcen jeglicher Segmente konkurrieren. In einer vollduplex-geschalteten Topologie erhält der Transmitter zu jeder Zeit und an jedem Ende der Verbindung Zugang zu einem Pfad mit der ganzen Bandbreite. Daher gibt es in dieser Topologie keinen Bedarf für Zugangskontrollmechanismen wie z.B. CSMA/CD oder Token Passing, da sich keiner der beiden Transmitter dem anderen gegenüber durchsetzt. Offensichtlich bietet diese Topologie die besten QoS-Fähigkeiten.

In einem *voll-duplex-geschalteten LAN* ist das MAC-Protokoll für den Zugriff unwichtig. Das MAC-Protokoll muß jedoch die Zugriffszeit berücksichtigen, die der Übertragungszeit des größten Paketes entspricht. Voll-duplex geschaltete Netzwerktopologien bieten sowohl für kontrollierte Last, als auch für garantierte Dienste gute QoS-Fähigkeiten, sofern sie innerhalb der Switches von geeigneten Queuing-Strategien unterstützt werden.

Sobald einem *einzelnen, geteilten CSMA/CD-Segment* ein CSMA/CD-Algorithmus präsentiert wird, ist der Versuch, einen garantierten Dienst einzuführen, aufgrund der fehlenden Kopplung zwischen mehreren Sendern einer Übertragungsstrecke ernsthaft gefährdet. Es gibt eine Reihe von Gründen, weshalb man keine bessere Lösung für dieses Problem anbietet. Erstens glaubt man nicht, daß dies ein wirklich lösbares Problem ist, da man dafür eine Reihe von Änderungen am MAC-Protokoll vornehmen müßte. Zweitens ist man nicht überzeugt, daß es sich dabei um ein wirklich interessantes Problem handelt. Drittens besteht der Kern der Campus-Netzwerke typischerweise eher aus auf Switches basierenden Lösungen als auf durch Repeater verbundene Segmente.

Viele der Argumente für sub-optimale Unterstützung von garantierten Diensten im geteilten Medium Ethernet gelten auch für das *halb-duplex geschaltete Ethernet*. Im Wesen ist diese Topologie ein Medium, das zwischen wenigstens zwei Sendern geteilt ist. Auch wenn diese beiden Sender eng miteinander verbunden sind und kooperieren,

besteht immer die Möglichkeit, daß der „best effort"-Verkehr des einen den reservierten Verkehr des anderen verdrängt.

In einem *halb-duplex geschalteten und geteilten Token Ring - Netzwerk* ist die Netzwerkzugriffszeit für hochpriorisierten Verkehr in jeder Station an die Formel $(N + 1)\cdot$ THTmax gebunden, wobei N die Anzahl der hochpriorisiert sendenden Stationen und THTmax die maximale „Token Holding Time" ist. Dabei ist leicht einzusehen, daß die Zugriffszeiten durch die Reduzierung von N oder THTmax verbessert werden können. Da die Zugriffszeit beschränkt ist, kann eine obere Grenze für die Ende-zu-Ende-Verzögerung angeboten werden, wie dies vom garantierten Dienst benötigt wird. Der an den garantierten Dienst gebundene Verkehr bekommt dabei einfach die für Benutzerdaten höchste Priorität.

In *halb-duplex geschalteten und geteilten IEEE 802.12-Netzwerken* (Demand Priority) basiert die Kommunikation zwischen Endknoten und Hubs sowie unter den Hubs selbst auf dem Austausch von Verbindungskontrollsignalen. Diese Signale werden benutzt, um den Zugriff auf das geteilte Medium zu kontrollieren. Die Netzwerkzugriffszeit für hoch-priorisierte Pakete ist im Grunde die Zeit, die gebraucht wird, um vor normal priorisierten Netzwerkdiensten senden zu können. Diese Zugriffszeit ist beschränkt und hängt von der Bitübertragungsschicht und der Topologie des geteilten Netzwerkes ab.

In den letzten Abschnitten wurde, eher nebenbei, die Notwendigkeit erwähnt, Prioritäten bei der Übertragung von Datenpaketen einzuführen. Durch die Einführung eines Feldes zur Prioritätensteuerung sollte es den Netzwerktopologien Token-Ring, FDDI und Demand Priority möglich sein, garantierte Dienste anbieten zu können. In den folgenden Abschnitten wird nun beschrieben, wie in den einzelnen Netzwerktopologien ein Feld zur Prioritätensteuerung, das user_priority-Feld, im Rahmen von ISSLL verwendet werden kann.

Das user_priority-Feld ist ein mit allen Rahmen assoziierter Wert im IEEE-802-Dienstmodell. Je nach Netzarchitektur ist es in den zu übertragenden Rahmen integriert oder auch nicht: In der Token-Ring-Topologie ist es z.B. im FC-Feld integriert, während dieses Feld in einem Ethernet-Rahmen nicht übertragen wird. Die Demand-Priority-Topologie (IEEE 802.12) kann, je nach benutztem Rahmenformat, dieses Feld implementiert haben: Wird das Token Ring-Rahmenformat zur Übertragung benutzt, so ist das user_priority-Feld implementiert, während dies bei der Benutzung des Ethernet-Rahmenformats nicht berücksichtigt wird. Der IEEE 802.1D-Standard definiert einen Weg, wie diese Bits über ein aus Ethernet, Token Ring, Demand Priority, FDDI oder anderen MAC-Schicht-Medien bestehendes gebridgetes Netzwerk übertragen werden können, indem ein erweitertes Rahmenformat benutzt wird. Ist ein explizites Feld zur Aufnahme der user_priority vorhanden, so muß die bestehende Hard- und Software der Schicht-2 nicht verändert oder erweitert werden. Höherentwickelte Schicht-3-Switches wären durch die Nutzung des user_priority-Feldes in der Lage, die zu übertragenden Pakete besser zu klassifizieren und dadurch die bestehenden Ressourcen effizienter zu nutzen. Das user_priority-Feld hat eine Größe von 3 Bit und bietet damit 8 verschiedene Kategorien, beginnend mit dem Wert 7 als höchste und wichtigste bis zu dem Wert 0 als der unwichtigsten Kategorie. Diese Einteilung in Kategorien ist in Tabelle 1 beschrieben. Desweiteren wird das user_priority-Feld benutzt, um die Dienstklassifizierungen der Schicht-3 auf die Schicht-2 abzubilden. Die Einteilung ist ebenfalls in Tabelle 1 beschrieben.

Anwendung	Benutzerpriorität	Dienst
Zeitunkritische Daten	0	Default (Best Effort)
	1	weniger als Best Effort
	2	
	3	
LAN-Management	4	Kontrollierte Last
Zeitkritische Daten	5	Garantierter Dienst, 100ms
Realzeitkritische Daten	6	Garantierter Dienst, 10ms
MAC-Rahmen	7	

Tabelle 1: Einteilung der Daten in Prioritäten

In Ethernet-Paketen gibt es kein explizites „traffic class“- oder „user_priority“-Feld, d.h. dieses Feld müßte in einem geeigneten, an dem Verkehrspfad liegenden Receiver oder Switch regeneriert werden oder aus Protokollinformationen einer höheren Schicht abgeleitet werden. Dies bedeutet, daß es nicht möglich ist, in einem Ethernet-Netzwerk Integrierte Dienste einzuführen.

Der Token Ring-Standard bietet ein Prioritätsmechanismus, der sowohl das Queuing der Pakete vor der Übertragung als auch den Zugang der Pakete zum geteilten Medium kontrollieren kann. Dieser Prioritätsmechanismus benutzt dabei Bits im AC (Access Control)-Feld und dem FC (Frame Control)-Feld des LLC-Rahmens. Token Ring benutzt auch das Konzept der „reservierten Priorität“ (Reserved Priority), das sich auf den Prioritätswert, mit dem eine Station den Token für die nächste Übertragung reserviert, stützt. Kreist ein freies Token im Ring, dann darf nur eine Station mit einer Access Priority, welches größer oder gleich der reservierten Priorität ist, das Token zur Übertragung im Ring benutzen. Desweiteren ist eine Token Ring-Station theoretisch in der Lage, sofern sie mehrere Rahmen unterschiedlicher Priorität zu versenden hat, anhand der im user_priority-Feld eingetragenen Priorität zu versenden.

Der FDDI-Standard bietet einen Prioritätsmechanismus, der sowohl für die Kontrolle der Paketübertragung als auch für die Kontrolle des Paketzugriffs auf das geteilte Medium benutzt werden kann. Der Prioritätenmechanismus ist dem oben beschriebenen Token Ring ähnlich.

IEEE 802.12 ist ein Standard für ein geteiltes 100 Mbps LAN. Datenpakete werden übermittelt, indem sie entweder das IEEE 802.3 oder der IEEE 802.5 Rahmenformat benutzen. Das MAC-Protokoll heißt dabei „Demand Priority“. Die Hauptcharakteristiken in Bezug auf QoS sind die Unterstützung von zwei Dienstprioritätsstufen, normale Priorität und hohe Priorität, und der Dienstauftrag für jede der beiden. Datenpakete von allen Netzwerkknoten (End-Hosts und Bridges/Switches) werden durch die Benutzung eines einfachen „Round Robin“-Algorithmus bedient.

Wenn das IEEE 802.3 Rahmenformat zur Datenübertragung benutzt wird, dann wird das user_priority im Start-Delimiter des IEEE 802.12 Datenpakets kodiert. Wird jedoch das IEEE 802.5 Rahmenformat benutzt, dann wird das user_priority zusätzlich in den YYY-Bits des FC-Feldes im IEEE 802.5 Paketkopf kodiert.

2.4 Zusammenfassung

Es wurde hier eine Möglichkeit beschrieben, integrierte Dienste im LAN-Bereich einzuführen. Das Reservierungsprotokoll RSVP sorgt dabei auf der Schicht-3 für die Reservierung der zur Verfügung stehenden Ressourcen. Die dem entsprechende, auf die Schicht-2 abgebildete Tätigkeit übernimmt der Subnetzbandbreitenmanager. Um die integrierten Dienste nutzen zu können, ist eine Prioritätenverwaltung auf der Schicht-2 unumgänglich. An der relativen Einfachheit des Ethernet-Rahmenformats jedoch scheitert die Prioritätenverwaltung im Ethernet.

2.5 Bewertung

Da der größte Teil der vorkommenden Netzwerke wohl Ethernet-, bzw. in neuester Zeit auch Fast-Ethernet-Netzwerke sind, kommt dort der Einsatz von integrierten Diensten nicht in Frage. Für die vergleichsweise wenigen Token-Ring-Netzwerke, in denen diese Technik eingesetzt werden könnte, wird die Industrie auf Basis der entstehenden Standards wahrscheinlich keine Soft- und Hardware entwickeln oder diese wird so teuer sein, so daß deren Einsatz unrentabel ist.

Literatur

[FeHu98] P. Ferguson und J. Huston. *Quality of Service*, Kapitel 7: The Integrated Services Architecture. Wiley. 1998.

[GPSS+98] A. Ghanwani, J.W. Pace, V. Srinivasan, A. Smith und M. Seaman. A Framework for Providing Integrated Services Over Shared and Switched IEEE 802 LAN Technologies. Technischer Bericht, IETF, März 1998. Internet Draft draft-ietf-issll-atm-framework-02.txt.

[SmSC97] A. Smith, M. Seaman und E. Crawley. Integrated Service Mappings on IEEE 802 Networks. Technischer Bericht, IETF, November 1997. Internet Draft draft-ietf-issll-is802-svc-mapping-01.txt.

[YHBB+98] R. Yavatar, D. Hoffman, Y. Bernet, F. Baker und M. Speer. SBM (Subnet Bandwidth Manager): A Protocol for RSVP-based Admission Control over IEEE 802-style networks. Technischer Bericht, IETF, März 1998. Internet Draft draft-ietf-issll-is802-sbm-06.txt.

Differentiated Services - oder wie das Internet schnell mit Dienstklassen ausgerüstet wird!

Thorsten Pastoors

Kurzfassung

Differentiated Services sind ein Ansatz, um im Internet gewisse Basisdienste einzuführen. Das Ziel ist es, Mechanismen und Standards zur Verfügung zu stellen, die es erlauben die Bandbreite des Internets unterschiedlichen Benutzern in kontrollierter Weise zur Verfügung zu stellen, um heutigen und zukünftigen Anforderungen besser gerecht zu werden. Da die Dienste der *Differentiated Services* einen aggregierten Datenverkehr und nicht einzelne Datenströme betrachten, können sie ohne Zustandsinformation pro Datenstrom im Inneren des Internets realisiert werden. Zusätzlich sollen sie auf die Hilfe von Signalisierung auf jedem Teilabschnitt verzichten. Diese Arbeit beschreibt zwei verschiedene Ansätze - zum einen die Einführung eines Dienstes für burstartigen Verkehr mit statistisch garantierter Bandbreite und zum anderen die Einführung einer Dienstklasse für Anwendungen mit konstanter Bitrate. Außerdem geht sie auf deren Gemeinsamkeiten und die Unterschiede ein, und stellt eine Architektur zur Integration beider Dienstklassen im Internet vor.

1 Motivation

Das heutige Internet baut auf dem IP-Protokoll auf. Dabei handelt es sich um einen Datagramm-Dienst. Ein Datagramm-Dienst ist ein verbindungsloser Dienst. Die zu übertragenden Daten werden in mehrere Pakete segmentiert und dann einzeln übertragen. Dabei wird jedes Paket völlig unabhängig von den anderen betrachtet. Ein Paket besteht aus einem Kopf und einem Rumpf. Der Kopf enthält verschiedene Kontrollfelder, wie z.B. die Quell-/Zieladressen und den Typ der Daten im Rumpf. Der Rumpf enthält die Daten. Diese Pakete durchlaufen auf dem Weg zum Ziel Router, die alle eingehenden Pakete in einer Warteschlange speichern und sie dann anhand der Zieladresse und ihrer internen Routingtabelle zum nächsten Router senden. Dabei kann es vorkommen, daß eine Warteschlange überläuft und alle neu ankommenden Pakete verworfen werden - ein Stau tritt auf. Dies ist ein Grund, warum man bei IP auch von einem unzuverlässigen Dienst spricht. Weitere Gründe sind erstens, daß Fehler in den Paketen nicht behoben werden, und zweitens, daß die einzelnen Pakete unterschiedliche Wege durch das Netz nehmen können und daher nicht unbedingt in der gleichen Reihenfolge beim Empänger ankommen wie sie beim Sender weggeschickt wurden.

Eine andere Bezeichnung für den IP-Dienst ist auch Best-Effort-Service. Die Eigenschaften des Dienstes werden durch 'gib mir immer soviel Bandbreite wie Du übrig hast und bediene mich so schnell wie Du kannst' sehr gut beschrieben. Der Benutzer kann also nie genau sagen, wieviel Bandbreite seinem Datenstrom vom Netz zur Verfügung gestellt wird und ob die Pakete überhaupt ankommen. Es existiert zwar im IP-Kopf auch ein 8 Bit großes Type-of-Service-Feld (TOS) in dem einzelne Bits für 'geringe Verzögerung', 'hohen Durchsatz', 'hohe Zuverlässigkeit' und für 8 Prioritätsstufen gesetzt werden können, doch gibt es für die Umsetzung dieser Wünsche keinerlei Garantie. Das Feld wird nämlich in der Regel überhaupt nicht ausgewertet.

Heutigen und zukünftigen Anforderungen der Benutzer und ihrer Anwendungen wird das Internet nicht mehr gerecht. Es werden vielmehr unterschiedliche Dienstklassen mit verschiedenen Merkmalen, wie z.B. kurze Ende-zu-Ende-Verzögerung, garantierte Bandbreite und geringer Jitter benötigt. Es ist die Aufgabe der "Differentiated-Services"-Arbeitsgruppe der IETF (Internet Engineering Task Force) Vorschläge für die 'schnelle' Einführung von Dienstklassen auszuarbeiten. Ihre Mitarbeiter haben inzwischen mehrere Vorschläge ausgearbeitet, die im folgenden vorgestellt und diskutiert werden sollen. Ihre Ziele waren hauptsächlich,

- das bestehende Paketformat - und hier vor allem den Paketkopf - unverändert zu lassen, d.h. einzelnen Felder höchstens eine neue Bedeutung zu geben,
- den Weiterleitungspfad einfach zu halten und
- die Komplexität so weit wie möglich an die Grenzen des Netzwerkes zu legen.

2 Grundlagen

Differentiated Services stellen skalierbare Dienste im Internet dar ([NiBl98]). Ihr Vorteil liegt darin, daß in den Zwischensystemen keine Zustandsinformation pro Datenstrom gespeichert werden muß, und daß keine Signalisierung auf jedem Teilabschnitt notwendig ist. Dies wird erreicht, indem in den Routern nur der aggregierte Datenverkehr betrachtet wird. Eine breite Palette von verschiedenen Diensten kann alleine durch die drei folgenden Aktionen realisiert werden:

- Bits im TOS-Feld werden an Netzwerk- und Verwaltungsgrenzen gesetzt, um über bestimmte Bitmuster eine Zuordnung der einzelnen Pakete zu Dienstklassen zu erhalten.
- Diese Bitmuster werden dazu benutzt, um zu entscheiden, wie die Pakete in den Routern behandelt werden.
- An den Netzwerkgrenzen werden die markierten Pakete entsprechend den Eigenschaften der Dienste überwacht und das Bitmuster wird gegebenenfalls neu angepaßt.

Das TOS-Feld der IPv4-Pakete bzw. das Class-Feld der IPv6-Pakete erhält im Modell der Differentiated Services eine neue Bedeutung. Es wird nun als Differentiated-

Services-Feld (DS) benutzt. In ihm können bestimmte Bits gesetzt werden, d.h. das Paket wird markiert. Das Bitmuster gibt dann an, zu welchem Dienst ein Paket gehört, und dient in den Routern als Index auf ein bestimmtes "Per-Hop"-Verhalten. Ein "Per-Hop"-Verhalten ist die Strategie, nach der sich ein Paket beim Weiterleiten in einem Router verhält. Anders gesagt, wird ein "Per-Hop"-Verhalten im DS-Feld kodiert und anhand der Kodierung eine gewisse Weiterleitungsstrategie in den Routern ausgewählt. Um Eigenschaften eines Dienstes vorhersagen zu können, müssen alle Router auf das kodierte "Per-Hop"-Verhalten reagieren, und zwar mit der jeweils gleichen Strategie. Die dabei angewandten Mechanismen können aber durchaus variieren. Jeder Router besitzt einen Satz Parameter, mit denen er steuert, wie die Pakete weitergeleitet werden.

Um *Differentiated Services* realisieren zu können sind außer der Markierung noch weitere Dinge nötig. Folgende funktionale Einheiten werden an den Netzwerkgrenzen und im Netzinneren gebraucht:

- *Klassifizierer*: wählt Pakete anhand bestimmter Felder des Paketkopfes, z.B. Quell-/Zieladresse oder Bitmuster im DS-Feld, aus und leitet sie entsprechend an andere Einheiten weiter.
- *Markierer*: setzt Bits im DS-Feld.
- *Überwacher*: überwacht das Verhalten eines Datenstroms und agiert aufgrund der Verhältnisse zwischen gemessenen und konfigurierten Werten (z.B. Datenrate und Bursts). Bei nicht konformen Paketen wird die Markierung zurückgesetzt oder sie werden verzögert oder sogar verworfen.
- *Shaper*: glättet einen burstartigen Datenstrom durch Verzögern oder Verwerfen.

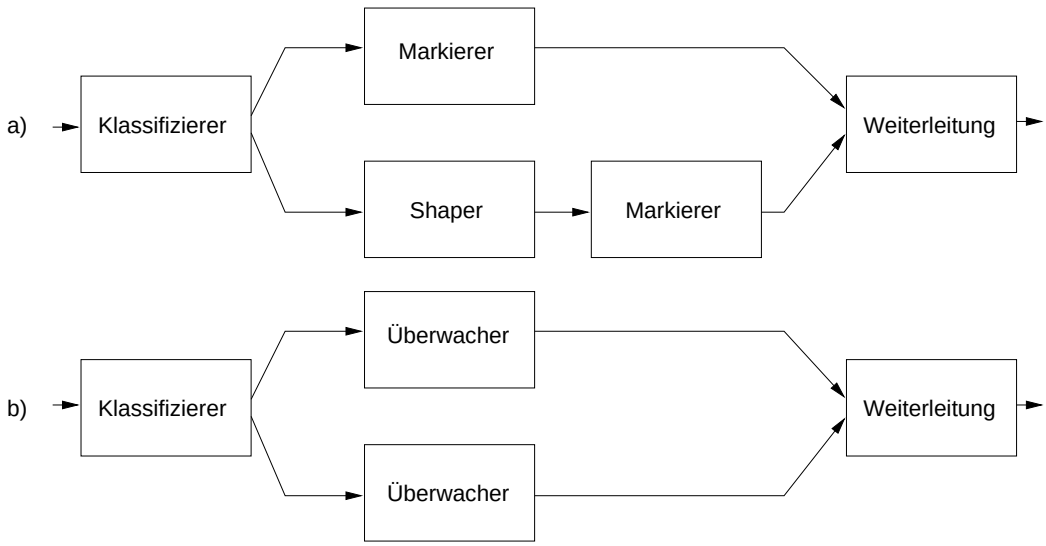

Abbildung 1: Zwei Beispiele für die Anordnung der funktionalen Einheiten

An verschiedenen Stellen im Netzwerk werden die funktionalen Einheiten kombiniert, um Ende-zu-Ende-Dienste, die durch das Verketten von "Per-Hop"-Verhalten entlang des Weges von der Quelle bis zum Ziel realisiert werden, anbieten zu können (siehe

Abbildung 1). Ein Beispiel ist eine Überwachungseinheit, die an der Grenze eines Netzwerks plaziert ist, und aus einem Klassifizierer, der das DS-Feld auswertet, und einem Überwacher besteht (siehe Abbildung 1 b).

Wie bereits oben erwähnt ist eine Signalisierung auf jedem Teilabschnitt nicht notwendig. Trotzdem müssen manche Parameter der funktionalen Einheiten, vor allem an den Netzwerkgrenzen, gesetzt werden, damit ein Netzwerk allen Benutzern die ausgehandelte Dienstgüte garantieren kann. Jedes Verwaltungsgebiet kann dabei die Bereitstellung der Bandbreite der Dienste in seinem Netz individuell regeln. Bei Ende-zu-Ende-Diensten über mehrere Verwaltungsbereiche hinweg müssen diese Bereiche aber Signalisierungsinformation austauschen können, um z.B. die Überwachungseinheiten zu konfigurieren. Das kann mit dem Protokoll RSVP, mit "Bandwidth-Brokern" (siehe Abschnitt 4.2) oder manuell durch die Netzwerkadministratoren erfolgen.

3 Die neuen Dienstklassen

Eine Gemeinsamkeit in den Vorschlägen zu den *Differentiated Services* besteht darin, daß der existierende Best-Effort-Service beibehalten und durch neue Dienstklassen ergänzt wird. Man erwartet, daß weiterhin ein Großteil des Internet-Verkehrs bestmöglich übertragen wird. Zum einen, weil viele, vor allem ältere Anwendungen entsprechend entwickelt wurden, und zum anderen, da nicht jeder Benutzer neue Dienstklassen benötigt oder bereit ist, für diese extra zu bezahlen. Denn es ist klar, daß mit der Einführung neuer Dienstklassen auch die Abrechnungspolitik im Internet ergänzt, oder sogar neu überdacht werden muß, um die Benutzer davon abzuhalten, nur noch die neuen Dienste zu fordern und zu benutzen.

In [ClWr97] und [NiJZ97] werden zwei neue, voneinander verschiedene Dienstklassen vorgestellt. Ein wichtiges Merkmal von beiden besteht darin, daß die bestehenden Anwendungen nicht geändert werden müssen, d.h. die Markierung der IP-Pakete wird beim Netzzugang und nicht in den Endsystemen vorgenommen.

3.1 Der Assured-Service

In [ClWr97] wird der Assured-Service als Ergänzung zum Best-Effort-Service vorgeschlagen. Er ist für burstartigen Verkehr geeignet und wird über ein Profil beschrieben. Ein Profil besteht aus einer mittleren Datenrate und einer maximalen Burstlänge. Z.B. "3 MBit/s mittlere Datenrate und eine Burstlänge von 1000 Byte". Die mittlere Datenrate wird dem Benutzer aber nur statistisch garantiert. Pakete, die konform zum Profil sind, werden nur sehr unwahrscheinlich verworfen. Pakete, die nicht konform sind, werden mit einer höheren Verlustwahrscheinlichkeit ausgeliefert. Konforme und nicht-konforme Pakete erfahren aber in den Routern keine Reihenfolgevertauschung untereinander. Der Dienst sichert dem Benutzer zu, daß die konformen (als IN markierten) Pakete den Empfänger mit einer hohen Wahrscheinlichkeit erreichen (aber nicht zu 100%, da die mittlere Bandbreite nur statistisch garantiert wird) . Die nicht markierten, nicht konformen Pakete (OUT-Pakete) desselben Datenstroms haben keine Privilegien. Sie werden bei Stausituationen wie normale Best-Effort-Pakete behandelt und können daher in den Routern verworfen werden.

Ein IP-Datenstrom des Assured-Service besteht also aus einzelnen Paketen, die entweder als IN oder OUT markiert sind. Diese Markierung - ein Bit genügt hier - soll im DS-Feld des IP-Kopfes geschehen. Da bei einem Best-Effort-Paket keine Bits gesetzt werden, soll die Kodierung so gewählt werden, daß ein als OUT markiertes Paket des Assured-Service nicht von einem Best-Effort-Paket zu unterscheiden ist. Die Markierung dient nun in den Routern dazu, OUT-Pakete - also auch Pakete des Best-Effort-Service - bei Stausituationen mit einer höherer Wahrscheinlichkeit als IN-Pakete zu verwerfen. Daher kann der Nutzer davon ausgehen, daß als IN markierten Pakete selbst bei Stau ziemlich sicher beim Empfänger ankommen und daß OUT-Pakete ankommen, falls kein Stau auftritt. Durch die Möglichkeit, einen Teil der Daten so gut wie sicher und einen anderen Teil 'auf gut Glück' zum jeweiligen Empfänger zu schicken und daher die Datenrate überzubelegen, ist dieses Dienstmodell besonders für burstartigen Verkehr geeignet.

Aber wie wird entschieden, welche Pakete als IN oder OUT markiert werden? Jedem Benutzer ist ein Dienstprofil zugeordnet, und sein Datenstrom wird dem Profil entsprechend markiert. Die Markierung wird von einer eigenen logischen Komponente, dem sogenannten Profile-Meter ausgeführt. Die Plazierung des Profile-Meter kann überall dort erfolgen, wo ein Benutzer-/Erbringer-Verhältnis besteht, also z.B. zwischen Endsystem und LAN, zwischen LAN und Internet Service Provider (ISP) oder zwischen ISP und ISP. Ein Datenstrom kann von seiner Quelle bis zu seinem Ziel durchaus mehrere Profile-Meter durchlaufen. Dabei unterscheidet ein Profile-Meter im Netzinnern an seinem Eingang keine unterschiedlichen Datenströme, z.B. anhand des IP-Quell/-Zieladreßpaares, sondern betrachtet den aggregierten Datenverkehr und behandelt alle eingehenden Pakete als zum selben Datenstrom gehörend und markiert diesen entsprechend dem eingestellten Profil. Der erste Profile-Meter auf dem Pfad vom Sender zum Empfänger jedoch enthält einen Klassifizierer, der einzelne Datenströme unterscheidet, und so eine feinere Abbildung der Dienstprofile auf Datenströme bietet.

Um tatsächlich eine Änderung gegenüber dem Best-Effort-Service zu erreichen, muß es auch einen Mechanismus in den Routern geben, der die Pakete anhand des IN-/OUT-Bits unterschiedlich behandelt. Im heutigen Internet wird in den Routern zur Zeit der RED-Algorithmus (Random Early Detection) eingeführt [FlJa93]. Er geht folgendermaßen vor: es wird nicht gewartet bis die Warteschlange in einem Router voll ist, um dann alle neu ankommenden Pakete zu verwerfen, sondern ab einer bestimmten Füllmenge werden neu ankommende Pakete mit einer kleinen aber zunehmenden Wahrscheinlichkeit verworfen. Dies führt über einen längeren Zeitraum betrachtet zu einem besseren Gesamtverhalten. Der neue Mechanismus (RIO, RED with IN and OUT) in den Routern für den Assured-Service ist eine Erweiterung des RED-Algorithmus. Wie bisher werden alle ankommenden Pakete, egal ob als IN oder OUT markiert, in einer einzigen Warteschlange gesammelt. Auch die Pakete, die zu einem Best-Effort-Service gehören, landen mit den Paketen des Assured-Services in der selben Schlange (Nach Definition sind die Best-Effort-Pakete und OUT-Pakete gleich markiert, s.o.). Für diese Warteschlange gibt es jetzt zwei RED-Algorithmen, je einer für die IN- bzw. OUT-Pakete. Dabei werden die OUT-Pakete schon bei einem viel kleineren Füllstand und auch mit einer höheren Wahrscheinlichkeit verworfen als die IN-Pakete. Die Parameter des RIO-Algorithmus sind im Idealfall so gewählt, daß zu keinem Zeitpunkt IN-Pakete verworfen werden müssen, d.h. daß der Nutzer davon ausgehen kann, daß seine IN-Pakete mit hoher Wahrscheinlichkeit beim Empfänger ankommen.

3.2 Der Premium-Service

Der Premium-Service [NiJZ97] ist für alle Anwendungen geeignet, die sich darauf verlassen, daß ihnen zu jedem Zeitpunkt eine feste Bandbreite exklusiv zur Verfügung steht. Es ist ein Dienst mit konstanter Bitrate und kurzer Verzögerung und wird über die gewünschte maximale Bitrate definiert. Diese darf vom Benutzer zu keinem Zeitpunkt überstiegen werden. Natürlich darf er aber weniger senden als erlaubt. In diesem Fall steht die übrige Bandbreite dann dem Best-Effort-Service zur Verfügung. Die Zugehörigkeit eines Paketes zum Premium-Service soll dabei durch ein Bit im DS-Feld angezeigt werden. Dieser Dienst kann beispielsweise dazu genutzt werden, virtuelle gemietete Leitungen zu realisieren.

Es werden auch hier die speziellen Komponenten aus Abschnitt 2 benötigt. Den Markierer, um das Premium-Bit in den Paketen zu setzen und den Shaper, um den Datenstrom zu glätten. Das Glätten ist wichtig, um Bursts zu vermeiden, die dazu führen könnten, daß die vereinbarte Datenrate überstiegen wird. Diese beiden Aufgaben soll der erste Router übernehmen, den der Datenstrom nach Verlassen des Endsystems passiert (der sogenannte *First-Hop-Router*, der aus einem Klassifizierer, einem Shaper und einem Markierer besteht). Um Bandbreite für einen Premium-Service bereitstellen zu können, müssen einzelne Datenströme, die verschiedenen Dienstprofilen angehören, auch unterscheidbar sein. Diese Unterscheidung sollte anhand der Quell- und Zieladresse oder irgendeiner anderen Kombination der Felder des IP-Paketkopfes getroffen werden. Es ist aber auch möglich z.B. die Portnummern in TCP oder UDP mit einzubeziehen, um nicht nur Bandbreite für die Kommunikation von einem Endsystem zu einem anderen zu reservieren, sondern sogar für die Kommunikation von zwei Prozessen auf unterschiedlichen Endsystemen. First-Hop-Router müssen die einzelnen Datenströme klassifizieren, glätten und das Premium-Bit setzen.

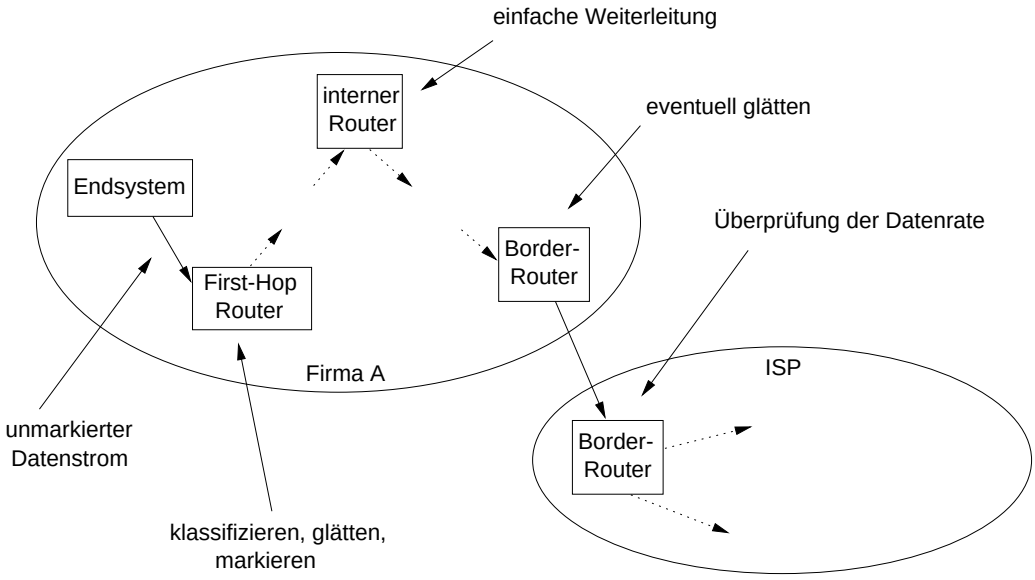

Abbildung 2: Der Pfad eines Datenstroms

Ein kritischer Punkt auf dem Pfad von der Quelle zum Ziel stellt die Grenze zwischen zwei Gebieten, die von unterschiedlichen Betreibern verwaltet werden, dar. Border-Router sind alle die Router, die an den Netzwerkgrenzen plaziert sind, und sowohl eine Verbindung zu einem Router innerhalb als auch eine Verbindung zu einem Router

außerhalb des selben Verwaltungsbereichs haben. Das Problem liegt darin, daß das Gebiet, in das der Verkehr weitergeleitet wird, in seinem Border-Router überprüfen muß, ob der Verkehr auch innerhalb der ausgehandelten Rate liegt. Alle Pakete, die diese Rate übersteigen, werden verworfen. Die Überprüfung ist wichtig, damit man den anderen Benutzern ihre ausgehandelte Bandbreite garantieren kann. Denn im Gegensatz zum Assured-Service ist die Bandbreite beim Premium-Service garantiert und darf nicht überbelegt werden. Der Router des anderen Gebietes, aus dem der Verkehr kommt, kann den Datenverkehr, der das Gebiet verläßt, noch einmal glätten, um eventuelle Bursts, die durch Verzögerungen in den internen Routern entstanden sind, auszugleichen, und damit dafür zu sorgen, daß die ausgehandelte Rate nicht überbelegt wird. Abbildung 2 zeigt einen Teil des Pfades vom Empfänger zum Sender, die Router, die passiert werden, und deren Aufgaben.

4 Ein Rahmenwerk zur Integration des Assured- und des Premium-Services

Die beiden neuen Dienstklassen, die in den Abschnitten 3.1 und 3.2 vorgestellt wurden sind orthogonal zueinander. Der Assured-Service bietet seinem Benutzer eine statistisch zugesicherte Bandbreite, den IN-Verkehr, die er aber auch überbelegen darf (OUT-Verkehr). Daher ist er für burstartigen Verkehr, von dem ein bestimmter Anteil unbedingt den Empfänger erreichen sollte, geeignet. Der Premium-Service bietet zu jeder Zeit eine feste Bandbreite, die aber nicht überschritten werden darf. Trotzdem sind beide Dienste in der Art und der Plazierung ihrer funktionalen Einheiten ziemlich ähnlich. [NiJZ97] schlägt deshalb eine Zwei-Bit-Architektur vor, welche die drei Dienste, Best-Effort, Assured und Premium, in sich vereinigt. Die zwei Bit sind dabei zur Anzeige der jeweiligen Dienstklasse bestimmt. Abbildung 3 zeigt eine mögliche Kodierung der zwei Bit im DS-Feld des IP-Paketkopfes, die aber erst in späteren Standardisierungsprozessen festgelegt werden soll. Ein wichtiger Punkt ist, daß in der Architektur keine Ände-

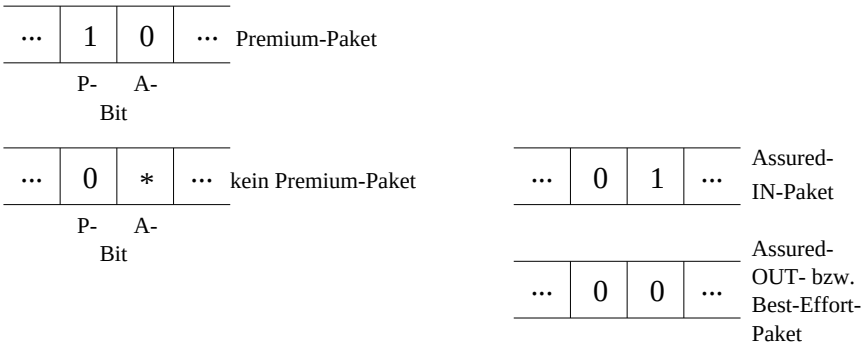

Abbildung 3: Eine Kodierungsmöglichkeit für Premium-, Assured- und Best-Effort-Service

rungen in den Endsystemen nötig sind. Die notwendigen Markierungen - das Setzen der P- und A-Bits - werden im Netzwerk, im First-Hop-Router, gemacht (siehe auch Abbildung 2).

4.1 Die neuen funktionalen Einheiten im Netz

Die Weiterleitungseinheit für den Datenfluß in einem Router der Zwei-Bit-Architektur besteht aus zwei Prioritätswarteschlangen. Einer hochpriorisierten für den Premium- und einer niederpriorisierten für den Assured-/Best-Effort-Verkehr. Die niederpriorisierte ist wie in Abschnitt 3.1 beschrieben mit dem RIO-Algorithmus ausgestattet. Die Pakete werden anhand des P-Bits klassifiziert und in einer der beiden Warteschlangen gespeichert.

Um die weiteren Mechanismen und Einheiten im Netz besser erklären zu können, wird anhand der Reihenfolge der Einheiten auf dem Pfad vorgegangen, den ein Datenstrom von einer Quelle zu einem Ziel durchläuft (siehe auch Abbildung 2). Als erstes durchläuft der Strom den First-Hop-Router. Zu seinen Funktionen gehören, die Pakete entsprechend ihrem Dienst zu markieren und sie weiterzuleiten. Zuerst müssen alle P- und A-Bits zurückgesetzt werden. Dies ist wichtig, um garantieren zu können, daß alle Pakete richtig markiert werden. Außerdem muß man sichergehen, daß kein Benutzer seine Pakete schon im Endsystem markiert, um eine bessere Dienstqualität zu bekommen, als er mit dem Netz vereinbart hat. Als nächstes werden die Pakete anhand einer Kombination der Paketkopffelder, z.B. des Quell-/Zieladreßpaares, klassifiziert, d.h. einem bestimmten Datenstrom zugeordnet. Für diesen Strom weiß der First-Hop-Router, ob er zu einem Premium- oder einem Assured-Service gehört. Entsprechend der Klassifikation werden die Pakete dann an den richtigen Markierer geschickt, der mit den im Profil vorgegebenen Parametern den Strom markiert. Die Parameter, beim Premium-Service die maximale Rate und beim Assured-Service die mittlere Rate und die Burstlänge, beschreiben den Datenstrom und müssen vor dem ersten Datenaustausch dem First-Hop-Router mittels Signalisierung übermittelt oder manuell vom Netzwerkadministrator gesetzt werden. Alle Pakete, die nicht klassifiziert

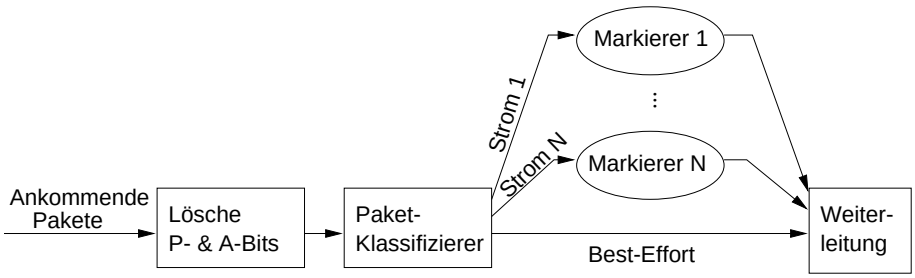

Abbildung 4: Funktionen des First-Hop Routers

werden konnten, sind Best-Effort-Pakete und durch das anfängliche Zurücksetzen der Bits schon richtig markiert (siehe Abbildung 3). Sie werden direkt an den Weiterleitungsblock geschickt, die anderen Pakete erst, nachdem sie die Markierer durchlaufen haben (siehe Abbildung 4).

Der interne Aufbau der Markierer kann für beide Dienstklassen, den Assured- und den Premium-Service, durch Token-Buckets realisiert werden. Ein Token-Bucket wird mit Token gefüllt. Die Füllrate entspricht dabei der Rate aus dem Benutzerprofil. Die Größe des Bucket muß man beim Assured-Service auf die maximale Burstlänge begrenzen, bei Premium sollte sie ein bis zwei Token betragen. Beim Assured-Service wird dann das A-Bit eines Paketes auf 1 (das bedeutet IN) gesetzt, falls ein Token vorhanden ist, und bleibt auf 0 (das bedeutet OUT), falls keiner vorhanden ist. Beim

Premium-Service wird bei Paketen, für die ein Token bereitsteht, das P-Bit auf 1 gesetzt. Pakete, für die kein Token bereitsteht, müssen solange verzögert werden, bis ein Token vorhanden ist. Tritt in dem Premium-Datenstrom ein Burst auf, kann es dazu kommen, daß die Warteschlange, in der die Premium-Pakete zwischengespeichert werden, überläuft. In diesem Fall werden die neu ankommenden Pakete verworfen. Der Paketstrom, der den First-Hop-Router verläßt, besteht aus einem geglätteten Paketstrom mit gesetztem P-Bit gemischt mit einem nicht geglätteten Best-Effort-Strom in dem manche Pakete ein gesetztes A-Bit haben. Es kann übrigens nicht vorkommen, daß der Datenstrom aus dem First-Hop-Router nur aus Premium-Paketen besteht und der Best-Effort-Service gänzlich unterdrückt wurde, da sowohl die Bursts als auch die Bandbreite des Premium-Service überwacht werden. Der Assured-Service kann im Gegensatz dazu den Best-Effort-Service ganz unterdrücken. Denn auch wenn die Bereitstellung von Bandbreite konservativ geschieht, um das Verhungern des unmarkierten Verkehrs zu vermeiden, können Bursts des Assured-Verkehrs bei Stausituationen den Best-Effort-Verkehr in einer Warteschlange verdrängen.

Die nächsten Knoten auf dem Pfad von der Quelle zum Ziel sind die internen Router (Abbildung 2), die völlig im Inneren eines Netzes liegen, das von einem einzigen Netzbetreiber verwaltet wird. Die einzige funktionale Einheit, die diese Router benötigen, ist die oben beschriebene Weiterleitungseinheit.

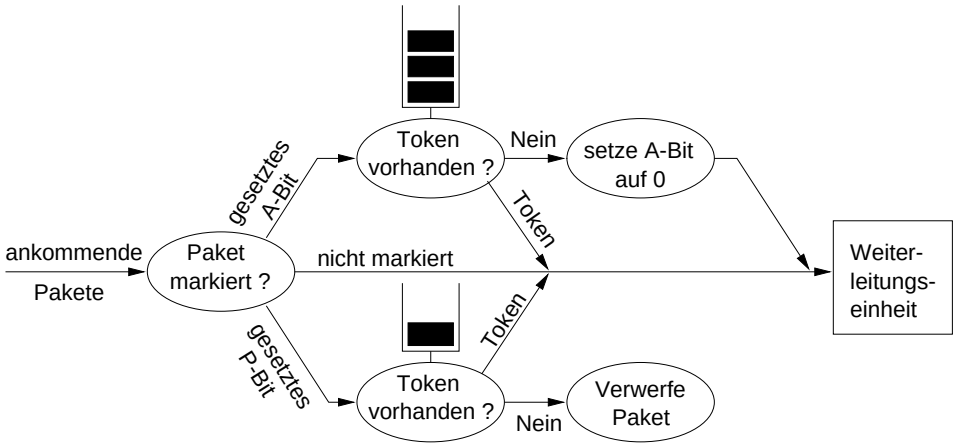

Abbildung 5: Überwachungseinheit am Eingang der Border-Router

Eine besondere Betrachtung verlangt der Übergang zwischen zwei Netzen, die von verschiedenen Organisationen verwaltet werden. Die Router, die eine Verbindung zu einem benachbarten Netz haben, werden Border-Router genannt. Um die Eigenschaften der Dienste Assured und Premium auch dann beibehalten zu können, wenn eine Kommunikation über verschiedene Netze erfolgt, muß zwischen je zwei benachbarten Netzverwaltungseinheiten ein gegenseitiger Vertrag abgeschlossen werden. Dieser Vertrag muß eine maximale Rate für den gesamten von einem Netz zum anderen fließenden Premium-, und eine mittlere Rate und Burstlänge für den Assured-Verkehr beinhalten. Die Werte können dabei durchaus für beide Richtungen unterschiedlich sein. Wichtig ist hierbei, daß insgesamt nur drei unterschiedliche Verkehrsarten von einem Netz in das andere fließen, d.h. alle Premium-Ströme werden als ein aggregierter Verkehr betrachtet, bei den beiden anderen Diensten ist das ebenso. Die Unterscheidung von Verkehrstypen wird alleine anhand des P-/A-Bitmusters getroffen (Abbildung 5). Am Eingang eines Netzes, d.h. in einem Border-Router muß sichergestellt werden, daß die beiden neuen "Differentiated-Service"-Verkehrstypen im Übereinstimmung mit den vereinbarten

Raten sind. Übersteigt der Premium-Strom die ausgehandelte Rate, werden alle überzähligen Premium-Pakete verworfen. Übersteigt der Assured-Strom die ausgehandelte Rate, wird bei allen überzähligen Assured-Paketen das A-Bit auf 0 zurückgesetzt. Diese Überwachungseinheit kann sehr leicht mit Hilfe des Token-Bucket-Algorithmus modelliert werden (siehe Abbildung 5). Am Ausgang eines Netzes (siehe Abbildung 2) könnte der ausfließende Premium-Strom noch einmal geglättet werden, um zu garantieren, daß keine Bursts auftreten, die den ausgehandelten Vertrag übersteigen. Das Glätten kann, wie oben bei den Markieren beschrieben, gemacht werden.

Es werden zusätzlich Mechanismen benötigt, um die spezifizierten Profile an den First-Hop-Router zu übermitteln. Spezifizierte Profile bestehen aus der Datenrate, der Burstlänge und aus dem Typ des Dienstes, also Premium oder Assured. Die Übermittlung kann z.B. mittels den Protokollen RSVP (ReSerVation Protocol), SNMP (Simple Network Management Protocol) oder manuell durch den Netzwerkadministrator erfolgen. Ein weiteres Problem ist die Zuteilung der Bandbreite des markierten Verkehrs im Internet. Die Zuteilung kann dabei statisch oder dynamisch erfolgen. Dabei deckt 'dynamisch' den Bereich von telefonischer oder per e-Mail übermittelter Anfrage bis hin zu einem eigenen Signalisierungsmodell ab. Für beide Zuteilungsarten besteht Bedarf. Z.B. kann ein virtuelles Kaufhaus statisch für jeden Kunden eine feste Bandbreite in oder aus seinem Web-Server bereitgestellt haben, um den Kunden eine gleichbleibende Geschwindigkeit beim Browsen zu bieten. Aber es muß auch die Möglichkeit der dynamischen Zuteilung geben, um auf spezielle Ereignisse reagieren zu können, wie z.B. aktuelle Benachrichtigungen.

4.2 Eine Architektur für Bandbreiten-Zuteilung

In [NiJZ97] wird auch eine Architektur zur Bandbreiten-Zuteilung vorgeschlagen. Die Grundidee liegt darin, Agenten, die Wissen über die Prioritäten und die Politik der Netzbetreiberorganisation besitzen, die Bandbreiten-Zuteilung vornehmen zu lassen. Es werden absichtlich zentrale Einrichtungen damit beauftragt, um den Gesamtüberblick bewahren zu können, der bei einzelnen Benutzern leicht verloren gehen kann. Die Agenten heißen Bandwidth-Broker. Aufgrund der Erfahrungen, daß multilaterale Verträge selten eingehalten werden, setzt die hier vorgestellte Architektur darauf, daß Ende-zu-Ende-Dienste auf mehreren bilateralen Abkommen aufgebaut werden. Ein weiterer Vorteil der Bandwidth-Broker besteht darin, daß die benötigte Zustandsinformation gering und zentral gehalten werden kann. Die beiden Aufgaben eines Bandwidth-Broker sind zum einen die Aufteilung der Bandbreite für den markierten Datenverkehr innerhalb seines Zuständigkeitsbereichs und die entsprechende Konfiguration der First-Hop-Router, und zum anderen die Kommunikation mit benachbarten Bandwidth-Brokern.

Der Grund, warum nur Bandwidth-Broker die First-Hop-Router konfigurieren dürfen, liegt daran, daß man ein sicheres System haben möchte. Die Bandbreiten-Zuteilung mit Hilfe von Bandwidth-Brokern muß nicht vollständig automatisiert sein, vor allem nicht in der Einführungsphase der Zwei-Bit-Architektur. Es ist möglich, daß durch den manuellen Eingriff von außen Parameter gesetzt werden. Wenn eine Zuteilung für einen bestimmten Datenstrom gewünscht wird, wird eine Anfrage an den Bandwidth-Broker X geschickt, die den Diensttyp, d.h. Premium oder Assured, die Datenrate, die maximale Burstlänge und das Quell-/Zieladreßpaar beinhaltet. Die Anfrage kann von einem

Benutzer oder einem benachbarten Bandwidth-Broker Y kommen. Der Bandwidth-Broker X prüft die Identität des Anfragenden und ob die geforderte Bandbreite frei ist. Werden die beiden Tests bestanden, wird die verfügbare Bandbreite entsprechend reduziert. Enthält die Datenstromspezifikation eine Zieladresse, die außerhalb des Zuständigkeitsbereichs des Bandwidth-Broker X liegt, muß der entsprechende benachbarte Bandwidth-Broker Y informiert werden. Der Bandwidth-Broker X konfiguriert den zuständigen First-Hop-Router mit den Parametern des Datenstroms, z.B. anhand welcher Kopffelder klassifiziert werden soll, und mit welcher Rate und welcher maximalen Burstlänge gesendet werden soll. Der Bandwidth-Broker Y ist dafür verantwortlich, daß der entsprechende Border-Router so konfiguriert wird, daß der Datenstrom passieren darf. Außerdem muß er alle zusätzlichen Konfigurationen innerhalb und falls nötigt auch alle Anfragen außerhalb seines Netzes vornehmen. Außerhalb dann, falls sein Netz nur als Transitnetz benutzt wird, d.h. weder die Quell- noch die Zieladresse in seinem Zuständigkeitsbereich liegen.

5 Schlußbemerkungen

Bei Sicherheitsfragen taucht mit den *Differentiated Services* ein neues Problem auf. Es kann nun nämlich Bandbreite gestohlen werden, wenn Signalisierungsdaten gefälscht oder vorgetäuscht werden, oder wenn versucht wird durch 'übermarkieren' mehr Pakete als erlaubt in das nächste Netzwerk zu senden. Deshalb müssen spezielle Vorkehrungen getroffen werden, um die Signalisierung zu sichern und um alle Datenströme, die in ein Netzwerk fließen, zu überwachen.

Wie schon angedeutet, muß auch über eine neue Abrechnungspolitik nachgedacht werden, denn mit der Zuteilung von Bandbreite wird dem Best-Effort-Service eine schlechtere Voraussetzung geboten. Daher müssen die Benutzer des Assured- bzw. Premium-Service mehr bezahlen, um dem Benutzer des Best-Effort-Service einen Grund zu liefern, bei dem weniger attraktiven Dienst zu bleiben.

Wichtig ist, daß zuerst Testumgebungen aufgebaut werden müssen, um Erfahrung zu sammeln und um die Akzeptanz der Dienste zu überprüfen. Denn trotz vieler theoretischer Überlegungen kann man nicht voraussagen, wie die Benutzer auf die neuen Dienste reagieren werden. Die in dieser Arbeit vorgestellte Architektur stellt nur eine Möglichkeit dar, viele andere sind denkbar und werden zur Zeit auch in einer mailing-list der "Differentiated-Services"-Arbeitsgruppe heftig diskutiert. Auch sind bisher wichtige Fragen, wie die zur Integration von Multicast und zur empfängerbasierten Bandbreitenreservierung noch nicht beantwortet.

Literatur

[ClWr97] D. Clark und J. Wroclawski. An Approach to Service Allocation in the Internet. *Internet Draft „draft-clark-diff-svc-alloc-00.txt“*, Juli 1997.

[FlJa93] S. Floyd und V. Jacobson. Random Early Detection Gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking*, August 1993.

[NiBl98] K. Nichols und S. Blake. Differentiated Services Operational Model and Definition. *Internet Draft „draft-nichols-dsopdef-00.txt“*, Februar 1998.

[NiJZ97] K. Nichols, V. Jacobson und L. Zhang. A Two-Bit Differentiated Services Architecture for the Internet. *Internet Draft „draft-nichols-diff-svc-arch-00.pdf“*, November 1997.