

Jörn Kruse

Internet-Überlast, Netzneutralität und Service-Qualität

Die Datenmenge im Internetverkehr wächst ständig. Insbesondere Anwendungen mit hohen Bitraten können dabei Staus verursachen. In diesem Fall werden bislang alle Pakete gleich behandelt. Dies führt aber bei verschiedenen Inhalten zu gravierenden Qualitätseinbußen. Wie können diese Überlastprobleme ökonomisch effizient gelöst werden? Inwieweit wird das klassische Internet modifiziert?

Die Datenmenge, die täglich im Internetverkehr bewegt wird, nimmt stark zu. Dies ist zu einem erheblichen Teil auf hochvolumige Dienste und Inhalte zurückzuführen, z.B. Up- und Download von Software-, Musik- und Video-Dateien etwa über Filesharing-Plattformen oder Internet-Fernsehen (IPTV) sowie auf Internettelefonie (VoIP) und Online-Spiele etc. Die Datenmenge führt temporär zu Überlastungen der IP-Infrastruktur, die erhebliche Qualitätsminderungen zur Folge haben (können) und im Kontext der Netzneutralitäts-Diskussion wirtschaftspolitisch relevant sind. Darauf konzentrieren sich die folgenden Überlegungen.

Die Ursprungsdaten bzw. die Datenströme der einzelnen Dienste und Inhalte werden zum Transport über das Internet in kleine Pakete zerteilt, die einzeln zum Empfänger geschickt werden. Die Datenpakete enthalten in ihrem Header die Informationen über das Ziel und die Position im Datenstrom, werden über verschiedene Router (die quasi die Vermittlungselemente im IP-Netz darstellen) und Übertragungswege geleitet und beim Empfänger wieder zusammengesetzt. Auf der Basis der ausgelesenen Header-Informationen schicken die Router die einzelnen Datenpakete auf dem jeweils geeigneten Weg zum Ziel. Diese Vorgänge werden von den internetspezifischen Protokollen TCP (Transmission Control Protocols) und IP (Internet Protocols) gesteuert. Alle über das Internet transportierten Dienste und Inhalte sind auf der Datenpaketenebene praktisch homogen und werden über eine universelle Infrastruktur zum jeweiligen Empfänger geleitet. Dies ist ein wesentlicher Grund für den rasanten technischen und wirtschaftlichen Erfolg des Internet.

Wenn bei einem hohen Datenaufkommen die Kapazität der Router bzw. der abgehenden Übertragungswege nicht ausreicht, entsteht ein Stau der Datenpakete an den Routern. Es kommt zunächst zu einer Zwischenspeicherung. Aus dieser werden die Pakete

nach dem First-in-first-out-Prinzip weitergeleitet. Eine Folge sind Datenverzögerungen (Delay)¹ und/oder Jitter.² Ein ähnlicher Effekt ergibt sich, wenn einzelne Infrastrukturelemente (z.B. Kabel) aufgrund von Störungen, etwa Naturereignissen (z.B. Erdbeben) oder technischen Schäden (Bagger etc.) ausfallen und deren Leistungen durch andere Übertragungswege und Router zusätzlich bewältigt werden müssen.

Wenn der Zwischenspeicher nicht mehr ausreicht, kommt es zum Verlust von Datenpaketen (Packet Loss). Einige der Dienste, nämlich solche, die als „elastisch“ bezeichnet werden, melden nicht angekommene Pakete an den Absender, so dass diese erneut gesendet und beim Empfänger eingefügt werden können.³ Der einzelne Datenverlust wird also quasi geheilt. Zu solchen Diensten gehören z.B. E-mail, Web-Surfen, Downloads etc. Andere (inelastische) Dienste haben derartige Nachsende-Mechanismen nicht, so dass nicht angekommene Pakete endgültig verloren sind. Dazu gehören insbesondere interaktive Dienste sowie IPTV (Internet-Fernsehen).

Delay, Jitter und Packet Loss können bei bestimmten Diensten deutliche Qualitätsminderungen für die Nutzer zur Folge haben. Bei anderen Diensten und Inhalten, die über das Internet realisiert werden, sind solche Wirkungen weit geringer oder gar nicht spürbar. Bei „qualitätssensitiven“ Diensten führt die Überlast zu erheblichen Qualitätsmängeln, die die Nachfrage deutlich beeinträchtigen, zu geringeren Erlösen, zu

¹ Der Delay erhöht die Zeitspanne, gemessen in Millisekunden, die ein Datenpaket vom Sender zum Empfänger benötigt. Vgl. zu den Qualitätsparametern insgesamt W. Brenner, M. Dous, R. Zarnekow, J. Kruse: Qualität im Internet. Technische und wirtschaftliche Entwicklungsperspektiven, St. Gallen 2007 (erhältlich über <http://www.hsu-hh.de/kruse/index>).

² Da im gleichen End-to-end-Strom die einzelnen Datenpakete insbesondere bei Überlast teilweise unterschiedliche Wege nehmen, treffen sie eventuell in anderen Abständen und Reihenfolgen im Ziel ein als denjenigen, in denen sie abgesendet wurden. Als Jitter (Latenzschwankungen) wird die Schwankung des Delays bezeichnet, gemessen als maximale Abweichung in Millisekunden.

³ In der Regel verwenden elastische Dienste das TCP-Protokoll, das einen Nachsende-Mechanismus implementiert hat.

Prof. Dr. Jörn Kruse, 59, ist Inhaber des Lehrstuhls für Wirtschaftspolitik an der Helmut Schmidt Universität in Hamburg.

geringerer Innovation und zu ökonomischer Ineffizienz. Dass diese Effekte gegenwärtig tatsächlich zu beobachten sind, kann man zunächst auf zwei Ursachen zurückführen:

1. *Flatrates*: Da die Internet-Preise der Endnutzer üblicherweise von der Menge des von ihnen generierten Datenverkehrs unabhängig und damit deren Grenzkosten null sind, decken sie bei Überlast nicht ihre volkswirtschaftlichen Grenzkosten, die dann (auch kurzfristig) größer als null sind. Dies gilt vor allem für einige Dienste mit besonders hoher Datenrate, insbesondere Peer-to-Peer-Filesharing (P2P), die in Deutschland mehr als zwei Drittel des gesamten Internetverkehrs ausmachen (siehe unten) und entsprechend stark an der Entstehung von Überlastsituationen beteiligt sind. Insofern sind die Flatrates ein erster Hauptgrund für die Überlastprobleme.
2. *Netzneutralität*: Netzneutralität bedeutet in enger Sicht, dass alle Datenpakete strikt gleich behandelt werden. In den USA wird intensiv darüber diskutiert, ob die Gleichbehandlung aller Datenpakete, unabhängig von deren Versendern, Diensten und Inhalten, regulatorisch vorgeschrieben werden sollte.⁴

Die Befürworter einer regulatorisch festgeschriebenen Netzneutralität argumentieren vor allem mit solchen diskriminierenden Verhaltensweisen, die negative Wirkungen auf den Wettbewerb bzw. wohlfahrtsmindernde Effekte zur Folge haben könnten. Sie fürchten, dass die Internet Service Provider (ISP) oder andere Netzbetreiber anderenfalls bestimmte Anwendungen (z.B. VoIP) oder Inhalte diskriminieren könnten, indem sie deren Datenpakete blockieren bzw. drosseln. Technisch wäre dies ohne weiteres möglich, da die entsprechenden Informationen aus den Headern der Datenpakete ausgelesen werden können. Die Anreize dazu könnten insbesondere dann bestehen, wenn die Internet Service Provider bzw. Netzbetreiber selbst Dienste anbieten, die zu den betreffenden Internetdiensten in einer Konkurrenzbeziehung stehen. Dies würde grundsätzlich den Wettbewerb beschränken und die ökonomische Effizienz mindern. Allerdings kann man davon ausgehen, dass derartige Diskriminierungspraktiken (1) für die Netzbetreiber, die in europäischen Ländern

(anders als in den USA) typischerweise im Wettbewerb stehen, ohnehin meist nicht vorteilhaft wären und (2) außerdem nach dem deutschen und europäischen Wettbewerbsrecht gar nicht zulässig wären.

Bei einem engen Verständnis von Netzneutralität können die Datenpakete nicht danach differenziert behandelt werden, ob Verzögerungen oder Verluste hohe oder geringe Qualitätsmängel oder andere Nachteile für die Nutzer mit sich bringen. Mit anderen Worten: Der ökonomische Wert des staufreien Transports eines bestimmten Datenpaketes könnte bei Netzneutralität nicht berücksichtigt werden. Dies macht bereits deutlich, dass eine Netzneutralitätsregulierung im Überlastfall nicht ökonomisch effizient sein kann.

Aufgrund der unterschiedlichen Qualitätssensitivitäten und Datenraten einzelner Dienste sowie der unterschiedlichen Zahlungsbereitschaft der Nutzer, kann es zu einem Effekt kommen, der als Crowding-Out bezeichnet wird.⁵ Dabei werden im Ergebnis einige ökonomisch hochwertige Dienste (d.h. solche mit hoher Konsumenten- und/oder Produzentenrente) durch geringwertige Dienste ganz oder vollständig verdrängt.

Dienstespezifische Datenraten, Qualitätsminderungen und ökonomische Werte

Eine wesentliche Ursache für den grandiosen Erfolg des Internet (sowie für Multimedia etc.) besteht darin, dass die Inhalte, die in der Ausgangsform (Schrift, Grafik, Musik, Fotos, Video, Businessanwendungen, Spiele etc.) ganz unterschiedliche Wahrnehmungskategorien und Funktionalitäten beinhalten, alle in den gleichen Typ von Datenpaketen transformiert und dieser Form über eine universelle Infrastruktur verschickt werden können. Beim Empfänger werden sie wieder in die diversen Anwendungen, Darstellungsformen und Funktionen zurücktransformiert. Auf der eigentlichen Infrastrukturebene, auf der die Überlast auftritt, sind die Datenpakete also homogen. Allerdings unterscheiden sich die einzelnen Dienste bezüglich ihrer (1) Datenraten, (2) Qualitätssensitivitäten, und (3) wirtschaftlichen Werte ganz erheblich.

1. Die Datenraten (Zahl der Datenpakete pro Zeiteinheit) sind vor allem bei Up- und Downloads besonders hoch, z.B. solche über P2P-Filesharing-Plattformen.⁶ Die Inhalte bestehen vor allem aus Videos,

⁴ Vgl. z.B. D. P. McCIure: Network Neutrality: Phantom Problem, Unintended Consequences, White Paper, US Internet Industry Association (USIIA), 14. März 2006; J. G. Sidak: A Consumer-Welfare Approach to Network Neutrality Regulation of the Internet, in: Journal of Competition Law and Economics, Vol. 2 (2006), Nr. 3, S. 349-474; B. van Schewick: Towards an Economic Framework for Network Neutrality Regulation, in: Journal on Telecommunications and High Technology Law, Vol. 5 (2007), Nr. 2, S. 329-392; R. Litan, H. Singer: Unintended Consequences of Net Neutrality Regulation, in: Journal on Telecommunications and High Technology Law, Vol. 5 (2007), Nr. 3, S. 533-572; R. Dewenter: Netzneutralität, Diskussionspapier Nr. 74, Fächergruppe Volkswirtschaftslehre, Helmut-Schmidt-Universität, Hamburg 2007.

⁵ Vgl. J. Kruse: Crowding-Out bei Überlast im Internet, Diskussionspapier Nr. 72, Fächergruppe Volkswirtschaftslehre, Helmut-Schmidt-Universität, Hamburg 2007, download <http://www.hsu-hh.de/kruse/index>.

⁶ Dies wird im Folgenden nach dem üblichen Sprachgebrauch pauschal als Peer-to-Peer (P2P) oder als Filesharing bezeichnet. Zu diesen gehören insbesondere (häufig großvolumige) Downloads und Uploads von Videos, Musik und Software. Dass die Inhalte überwiegend illegal (fehlende Copyrights) sind, soll hier nur am Rande erwähnt werden.

Musik und Software.⁷ Derartige P2P-Plattformen machen einen erheblichen Teil (in Deutschland mehr als zwei Drittel)⁸ des gesamten Internetverkehrs aus.

- Die Folgen der Überlast sind je nach betrachtetem Dienst sehr unterschiedlich.⁹ Während (in Grenzen) der Delay für die Qualität bestimmter Internetdienste, wie z.B. E-Mails, Webseitenabruf oder Downloads fast keine Rolle spielt, hat er bei zeitkritischen Diensten einen maßgeblichen Einfluss auf die Dienstqualität. Zu diesen zählen insbesondere interaktive Dienste (Voice over IP, Videospiele etc.).¹⁰ Ähnliches gilt für Jitter und Packet Loss.¹¹ Während elastische Internetdienste (z.B. Filesharing, E-mails, Webbrowsing) einen Paketverlust erkennen und durch das erneute Versenden des Datenpaketes heilen können, besteht bei interaktiven, zeitkritischen Anwendungen (z.B. Voice over IP, Videotelefonie, Onlinespiele) diese Möglichkeit nicht. Sprach- und Videodienste können Packet Loss nur teilweise kompensieren. Darüberhinaus führt es zu Bild- und Tonstörungen, die schnell ein für fernsehgewöhnte Zuschauer intolerables Ausmaß annehmen können.
- Der wirtschaftliche Wert, den der Transport eines Datenpaketes hat, unterscheidet sich nach einzelnen Diensten und Nutzern. Das Gleiche gilt auch für die zusätzliche Zahlungsbereitschaft der Nutzer dafür, dass die Datenpakete ohne Staufolgen (Verzögerung, Jitter und/oder Paketverlust) ankommen. Diese ist bei geringer Qualitätssensitivität ebenfalls gering und steigt mit der Qualitätssensitivität und dem Wert für die Nutzer.

Der wirtschaftliche Wert wird ökonomisch durch die Summe aus Konsumenten- und Produzentenrente (Wohlfahrt) gemessen. Für die Anbieter steht dabei die Zahlungsbereitschaft der Kunden bzw.

die Nachfragefunktion im Vordergrund, die das Erlöspotential und ihre Gewinnmöglichkeiten bestimmen, und damit auch ihre Anreize für Innovationen.

Bei vielen Businessanwendungen, Voice over IP und anderen interaktiven Diensten besteht häufig eine hohe Zahlungsbereitschaft für den Transport eines Datenpaketes, so dass man sie als hochwertige Dienste bezeichnen kann. Dagegen produzieren z.B. die Filesharing-Plattformen in der Regel geringwertige Dienste, da die Nutzer bzw. die Werbetreibenden für den Transport eines Datenpaketes nur eine geringe Zahlungsbereitschaft haben und die erzeugte Wohlfahrt klein ist.

Crowding-Out hochwertiger Dienste

Betrachten wir den Crowding-Out-Effekt anhand zweier Dienste A und B.¹² A ist ein Dienst mit hoher Datenrate, geringer Qualitätssensitivität und geringem ökonomischen Wert (z.B. Downloads und Filesharing). B ist (unabhängig von der Datenrate) ein Dienst mit hoher Qualitätssensitivität und hohem ökonomischen Wert (z.B. Business-Kommunikation, Voice over IP und andere interaktive Dienste).

Angenommen, die Datenmenge von A ist hoch und nimmt weiter zu, was insbesondere bei Endnutzer-Flatrates plausibel ist, da hier die Preise (Grenzkosten) pro Datenpaket für die Nutzer null sind. Dies hat zur Folge, dass die Überlastsituationen immer häufiger auftreten und sich damit die Qualitätsminderungen für B immer öfter und stärker negativ bemerkbar machen. Damit wird die Nachfrage für B-Dienste immer geringer und Konsumenten- und/oder Produzentenrente, Erlöse etc. sinken. In letzter Konsequenz werden die hochwertigen Dienste durch geringwertige Dienste eventuell vollständig verdrängt (Crowding-Out). Innovative Dienste, die hohe Anforderungen an die Transportqualität haben, werden gar nicht erst entwickelt und eingeführt.

Fazit: Ein solcher Crowding-Out-Effekt kann unter den gegenwärtigen Bedingungen also auch von solchen Diensten verursacht werden, für die selbst nur eine sehr geringe Zahlungsbereitschaft vorhanden ist und die nur minimale Konsumenten- und/oder Produzentenrenten und nur geringe Umsätze erzeugen, also ökonomisch nur einen geringen Wert aufweisen.

Die daraus entstehenden ökonomischen Ineffizienzen sind eine unmittelbare Folge von Flatrates und Netzneutralität. Es bieten sich die folgenden Lösungsoptionen an: Overprovisioning, separate Netze, Volumentarife und Priority Pricing (Quality-of-Service-Konzepte).

⁷ In Deutschland sind dies insbesondere BitTorrent und eDonkey, die 66,7% bzw. 28,6% des P2P-Verkehrs ausmachen. Vgl. H. Schulze, K. Mochalski: The Impact of P2P File Sharing, Voice over IP, Skype, Joost, Instant Messaging, One-Click Hosting and Media Streaming such as YouTube on the Internet, Ipoque Internet Study 2007, S. 4.

⁸ Nach der Ipoque-Studie 2007 sind die quantitativen Anteile für Deutschland: P2P 69,25%, Webbrowsing 10,05%, Media Streaming (inklusive YouTube etc.) 7,75%, VoIP 0,92%, E-Mail 0,37%. Vgl. H. Schulze, K. Mochalski, a.a.O., S. 2.

⁹ Vgl. W. Brenner, u.a., a.a.O.

¹⁰ Bei Voice-over-IP gelten Verzögerungen bis zu ca. 150 Millisekunden für die Telefonnutzer als akzeptabel, darüber hinaus nicht mehr. Bei Online-Spielen führt bereits ein Delay von 50-100 ms zu Qualitätseinbußen. Höhere Werte stören das Spiel gravierend, so dass der Dienst praktisch nicht mehr nutzbar ist.

¹¹ Ein hoher Wert für Jitter führt bei zeitkritischen Diensten, wie z.B. Voice-over-IP oder Videoübertragung, zu einer deutlichen Beeinträchtigung der Servicequalität. Nur wenig oder gar nicht bemerkbar sind die Folgen derartiger „Unregelmäßigkeiten“ des Pakettransports z.B. bei Filesharing, E-mails oder Webbrowsing.

¹² Vgl. ausführlicher J. Kruse, a.a.O.

Overprovisioning und optimale Kapazität

Grundsätzlich könnten die meisten Überlastprobleme vermieden werden, wenn immer so große Überkapazitäten vorgehalten werden, dass jederzeit (d.h. auch bei sehr kurzfristiger Spitzenlast) mehr Bandbreite verfügbar ist als Verkehrslast auftritt, so dass alle Datenpakete immer sofort weitergeleitet werden können. Dies wird als Overprovisioning bezeichnet. Außerdem könnten zur weitgehenden Vermeidung der Folgen von störungs- oder katastrophenbedingten Kapazitätsausfällen (kapazitätsinduzierte Überlast) grundsätzlich die einzelnen Netzelemente redundant ausgelegt werden.

Ein Overprovisioning der Kapazitäten auf eine potentielle maximale Spitzenlast und eine exzessive Redundanzvorsorge für Störungen erfordert hohe Reservekapazitäten und verursacht entsprechend hohe Investitionsaufwendungen für die Netzbetreiber. Diese haben keine wirtschaftlichen Anreize für die entsprechenden Investitionen, da sie über die Erlöse nicht refinanzierbar wären.

Es ist offensichtlich, dass Overprovisioning volkswirtschaftlich ineffizient ist.¹³ Ein erheblicher Teil der Kapazitäten würde praktisch nie genutzt werden und die Überkapazitäten sind durch die dadurch erzeugten Qualitätseffekte und die Nutzensteigerungen für die Kunden nicht zu rechtfertigen. Dabei muss berücksichtigt werden, dass ein ganz wesentlicher Teil der induzierten Überkapazitäten durch Dienste verursacht wird, deren Nutzer nur eine geringe Zahlungsbereitschaft haben.

Bei Overprovisioning entsteht im Zeitablauf ein problematischer Selbstverstärkungseffekt. Je mehr Kapazitäten und Bandbreiten dadurch zur Verfügung stehen, desto mehr datenintensive Applikationen und entsprechende Nutzungsgewohnheiten werden sich herausbilden. Dies lässt bei Flatrates die Frage offen, ob solche Dienste und Konsummuster sich auch entwickeln würden, wenn die Nutzer die von ihnen verursachten Kosten tragen müssten. Das Ergebnis wäre ein „Fass ohne Boden“, da sich die Datenraten weiter erhöhen (z.B. HD-Video) und die Nutzung hoch-datenratiger Dienste weiter zunehmen würden.

Es wäre ökonomisch unsinnig, die Netzbetreiber durch eine Regulierung zu zwingen, auf Dauer die Kosten für die P2P-Heavy-User zu tragen, um Overprovisioning zu realisieren.

¹³ Vgl. dazu ausführlicher J. Kruse, a.a.O.; W. Brenner, u.a., a.a.O.

Separate Netze für qualitätssensitive Dienste

In der wirtschaftlichen Realität würden einzelne große Marktteilnehmer sich jedoch nicht mit den negativen Folgen der „Netzneutralitäts-Internet-Allmende“ zufrieden geben, sondern selbst tätig werden. Wenn die Überlastfolgen eintreten oder sich abzeichnen, haben die Anbieter (bzw. die Nutzer) hochwertiger, qualitätssensitiver Dienste wirtschaftliche Anreize, in eigene IP-Netze zu investieren, um die Qualitätsparameter zu realisieren, die für die technische und wirtschaftliche Funktionsfähigkeit ihrer Dienste erforderlich sind. Dies setzt keine eigenen physischen Netze voraus, sondern kann auf der Basis virtueller Netze erfolgen. In der Praxis würde dies eher so realisiert, dass bestimmte Dienste- und Inhalteanbieter Verträge mit den Netzbetreibern über die Bereitstellung separater (bzw. exklusiver) IP-Netze bestimmter Bandbreiten und Verfügbarkeiten schließen, was teilweise bereits erfolgt ist.

Im Ergebnis würden damit für einzelne Dienste oder Inhalte (bzw. Dienste- und Inhalte-Anbieter) eigene IP-Infrastrukturen errichtet und damit von der generellen IP-Infrastruktur separiert. Mit anderen Worten, das klassische Internet mit universeller Netzinfrastruktur als Common-Use-Ressource würde aufhören zu bestehen. Dadurch würden die volkswirtschaftlichen Gesamtkosten der IP-Netze ansteigen, da die Economies of Scope der gemeinsamen Nutzung einer IP-Infrastruktur durch viele Dienste nicht mehr ausgeschöpft werden könnten. Es müssten insgesamt höhere Kapazitäten erstellt werden. Da die Kapazitätsinvestitionen im exklusiven Bereich (mehr oder minder) dienstespezifisch wären, würden die Risiken der Investitionen ansteigen, was sich in höheren Kosten und Preisen niederschlagen würde.

Die Abkehr von einer universellen IP-Infrastruktur würde also volkswirtschaftliche Ineffizienzen mit sich bringen. Eine solche Erstellung eigener Infrastrukturen für die qualitätssensitiven, ökonomisch wertvollen Dienste würde allerdings auch Rückwirkungen auf die verbleibende IP-Restinfrastruktur haben. Die Anreize der Netzbetreiber, in sie zu investieren, würden sinken, und die Attraktivität des Netzes sowie die Zahl der Nutzer würden geringer.

Wenn die Betreiber der bisherigen Universal-IP-Netze keine entsprechenden Maßnahmen ergreifen (siehe unten), würde die Separierung fortschreiten. Mehr und mehr Dienste würden „auswandern“. Schon bald würden die Folgen sich auch bei den A-Diensten zeigen, deren Zahlungsbereitschaft nicht ausreicht, ihre eigene Infrastrukturnutzung zu finanzieren. Bei

Weiterbestehen von Flatrates wären die Netzbetreiber gezwungen, die A-Dienste zu drosseln, um die Internet-Funktionsfähigkeit zu retten.

Volumentarife statt Flatrates

Eine wesentliche Ursache der Qualitätsprobleme besteht im Vorherrschen von Flatrates, die für die Versender von Datenpaketen Grenzausgaben von null zur Folge haben, während die ökonomischen Grenzkosten (kurzfristig zu Überlastzeiten) durchaus nicht null sind. Die langfristigen Grenzkosten sind ohnehin immer größer als null.

Eine naheliegende Lösung ist die Ersetzung der Flatrates durch volumenbasierte Tarife. Dies würde in der Tat einen wesentlichen Teil der Probleme lösen. Allerdings wäre eine funktionsfähige Implementierung, die das genannte Überlast-Qualitätsproblem löst, unter den spezifischen Nutzungsbedingungen und Lastschwankungen des Internet nur schwer zu realisieren. Dies folgt schon aus den extrem kurzen Zeitspannen, die eine Mengenreaktion auf knappheitsorientierte Ad-hoc-Preise stark erschweren.

Da die Netzlast im Zeitablauf stark variiert, würde die Festsetzung ökonomisch adäquater Preisstrukturen (System von Spitzenlastpreisen) eine Prognose der Nachfragefunktionen zukünftiger Zeitpunkte erfordern, was nur partiell möglich ist, da wesentliche Elemente der Nachfrage nicht zyklisch oder anderweitig prognostizierbar, sondern nahezu erratisch sind. Temporär „zu niedrige“ Preise würden die Qualitätsmängel bei qualitätssensitiven Diensten weiter bestehen lassen. Temporär „zu hohe“ Preise würden einzelne Dienste ausschließen, obwohl noch freie Kapazitäten vorhanden (und somit die kurzfristigen Grenzkosten null) wären.

Aber selbst für den allenfalls theoretisch denkbaren Fall, dass die jeweiligen Preise genau die temporären Nachfragefunktionen (Verkehrslasten) widerspiegeln, alle Nutzer die Preise kennen und ihre Verkehrsmengen-Nachfrage entsprechend ausrichten, was aufgrund des zeitkritischen Charakters in aller Regel nicht der Fall ist, wären die Ergebnisse nicht effizient. Folgendes Problem besteht nämlich weiter: Im Internet sind die temporären Überlastphasen häufig extrem kurz. Sie bestehen zum Teil nur für wenige Sekunden, während danach wieder Router- bzw. Übertragungswege-Kapazitäten verfügbar sind (falls nicht der P2P-Verkehr zu einer längeren Spitzenauslastung führt). Bei einigen nicht-qualitätssensitiven Diensten können die Datenpakete getrost über kurze Zeitspannen aufgehalten werden, ohne dass fühlbare Qualitätsminderungen eintreten. Das Gleiche gilt für verlorene Datenpakete von Diensten, bei denen das Protokoll dafür sorgt, dass sie nachgeordnet und somit ersetzt werden. Zu

diesen gehören nicht nur E-Mails, sondern vor allem Downloads und P2P-Filesharing-Dienste.

Wenn die Datenpakete solcher Dienste also warten, bis wieder Router-Kapazität frei ist, würden sie keinerlei Staukosten verursachen (also kurzfristige Grenzkosten von null aufweisen). Insofern wäre es ökonomisch ineffizient, sie durch einen positiven Preis vom Transport auszuschließen, der für alle Dienste gilt und Preisbestandteile enthält, die auf den Überlastexternalitäten beruhen.¹⁴

Priorisierung, Priority Pricing und Quality of Service

Aufgrund der stark unterschiedlichen Qualitätssensitivität auf der Diensteebene besteht das eigentlich relevante Problem in einer adäquaten Priorisierung der einzelnen Datenpakete bei Überlast, auch wenn eine solche extrem kurzfristig auftritt. Dabei geht es um eine Priorisierung derart, dass einerseits Datenpakete qualitätssensitiver, hochwertiger Dienste möglichst sofort weitergeleitet werden, und andererseits Datenpakete nicht-qualitätssensitiver, geringwertiger Dienste gegebenenfalls warten müssen und erst verzögert weitergeleitet oder später ersetzt werden.

Die Priorisierung könnte grundsätzlich auf verschiedene Weise erfolgen. Eine Option wäre, bestimmte Dienste a priori bestimmten Prioritätsklassen zuzuordnen. Das wäre allerdings eine eher technokratische Lösung, die zahlreiche Zuordnungskonflikte und Anreize für das „Tarnen“ von Datenpaketen mit sich bringen würde. Man sollte den einzelnen Kunden die Wahl lassen, ob sie z.B. Voice over IP in Best-Effort- oder in hoher Sprachqualität nutzen wollen, d.h., die individuelle Zahlungsbereitschaft (für prioritäre Verarbeitung, also für das Vermeiden von Qualitätsminderungen bei den einzelnen Diensten) sollte bei der Priorisierung eine wesentliche Rolle spielen.

Die ökonomisch überzeugende Lösung besteht in einer wertbezogenen Steuerung der Priorisierung, dem sogenannten Priority Pricing. Darunter versteht man die Preissetzung für das Recht, vorrangig bedient zu werden, falls Angebotsengpässe bestehen sollten. Der höhere Preis für eine gegebenenfalls bevorzugte Bedienung gilt also generell, d.h. unabhängig vom konkreten Eintritt der Überlast. Dazu erhalten die Header der Datenpakete von den Versendern (bzw. den Internet-Nutzern) entsprechende Prioritätsinformationen, die von den Routern gelesen und bei Überlast zum Kriterium der Weiterleitung gemacht werden.

Die Kunden können sich dann ex ante zwischen verschiedenen Prioritätsklassen entscheiden, die eine unterschiedliche Priorität der Datenpakete gewähren

¹⁴ Vgl. J. Kruse, a.a.O.

und unterschiedliche Preise haben. Die Zahlungsbereitschaft für Priorität hängt einerseits von der Qualitätssensitivität eines Dienstes ab. Nur die Anbieter qualitätssensitiver Dienste werden überhaupt eine Veranlassung haben, für Priorität zu zahlen. Die Nutzer nicht-qualitätssensitiver Dienste (E-mail, Webbrowser, Filesharing) haben in der Regel keine nennenswerten Zahlungsbereitschaft für Priorität und werden den Internettransport besonders kostengünstig erhalten. Andererseits werden auch die Anbieter qualitätssensitiver Dienste nur dann eine nennenswerte Zahlungsbereitschaft für Priorität haben, wenn die Nutzer der Dienste (bzw. die Werbetreibenden) ihrerseits eine entsprechende Zahlungsbereitschaft für die Qualität dieser Dienste aufweisen. In aller Regel werden also nur die Anbieter bestimmter hochwertiger Dienste eine hohe Priorität wählen.

In der Praxis vieler heterogener Dienste mit entsprechend unterschiedlicher Qualitätssensitivität bezüglich aller drei Parameter (Delay, Jitter und Packet-Loss) und unterschiedlichen Anforderungen bezüglich der einzelnen Parameter wird sich auf wettbewerblichen Märkten vermutlich ein Quality-of-Service-System mit verschiedenen Qualitätsklassen und unterschiedlichen Preisen herausbilden. Ein Beispiel für ein System mit den vier Qualitätsklassen Interaktiv, Multimedia, Critical und Best-Effort ist in der Studie von Brenner u.a. beschrieben.¹⁵

Die Einordnung eines bestimmten Dienstes in eine Qualitätsklasse wird grundsätzlich durch den Diensteanbieter vorgenommen. Da mit unterschiedlichen Qualitätsklassen auch unterschiedliche Transportpreise verbunden sein werden, erfolgt die Zuordnung unter den üblichen Kosten-/Nutzen-Gesichtspunkten. Grundsätzlich kann der Diensteanbieter dies auch durch seine Kunden entscheiden lassen, indem er ihnen verschiedene Varianten zur Wahl anbietet.

Ein solches Qualitätsklassenmodell führt grundsätzlich zu einer ökonomisch effizienten Rationierung im Überlastfall und vermeidet die oben genannten Crowding-Out-Probleme.

Nach der Implementierung eines Quality-of-Service-Systems, das zu einer prioritären Bedienung der qua-

¹⁵ Die Qualitätsklasse „Interaktiv“ eignet sich für Dienste, die niedrige Delay- und Jitter-Werte erfordern, aber nicht sehr bandbreitenintensiv sind. Hierzu zählen beispielsweise Voice-over-IP-Dienste oder Online-Spiele. Die Qualitätsklasse „Multimedia“ ist insbesondere für Audio-/Video-Streaming-Dienste, wie Video-on-Demand, IPTV oder Internet-Radio, geeignet, die eventuell hohe Bandbreiten benötigen und gewisse Mindestanforderungen an Delay, Jitter und Packet Loss haben. Die Qualitätsklasse „Critical“ ist geeignet für kritische Geschäftsanwendungen, die die Internet-Infrastruktur zur Kommunikation nutzen. Die Bandbreite kann variabel gewählt werden. Delay, Jitter und Packet Loss sind gering. Die Qualitätsklasse „Best-Effort“ umfasst im Wesentlichen alle nicht qualitätssensitiven Dienste, wie z.B. Web-Browsing, E-Mail oder Downloads. Vgl. W. Brenner, u.a., a.a.O.

litätssensitiven Datenpakete führt, ist im Übrigen (falls es z.B. aus Marketinggründen solche Anreize gibt) auch eine Flatrate für Best-Effort-Service ökonomisch weniger nachteilig. Dann kann es nicht mehr zu einem Crowding-Out-Effekt zu Lasten hochwertiger Dienste kommen. Für den Quality-of-Service-Verkehr selbst ist eine Flatrate allerdings ineffizient.

Wenn man an einem strikten Netzneutralitäts-Konzept festhalten würde, wäre ein ökonomisch effizientes Quality-of-Service-Konzept nicht realisierbar. Bei Quality of Service werden die Datenpakete bei Überlast zwar differenziert behandelt (und nicht gleich), aber keine Nutzer diskriminiert, sondern alle Nutzer bei gleicher Zahlungsbereitschaft auch gleich behandelt. Da eine strikte Netzneutralitäts-Regel ökonomische Effizienz verhindern würde, ist sie abzulehnen. Dennoch gewährleistet das allgemeine Wettbewerbsrecht, dass bestimmte Dienste nicht willkürlich durch die Netzbetreiber gedrosselt oder blockiert werden können. Im Regelfall wäre jedoch zu erwarten, dass ohnehin der Wettbewerb der Netzbetreiber ein solches Verhalten verhindern würde.

In diesem Zusammenhang ist darauf hinzuweisen, dass bei der Einführung einer verbindlichen Netzneutralitäts-Regulierung die Effekte von dienstespezifischen Exklusiv-Netzen eintreten würden, die zu höheren volkswirtschaftlichen Gesamtkosten führen würden. Insofern verhindert ein Quality-of-Service-Konzept im Universal-Internet eine ineffiziente Exklusivlösung.

Zusammenschaltungsregime beim Quality-of-Service-Konzept

Da bei vielen Internet-Diensten die Datenpakete über die Netze verschiedener Betreiber transportiert werden, stellt sich die Frage nach der Realisierung des Quality-of-Service-Konzepts über Netzgrenzen hinweg und somit auch nach dem anzuwendenden Zusammenschaltungsregime.

Es ist für eine Quality-of-Service-Konzeption von elementarer Bedeutung, dass die Zusammenschaltungsentgelte bei der Paketübergabe zwischen zwei Netzen davon abhängig sind, ob bestimmte (vereinbarte) Qualitätsparameter zuverlässig eingehalten werden oder nicht.

Der Ausgangspunkt dafür ist die Kennzeichnung der Datenpakete mit der gewünschten (bzw. erforderlichen) Qualitätsklasse durch den Versender, bzw. durch dessen Internet Service Provider. Dieser muss seinerseits bei der Übergabe des Verkehrs an den nächsten Netzbetreiber dafür Sorge tragen, dass jener die Datenpakete ebenfalls so prioritär behandelt, dass die gewünschten Qualitätsmerkmale realisiert werden.

Er wird deshalb seinen Qualitätsverkehr nur an solche Netze weitergeben, die dies gewährleisten.

Da die nachhaltige Realisierung und Gewährleistung von definierten Qualitätsparametern höhere Kosten als für einen Best-Effort-Verkehr verursacht, wird ein empfangender Netzbetreiber dem übergebenden Netz eine solche Gewähr nur dann geben, wenn dieser dafür entsprechend zahlt. Mit anderen Worten: Die Realisierung des Prinzips des Sending Party's Network Pays (SPNP) ist eine Voraussetzung für die erfolgreiche Implementation eines Quality-of-Service-Konzeptes.

Der einzelne Ausgangs-Internet-Service-Provider stellt seinen Kunden den Qualitätsverkehr der von ihnen gewählten Quality-of-Service-Klassen in Rechnung. Da die Anbieter von kommerziellen Diensten und Inhalten den größten Teil aller Quality-of-Service-Datenpakete versenden, werden sie auch den überwiegenden Teil der Kosten tragen. In welcher Weise sie diese refinanzieren, ist eine Frage ihres Geschäftsmodells und nicht Gegenstand des Zusammenschaltungsregimes.¹⁶

Es ist offensichtlich, dass ein Quality-of-Service-Konzept (bzw. ein wertbezogenes Priorisierungskonzept) vor allem bei einem SPNP-Zusammenschaltungsregime voll funktionsfähig wäre. Es stellt sich jedoch die Frage, ob SPNP nicht generell bei der IP-Zusammenschaltung angewendet werden sollte. Schließlich gilt ganz allgemein, dass die Kosten (in kurz- und langfristiger Hinsicht) von der Zahl der Datenpakete abhängen. Allerdings könnte dies den Effekt haben, dass insbesondere die nicht-kommerziellen Anbieter z.B. von Webseiten nicht bereit wären, die Kosten für den Abruf ihrer Seiten zu tragen. Dies könnte den Verlust des Zugangs zu vielen Webseiten und somit volkswirtschaftliche Ineffizienzen zur Folge haben. Dafür müsste ein kompensatorisches Abrechnungssystem installiert werden. Für den Quality-of-Service-Datenverkehr sollte das Sending-Party's-Network-Pays-Prinzip gelten.

Die praktische Umsetzung des Quality-of-Service-Konzeptes sollte dem Marktprozess überlassen bleiben, d.h. den kommerziellen Transaktionen zwischen den Netzbetreibern, den Dienste- und Inhabeanbietern, den Werbetreibenden und den Nutzern.¹⁷

¹⁶ Dass das Sending-Party's-Network-Pays-Prinzip für das Quality-of-Service-Konzept das geeignete Modell ist, ergibt sich auch aus folgender Überlegung: Nehmen wir an, wir hätten ein Receiving-Party-Pays-Modell im Internet. Dann gäbe es Anreize, grundsätzlich Datenpakete mit hohen Qualitätsklassen zu übergeben und dafür zu kassieren. Dies müsste der übernehmende Netzbetreiber seinerseits seinen Kunden in Rechnung stellen, was diese aber häufig ablehnen werden – insbesondere, wenn es sich um unerwünschte Datenpakete bzw. um eine unerwünscht hohe Qualitätsklasse handelt.

¹⁷ Vgl. hierzu verschiedene Geschäftsmodelle in W. Brenner, u.a., a.a.O.