10-Gigabit-Ethernet-Triple-Play-Switches – Moderne LAN-Switches müssen nicht nur Bandbreite satt liefern, sie sollen auch den Anforderungen der verschiedenen IP-Anwendungen gewachsen sein. Mechanismen wie die Datenpriorisierung und das Bandbreitenmanagement machen dies möglich.

Wettstreit der Multitalente

Moderne Kommunikationsnetze stellen aktive Komponenten vor immer anspruchsvollere Aufgaben. Mit dem Siegeszug der IP-Telefonie hält die erste Echtzeitanforderungen stellende Anwendung flächendecken Einzug in die Welt der Ethernet-basierenden Unternehmensnetze. Und auch die Integration von Video-over-IP steht in vielen Bereichen schon unmittelbar bevor. Dann sind da noch die klassischen Datenapplikationen, und deren Anforderungen an das Netzwerk werden auch immer anspruchsvoller.

Um diesen Herausforderungen zu begegnen haben die Switch-Hersteller ihre Systeme zügig weiter entwickelt. Mit der Einführung von 10-Gigabit-Ethernet ist das gute alte Ethernet nochmals um den Faktor 10 schneller geworden. Und Mechanismen wie die Datenpriorisierung und das Bandbreitenmanagement sollen für eine intelligente Ausnutzung der zur Verfügung gestellten Ressourcen im gesamten Unternehmensnetz sorgen.

Störfaktoren im LAN

Die Übertragung von einem Endpunkt im Netzwerk zum anderen erfordert eine gewisse Laufzeit. Dabei gibt es zunächst einen festen Teil, der durch die Auswahl der zu verwendenden Codecs, also der Sprach-Digitalisierungs-Algorithmen, und der Netzwerkkomponenten beeinflussbar und ziemlich gut berechenbar ist. Dieser wird durch die Zeit, die die Kodierungsalgorithmen an beiden Endpunkten benötigen, durch die Hardware-Durchlaufzeit auf den beteiligten End- und Knotenpunkten und durch die rein physikalischen Übertragungsgeschwindigkeiten der verschiedenen Medien über bestimmte Entfernungen festgelegt. Zusätzlich entstehen Verzögerungen beispielsweise durch volle Warteschlangen bei Überlast oder durch die eventuelle Wahl alternativer Routen zum Zielpunkt. Die beiden letzteren Punkte können auch die Ursache für zwei andere Übertragungsfehler sein. Beim sogenannten Jitter treffen Pakete, die in regelmäßigen Intervallen in das Netz geschickt werden, in unregelmäßigen Abständen beim Empfänger ein. Ist bei isochromem Datenverkehr wie der IP-Sprachübertragung ein Paket zu schnell am Ziel, dann kann es für die Ausgabe noch nicht verwendet werden. Kommt es dagegen später als erwartet, können Lücken in der Sprachwiedergabe entstehen. Diesem Jitter kann man durch den Einsatz eines Jitter-Buffers entgegenwirken, der Pakete aus dem Netz entgegennimmt und verzögert aber gleichmäßig an die Dekodiereinheit weitergibt. Natürlich erhöht sich dadurch auch der Delay. Treffen die Pakete beim Empfänger in einer anderen Reihenfolge ein, als vom Sender beabsichtigt, spricht man von einem Sequence-Error. Häufigste Ursache hierfür ist, dass einige zu einer Übertragung gehörende Pakete auf Grund einer Überlast regeroutet werden und so ihr Ziel auf einem anderen, möglicherweise langsameren Weg erreichen. Wie gut solche Fehler in der Paket-Reihenfolge ausgeglichen beziehungsweise überspielt werden können, hängt in erster Linie von der Länge des Jitter-Buffers ab.

Gehen bei der Übertragung Pakete ganz verloren (Packet-Loss), dann sind die Auswirkungen um so größer, je höher die Anzahl der Sprachdaten-Bytes in dem verlorenen Paket war und je stärker der Codec komprimiert. Gehen mehrere aufeinander folgende Pakete verloren (Consecutive-Packet-Loss), sind die Auswirkungen auf die Sprachqualität deutlich stärker, als wenn die Verluste gleichmäßig streuen. Diese Verlustart tritt überwiegend in Burst-Situationen auf. Die Ursache für Packet-Loss liegt häufig darin, dass auf dem Übertragungsweg Bandbreitenengpässe auftreten und durch länger dauernde Bursts Warteschlangen überlaufen, weshalb dann Pakete verworfen werden, oder Pakete in den Warteschlangen so weit verzögert werden, dass sie nicht mehr über den Jitter-Buffer sinnvoll versendet werden können. Werden die Jitter-Buffer sehr groß ausgelegt, um entsprechende Netzwerkfehler wie Sequence-Errors oder Jitter auszugleichen, führt diese Technik selbst zu einer zu großen Verzögerung, die dann gleichfalls die Echtzeitkommunikation stört. Jitter-Buffer verringern also Probleme, die durch Jitter und Squenz-Error entstehen können, erzeugen aber ihrerseits zusätzliche Delay-Zeiten. Gute Endgeräte verwalten den Jitter-Buffer daher dynamisch.

Bei entsprechender Überlast im Netz sind Datenverluste ganz normal, jedoch sollen sie durch die Priorisierungsmechanismen in der Regel auf nicht echtzeitfähige Applikationen verlagert werden. Arbeitet diese Priorisierung nicht wie vorgesehen, kommt es auch im Bereich der hoch priorisierten Sprachdaten zu Verlusten. Für eine realitätsnahe und aussagefähige Auswertung der Messergebnisse ist es darüber hinaus entscheidend zu wissen, welche Framegrößen in welchen Verteilungen in realen Netzwerken vorkommen. Analysen der Verteilung der Framegrößen, beispielsweise für das NCI-Backbone oder von den Applikationen her typischer Business-DSL-Links haben ergeben, dass rund 50

Prozent aller Datenrahmen in realen Netzwerken 64 Byte groß sind. Die übrigen rund 50 Prozent der zu transportierenden Datenrahmen streuen über alle Rahmengrößen von 128 bis 1518 Byte.

Für Echtzeit-Anwendungen wie die Voice- oder Video-Übertragung ist zunächst das Datenverlustverhalten von entscheidender Bedeutung. Ab 5 Prozent Verlust ist je nach Codec mit deutlicher Verschlechterung der Übertragungsqualität zu rechnen, 10 Prozent führen zu einer massiven Beeinträchtigung, ab 20 Prozent Datenverlust ist beispielsweise die Telefonie definitiv nicht mehr möglich. So verringert sich der R-Wert für die Sprachqualität gemäß E-Modell nach ITU G.107 schon bei 10 Prozent Datenverlust um je nach Codec 25 bis weit über 40 Punkte, also Werte, die massive Probleme im Telefoniebereich sehr wahrscheinlich machen. Auf Grund ihrer Bedeutung für die Übertragungsqualität haben wir daher das Datenrahmenverlustverhalten als primäres K.O.-Kriterium für unsere Tests definiert. Die Parameter Latency und Jitter – die wir standardmäßig ebenfalls messen – sind dann für die genauere Diagnose und weitere Analyse im Einzelfall wichtig. Sind jedoch die Datenverlustraten von Hause aus schon zu hoch, können gute Werte für Latency und Jitter die Sprachqualität auch nicht mehr retten. Dafür, dass es zu solchen massiven Datenverlusten im Ethernet-LAN erst gar nicht kommt, sollen entsprechend gut funktionierende Priorisierungsmechanismen sorgen. Sie tun dies aber durchaus nicht immer, wie die Erfahrung aus vorhergehenden Tests zeigt.

Qualitätssicherung im LAN

Ethernet-basierte LANs ermöglichen eine kostengünstige durchgehende Netzwerk-Lösung vom Backbone-Bereich bis an die Endgeräte am Arbeitsplatz und sind nicht zuletzt aus diesem Grund weltweiter Standard in praktisch allen Unternehmen und Institutionen. Allerdings bietet das Ethernet auf Layer-2 und -3 nicht die selbe Übertragungsqualität wie von Hause aus echtzeitfähige Technologien, beispielsweise ATM. Das Ethernet-Protokoll ist zwar einfacher und arbeitet mit geringerem Overhead, die Übertragungen erfolgen aber ohne vorherigen Aufbau einer Verbindung oder die Aushandlung der Qualität der Übertragungsstrecke von Endpunkt zu Endpunkt. Für alle Applikationen wird nur eine Best-Effort-Behandlung – so gut, wie eben möglich – bereitgestellt, unabhängig von deren tatsächlichen Anforderungen oder den Anforderungen der Nutzer. Die Absicherung einer Verbindung erfolgt – wenn überhaupt – erst in Protokollen höherer Ebenen, wie dem TCP. In Ethernet-Netzen und unter Verwendung des TCP/IP-Protokolls – und damit auch im Internet, Intranet oder Extranet – gibt es also keine garantierten Verbindungseigenschaften. Deshalb ist auch die Implementierung von Quality-of-Service oder kurz QoS, wie man es von ATM kennt, nicht möglich. Trotzdem versuchen die Ethernet-Produktentwickler durch Priorisierung und Reservierung von Ressourcen auch in IP-Netzen verschiedene Serviceklassen, die Class-of-Service oder CoS, zu etablieren. Allgemein sind zwei Wege zu unterscheiden, Service-Qualitäten zu realisieren, zum einen über die Reservierung von Netzwerkressourcen, die Resource-Reservation, und zum anderen über eine bevorzugte Behandlung be-

stimmter Pakete bei deren Weiterleitung, die Daten-Priorisierung.

Grundlage für letztere ist die Entscheidung, welches Paket welche Priorität besitzt. Diese Entscheidung kann auf Grundlage der generell zur Verfügung stehenden Informationen aus den Headern der Ebenen 2, 3 oder 4 erfolgen. So ist es möglich, den Verkehr beispielsweise hinsichtlich der Quell- und Zieladressen (MAC oder IP) oder der Protokoll- und Portnummern einzuteilen, natürlich in Abhängigkeit davon, bis in welche Ebene das Netzwerkgerät die Protokoll-Header analysieren kann. Geht man einen Schritt weiter, kann man in den Protokoll-Headern der verschiedenen Ebenen bestimmte Bits gezielt setzen und so die Zugehörigkeit eines Paketes zu einer Prioritätsklasse kennzeichnen.

Die Hierarchie der Prioritätsentscheidungen auf den verschiedenen Layern, die ja durchaus widersprüchlich sein kann, ist für jeden Switch intern gelistet und entweder frei konfigurierbar oder fest vorgegeben. Zu beachten ist auch, dass Layer-2-Priorisierungen auf dem Weg durch ein LAN in der Regel verloren gehen, sobald sie auf Layer-3 geswitched beziehungsweise geroutet werden. Die Konfiguration des aktiven Netzwerks, das intelligent die Priorisierungsmechanismen nutzen soll, ist daher gerade in heterogenen Umfeldern nicht gerade trivial. Häufig wird der IT-Verantwortliche gut beraten sein, wenn er sich schon aus Gründen einer vollständigen Kompatibilität für ein Netzwerk aus einer Hand entscheidet. Bei größeren Netzen ist auch eine entsprechende CoS-Management-Software unerlässlich, um die zur Verfügung stehenden Priorisierungsmechanismen auch wirklich effizient nutzen zu können.

Qualitätssicherung auf Ebene 2

Für die Ebene 2, den Data-Link-Layer, ist in der Spezifikation IEEE 802.1Q die VLAN-Funktionalität beschrieben, die eigentlich dazu gedacht ist, auf Switches virtuelle LANs einzurichten und so unabhängig von der physikalischen Struktur eine logische Unterteilung des Netzwerks zu erhalten. Die Zuordnung eines Paketes zu einem VLAN erfolgt mit Hilfe einer Marke, dem Tag, im Layer-2-Header. In diesem Tag ist neben der VLAN-ID unter anderem auch ein Feld »User Priority« vorgesehen. Die Nutzung dieses 3-Bit-Feldes zur Einteilung der Pakete in acht mögliche Prioritätsklassen ist in der Spezifikation IEEE 802.1p festgehalten. – »Mögliche« Prioritätsklassen sind dies, weil die tatsächlich zur Verfügung stehenden Warteschlangen unterschiedlicher Priorität von der Hardware des Switches abhängt. Die meisten Systeme bieten mindestens vier verschiedene Queues.

Qualitätssicherung auf Ebene 3

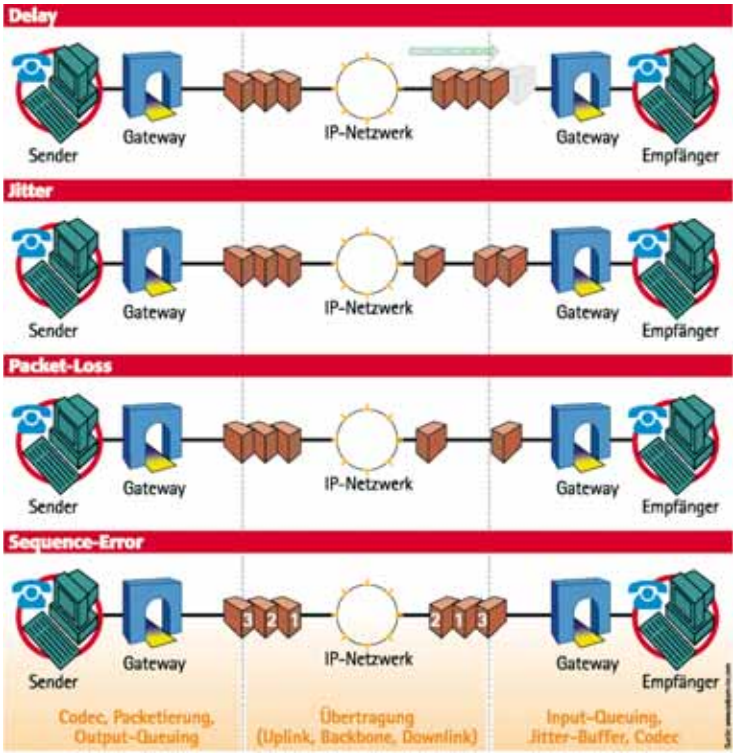
Eine weitere Möglichkeit der Zuordnung eines IP-Paketes ist die Nutzung des Type-of-Service-Byte, kurz ToS, im IP-Header Version 4. Dazu sind zwei Varianten beschrieben. RFC 791 definiert mit den Bits 0 bis 2 acht Klassen, von »Routine« über »Immediate« bis zu »Network-Control«. Pakete mit einem höheren Octal-Wert in diesem 3-Bit-Feld werden vorrangig behandelt (IP-Precedence). Variante 2 verwendet die Bits 3 bis 6, um eine normale und vier um besondere Service-Klassen zu kennzeichnen. Festgehalten ist dies in RFC 1349. Ungünstigerweise wird dieses vier Bit große Teilfeld des ToS-Byte ebenfalls als Type-of-Service bezeichnet. Es gibt also im IP-Header ein ToS-Byte und darin enthalten ist ein ToS-Feld. Pakete können anhand des ToS-Feldes entsprechend der eingestellten Klasse Warteschlangen unterschiedlicher Priorität zugeordnet werden. Im IP-Header Version 6 ist ebenfalls ein Byte für eine Klasseneinteilung vorgesehen. Es wird treffend als »Class« bezeichnet und könnte ähnlich verwendet werden.

Eine Arbeitsgruppe der IETF stellte 1997 eine alternative Implementation des ToS-Byte vor. Auch bei den Differentiated-Services, kurz Diffserv, wird dieses Byte dazu verwendet, um Pakete mit Markierungen zu versehen, die dann auf den Netzwerkknotenpunkten eine bestimmte Behandlung bei der Weiterleitung zum nächsten Knoten bewirken (Per-Hop-Behavior). Dazu erhält dieses Byte im IP-Header per Definition eine neue Bedeutung und wird in diesem Anwendungsfall dann als Differentiated-Service-Byte oder kurz DS-Byte bezeichnet.

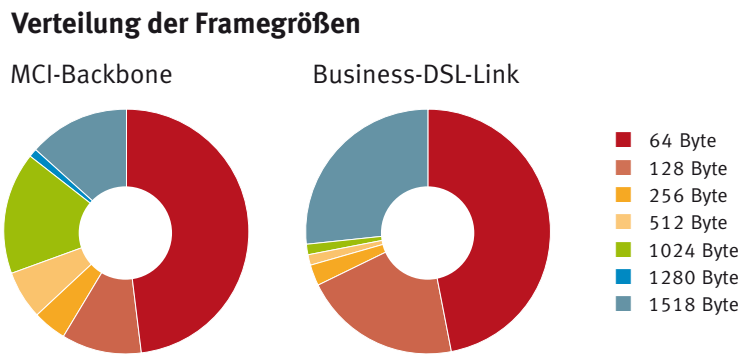
Die Diffserv-Spezifikation nach RFC 1349 definiert sechs Bits, die dazu dienen, den Differentiated-Services-Code-Point festzulegen. Diese sechs Bits werden genutzt, um verschiedene Service-Klassen zu definieren. Jede Netzwerkkomponente entscheidet anhand dieser Bits, wie die entsprechenden Pakete zu behandeln sind und steuert das Per-Hop-Behavior. Die sechs Bits sind nochmals in zwei mal drei Bits unterteilt. Diese Struktur ist in RFC 1349 festgeschrieben, aber letztendlich ist es den Herstellern beziehungsweise den Netzwerkadministratoren freigestellt, wie sie diese Bits genau nutzen. Eine sinnvolle Diffserv-Anwendung ist daher nur möglich, wenn ein Managementsystem durchgängig die notwendigen Service-Klassen-Zuordnungen steuert. Die insgesamt 64 Codierungs-Möglichkeiten müssen auf die vorhandenen Hardware-Queues beziehungsweise auf die zur Verfügung stehenden Links abgebildet werden und so dafür sorgen, dass die unterschiedlichen Dienste mit der gewünschten Qualität übertagen werden können. Diese Mechanismen müssen in einer Domäne konsistent arbeiten und zwischen verschiedenen Domänen durch Mapping gesichert werden.

Die Funktionsweise bei der Priorisierung ist im Grunde immer die gleiche. Pakete werden auf

Störfaktoren im Ethernet-Netzwerk



Störfaktoren wie Packet-Loss, Delay oder Jitter können die Übertragung von Real-Time-Applikationen empfindlich stören und beispielsweise die VoIP-Telefonie unmöglich machen.



Analysen der Verteilung der Framegrößen, beispielsweise für das NCI-Backbone oder von den Applikationen her typischer Business-DSL-Links haben ergeben, dass rund 50 Prozent aller Datenrahmen in realen Netzen 64 Byte groß sind.

den Gateways und Knotenpunkten anhand dieser Unterscheidungsmerkmale in den Headern den Warteschlangen oder Queues unterschiedlicher Priorität zugeordnet. Die Queues höherer Priorität werden dann entsprechend der Policy der jeweiligen Queuing-Mechanismen bevorzugt weitergeleitet. Welches Prinzip dieser Bevorzugung zu Grunde liegt, ist unterschiedlich. In vielen Fällen sollte eine Priorisierung aber nicht ohne eine Festlegung einer gewissen Bandbreite erfolgen. Diese könnte beispielsweise so aussehen, dass die Queue mit der höchsten Priorität nur eine bestimmte maximale Bandbreite erhält. Sonst kann es passieren, dass bei einer Überlast ausschließlich hoch eingestufte Pakete transportiert werden, während sich die Pakete in den unteren Queues stauen, bis sie verworfen werden. Das Festlegen einer minimalen Bandbreite für Pakete niedrigerer Priorität erfüllt den selben Zweck. Moderne Switches verschieben diese Grenzen dynamisch, abhängig vom momentanen Verkehr. Zu beachten ist jedoch, dass bei voller Ausreizung der entsprechenden Bandbreiten und der den Queues zugeordneten Buffern auch Pakete höherer Priorität keine Chance mehr haben, transportiert zu werden und ebenso verfallen können. Hierin liegt ein grundsätzlicher Nachteil der Ethernet-Technologie. Obwohl eigentlich alle aktuellen Netzwerkgeräte das ToS- beziehungsweise DS-Byte auswerten können, ist diese Funktion in den seltensten Fällen aktiviert und wird höchstens im In-House-Bereich oder anderen abgegrenzten und kontrollierbaren Umgebungen genutzt.

Queuing-Mechanismen

Die Hersteller von Switches verwenden oft eigene Namen für die CoS-Queuing-Strategien oder ändern die eigentliche Strategie nach ihren Vorstellungen ab. Oft werden auch verschiedene Strategien miteinander kombiniert, um die Ergebnisse zu verbessern. Die ursprünglichen Queuing-Strategien sind:

- First-In First-Out (FIFO),
- Stirct- und Rate-Controlled-Priority-Queuing (PQ),
- Fair-Queuing (FQ),
- Weighted-Fair-Queuing (WFQ),
- Weighted-Round-Robin-Queuing (WRR), auch als Class-Based-Queuing (CBQ) bezeichnet und
- Deficit-Weighted-Round-Robin-Queuing (DWRR).

Für die für unseren Switch-Vergleichstest unterstellten Anwendungsgebiete ist eine gleichmäßige, konfigurierbare Verteilung der Prioritäten erwünscht. DWRR bietet sich hier als Scheduling-Strategie an, da es sehr fair und auch schnell genug ist, um in Switches zu arbeiten, die Anschlüsse mit Gigabit-Geschwindigkeit haben. Dieses Verfahren erscheint wegen seiner Geschwindigkeit durch hardwaremäßige Implementierung und durch die Art des Schedulings von der Theorie her das am besten geeignete Verfahren zu sein.

Da DWRR recht komplex in der Realisierung ist, kann man es natürlich auch mit einem einfachen WRR ersetzen. Allerdings sollte man in diesem Fall bedenken, dass Qualitätsklassen mit vorwiegend kürzeren Datenframes weniger benachteiligt weitergeleitet werden. Das heißt, wenn zwei Klassen die gleiche Bandbreite zugewiesen bekommen, wird die Klasse mit den längeren Frames mehr resultierende Bandbreite erhalten als die Klasse mit den kürzeren Frames. Der Grund dafür liegt im Verfahren selbst.

Als Ersatz für das DWRR kann man auch eine WFQ-Strategie verwenden. Das Resultat ist dabei das gleiche, nur eignet sich DWRR besser für Switches im Gigabit-Bereich als WFQ. Der Grund dafür liegt in der Implementierung. DWRR kann hardwaremäßig implementiert werden, WFQ aber nur softwaremäßig. Softwaremäßige Implementierungen sind in der Regel aber deutlich langsamer als hardwaremäßige.

Um bei extrem überlasteten Netzen auch eine Qualitätsklasse zu haben, die von den anderen unbeeinflusst geswitcht wird, bietet sich die Zusammenarbeit des DWRR – oder auch WRR beziehungsweise WFQ – mit einer Strict-PQ-Strategie an. So können die Daten zur Netzwerkkontrolle, wie sie in der CoS-Klasse 7 in dem Standard 802.1D gefordert, jederzeit ihr Ziel erreichen. Da vier verschiedene Qualitätsklassen untersucht werden sollen, bietet es sich an, drei Queues mit DWRR zu versehen und eine Queue mit Strict-PQ. Die mit DWRR verwalteten Queues werden für die CoS-Klassen 1, 3 und 5 verwendet. Sie entsprechen Daten für Hintergrundverkehr mit geringer Priorität, Excellent-Effort für hervorragende Übertragung und Videodaten, die in Echtzeit mit höchster Priorität übertragen werden sollten.

In diesem Aufbau für die Scheduling-Strategien können dann verschiedene Gewichtungen eingestellt werden. Dabei sollte die Gewichtung des Strict-PQ nie verletzt werden. Bei Überlast in allen Prioritäten kann es in den anderen Queues theoretisch dazu kommen, dass diese ihre gewünschten Gewichtungen nicht einhalten können.

Edge- und Core-Switches

Bei der Priorisierung mit Bandbreitenmanagement sind zwei grundsätzlich unterschiedliche Verhaltensweisen der Switches zu unterscheiden: Rate-Limited-Switching und Weighted-Switching. Beim Rate-Limited-Switching garantiert der Switch den entsprechend konfigurierbaren Bandbreitenanteilen der einzelnen Prioritäten nicht nur eine Mindestbandbreite, er »deckelt« quasi auch die Durchsätze, indem er übrige Bandbreiten, die ein Dienst, dem sie zur Verfügung stehen, derzeit nicht benötigt, auch nicht für andere Dienste verfügbar macht. Eine solche Funktionalität ist insbesondere für den Edge-Bereich je nach Policy unverzichtbar, weil es so möglich ist, der Entstehung von Überlasten bereits in der Netzwerkperipherie vorzubeugen. Switches im Core-Bereich sollten dagegen die Mindestbandbreiten für die ihnen zugeordneten Dienste reservieren. Wenn diese Dienste die ihnen zustehenden Bandbreiten aber nicht benötigen, dann sollten andere Dienste freie Bandbreiten über die ihnen selbst zustehenden hinaus ruhig nutzen können. Ansonsten wird die insgesamt im Core-Bereich zur Verfügung stehende Bandbreite unnötig verringert. Ein solches Verhalten kann aber auch im Rahmen der Policy erwünscht sein. Um den verschiedenen Mechanismen gerecht zu werden, haben wir lediglich das Verhalten der Systeme bei Volllast gewertet, da dann unabhängig vom verwandten Mechanismus die gleichen Maximalbandbreiten für die jeweilige Priorität eingehalten werden müssten.

Prof. Dr. Bernhard G. Stütz,
dg@networkcomputing.de