Knieps, Günter

Conference Paper

Network Neutrality and the Evolution of the Internet

21st European Regional Conference of the International Telecommunications Society (ITS): "Telecommunications at New Crossroads: Changing Value Configurations, User Roles, and Regulation", Copenhagen, Denmark, 13th-15th September 2010, No. 19

Provided in Cooperation with:

International Telecommunications Society (ITS)

Suggested Citation: Knieps, Günter (2010) : Network Neutrality and the Evolution of the Internet, 21st European Regional Conference of the International Telecommunications Society (ITS): "Telecommunications at New Crossroads: Changing Value Configurations, User Roles, and Regulation", Copenhagen, Denmark, 13th-15th September 2010, No. 19, International Telecommunications Society (ITS), Calgary

This Version is available at:
https://hdl.handle.net/10419/44343

21st European Regional ITS Conference
Copenhagen, 13-15 September 2010

Günter Knieps

Network neutrality and the evolution of the Internet

Abstract

In order to create incentives for Internet traffic providers not to discriminate with respect to certain applications on the basis of network capacity requirements, the concept of market driven network neutrality is introduced. Its basic characteristics are that all applications are bearing the opportunity costs of the required traffic capacities. An economic framework for market driven network neutrality in broadband Internet is provided, consisting of congestion pricing and quality of service differentiation. However, network neutrality regulation with its reference point of the traditional TCP would result in regulatory micromanagement of traffic network management.

JEL codes: D85, L51, L86, L96

Key words:

Broadband Internet, network neutrality, quality of service differentiation, congestion pricing, interclass externality pricing, interconnection agreements

Prof. Dr. Günter Knieps, Albert-Ludwigs-Universität, Institute of Transport Economics and Regional Policy, Platz der Alten Synagoge, 79085 Freiburg i. Br., Germany, guenter.knieps@vwl.uni-freiburg.de

1. Mandatory versus market driven network neutrality

The evolution of the Internet is characterized by the transition from narrowband to broadband Internet. The narrowband Internet provided low speed access for services like e-mails, download of small documents etc. Thus, all applications were time insensitive and required similar low traffic capacity. In contrast, broadband Internet provides high speed access for a large scope of heterogeneous applications. Some applications, such as interactive video-gaming, voice over IP or video streaming are time sensitive, whereas other applications, such as content distribution or e-mailing are time insensitive. Some applications are capacity intensive, such as peer-to-peer exchange of videos or video on demand. Others need only little capacity, such as voice over IP or e-mails.

The recent world wide net neutrality debate has shifted public attention to the challenges faced by the traditional Internet which transports data packages on the basis of the best-effort transmission control protocol (TCP), assigning all data packets equal priority. TCP manages end-to-end connections by limiting the traffic offered by a sender when it detects congestion (Cerf, Kahn, 1974; Jacobson, 1988). Best-effort average traffic quality results endogenously, depending (positively) on capacity and (negatively) on traffic without quality of service guarantee of the data packet transmission (Cremer, Rey, Tirole, 2000, 455 f.). Due to the transition from narrowband access to broadband access, congestion and heterogeneous requirements for traffic qualities become increasingly important (Lehr, McKnight, 2002). As a consequence, best-effort average quality networks cannot be expected to provide the necessary allocation mechanisms to fulfil the heterogeneous requirements for traffic qualities.

Network neutrality is often considered as a rather vague concept with no generally accepted unique definition. However, generally the term “network neutrality” is used as a regulatory concept, addressing what deviations should be permitted from the traditional best-effort TCP (Schwartz, Weiser, 2009, 1). According to the OECD the notion of network neutrality “has recently been used to describe a data network that assigns all transmissions equal priority as they are passed along the network” (OECD, 2006, 3). This is a plea against traffic shaping within the Internet and the resultant challenge to the traditional best-effort transmission. According to the European Commission declaration on net neutrality the focus of net neutrality is “the creation of safeguard powers for national regulatory authorities to prevent

the degradation of services and the hindering or slowing down of traffic over public networks".[1] In October 2009 the U.S. Federal Communications Commission proposed network neutrality regulations in order to implement a principle of non-discrimination: "We understand the term "nondiscriminatory" to mean that a broadband Internet access service provider may not charge a content, application, or service provider for enhanced or prioritized access to the subscribers of the broadband Internet access service provider, …. We propose that this rule would not prevent a broadband Internet access service provider from charging subscribers different prices for different services" (FCC, 2009, 42).

The network neutrality debate is confronted with several fallacies: The first fallacy is the lack of differentiation between mandatory network neutrality and market driven network neutrality. Mandatory network neutrality consists of ex ante regulation of traffic management based on the traditional TCP. In contrast, market driven network neutrality means an entrepreneurial search for traffic allocation in such a way that there are no incentives for the Internet traffic service provider to discriminate between possible network applications on the basis of network capacity requirements. This is the case, if any application is charged according to the opportunity costs of traffic capacities it requires.

The second fallacy is the lack of differentiation between the impact of TCP in narrowband and broadband Internet. In the narrowband Internet the best-effort TCP fulfils the criteria of market driven network neutrality. Since all applications are homogeneous with respect to transmission quality and transmission capacity, the TCP creates no incentives to discriminate between different applications. In contrast, in the broadband Internet with its many and heterogeneous applications the TCP creates large discrimination potentials. On the one hand, low capacity applications are discriminated against by high capacity applications, on the other hand, time sensitive applications are discriminated against by time insensitive applications.

The third fallacy is to destroy market driven network neutrality by network neutrality regulation. Market driven network neutrality requires an evolutionary search for price and quality differentiation in order to reflect the opportunity costs of traffic capacity. Irrespective of how network neutrality regulation would be implemented in detail, it would limit the entrepreneurial flexibility with respect to the design of Internet architecture, traffic quality differentiation, and flexible transmission pricing.

---

[1] Commission declaration on net neutrality, Official Journal of the European Union, C 308/2, 18.12.2009

The fourth fallacy is the statutory prohibition under the heading of network neutrality of providers of Internet access services charging providers of Internet application services for enhanced or prioritized access. The focus of the debate is on whether Internet application providers should be protected from the abuse of market power of Internet access providers (Economides, 2008, 210; FCC, 2009, 30). Instead of forbidding price and quality differentiation of Internet traffic providers, it is necessary to regulate market power at its roots, meaning the remaining monopolistic bottleneck components within the local loop in the telecommunications network. The complementary Internet traffic markets are under the constraint of both active and potential competition. This includes active and potential competition between alternative Internet access service providers as well as between Internet backbone service providers (Faratin et al., 2007; Knieps, Zenhäusern, 2008, 127 ff.).

The fifth fallacy became known as the “dirt road” fallacy. FCC (2009, 30 f.) argued that price and quality discrimination would create incentives for Internet access service providers to reduce or fail to increase the transmission capacity available for standard best-effort Internet access service relative to higher quality services in order to increase their revenues. Contrary to this claim of ad hoc discrimination between high quality and low quality users, market driven network neutrality provides incentives to Internet traffic providers to offer a consistent choice of user charges and capacity allocation.

In the subsequent section 2 the discriminatory potentials of the TCP in broadband networks are pointed out. In order to avoid inefficient application restrictions, a shift from the traditional best-effort TCP towards more intelligent network architectures is required, allowing traffic shaping and prioritization of data packets. In section 3 the potentials of congestion pricing in broadband Internet are considered. Whereas under TCP each packet has an equal chance of getting through or being dropped, under congestion pricing dropping is not randomly but according to the willingness to pay indicated in the header of IP packets. Section 4 is devoted to quality differentiation of Internet traffic in order to allow priority pricing for time sensitive applications. This allows combining congestion pricing and quality of service differentiation. From the perspective of price differentiation of different service qualities, traffic prices should fulfil their incentive function in such a way that users with high preference for quality (low congestion) have the possibility to get premium quality transport. Finally, the role of quality of service based interconnection agreements based on interclass externalities is discussed.

2. The discrimination potentials of TCP in broadband Internet

A major characteristic of TCP is that it only controls the sending rate for a single traffic flow. Flow rate fairness is based on Jacobson's (1988) congestion and control mechanisms. TCP works by constantly increasing its rate until some link along the way to the receiver cannot handle the traffic flow and has to drop the packet. The sending computer halves its rate when retransmitting the missing packets. Since TCP controls each traffic flow separately, it cannot differentiate between heavy and light users of capacity. In particular, TCP does not take into account whether there are multiple TCP flows running on a single end-node. Thus, TCP does not take into account the aggregated usage of traffic capacity from a computer during a given time interval (Bauer et al., 2009, 3). Although the TCP tries to share the bit rate equally within the traffic flow, due to randomization of packet dropping it merely gives an illusion of fairness. A user gets multiple capacity shares if he runs multiple data flows at once. TCP gives much higher shares of capacity to the heavy users and much less to the light users (Briscoe, 2008).

A strategy of traffic service providers consists in network user restrictions with the goal of limiting the capacity consumption of heavy users. A survey of broadband usage restrictions has been provided by Wu (2003, 158ff.). Contractual restrictions on providing content effected the end user's sharing of content in contrast to simply downloading content. The restrictions favoured client-server applications over peer-to-peer applications. Other restrictions on applications have been prohibitions on applications for commercial use, restricting the number of computers that can be attached to a single connection and controlling the deployment of home wireless networks. Architectural restrictions may exist, due to the allocation of asymmetric bandwidth by designing networks to provide more downstream bandwidth than upstream, such that end-users can download more data packets than upload.

In the meantime, the degree of asymmetry of traffic capacity consumption between heavy and light users is enormously increasing. An illustrative example is provided by the Comcast-Case. Comcast is the leading provider of cable television and the number two provider of high-speed Internet connections in the U.S. The members of Free Press and Public Knowledge are subscribers of Comcast high-speed Internet access and many use peer-to-peer

applications through Comcast or another network provider. The formal complaint in October 2007 was: “Comcast is secretly degrading peer-to-peer protocols, threatening to undermine the Internet’s open and interconnected character, discourage broadband use, and crippling the innovation the Internet has made possible” (FCC, 2007, 1). In August 2008 the Federal Communications Commission (FCC) addressed the question whether it would be reasonable network management praxis for Comcast to interfere with its customers’ applications of BitTorrent (FCC, 2008). Unlike traditional methods of file sharing which require establishing a single TCP connection between a user’s computer and a single server, BitTorrent is a peer-to-peer networking protocol employing a decentralized distribution approach, all via TCP connections. Each computer in the BitTorrent swarm is able to download content from other computers in the swarm and each computer also uploads contents for the members of the peer-group. Moreover, a computer can download different portions of the same content from multiple computers simultaneously. While Comcast claimed that its interference into BitTorrent’s applications were required to manage scarce network capacity, the opponents claimed that Comcast had arbitrarily blocked subscribers’ access to applications, not applying a consistent congestion based approach. The FCC decided that Comcast’s network management practices in the BitTorrent-Case would be considered unreasonable and should not continue.[2] In this context the FCC suggested some ad-hoc solutions, such as capping the average user’s capacity and charging the heaviest users’ overage fees, or to throttle back the connection speed of high capacity users. However, these ad-hoc suggestions by the FCC for managing network traffic would not guarantee market driven network neutrality, because the opportunity costs of capacity usage were not consistently taken into account. In order to provide the proper incentives for network usage, congestion pricing models become relevant.

The provision of time sensitive applications needs guaranteed timely and steady packet delivery. The traditional TCP is not able to provide prioritization of data packets and quality of service guarantees. Thus, best-effort TCP transmission quality is not sufficient to guarantee the provisions of time sensitive applications; this entails a further important discrimination potential of TCP. In order to provide market driven network neutrality the transition to more

[2] In a petition for review of the FCC order the United States Court of Appeals decided that the Federal Communications Commission (FCC) would have no statutorily mandated responsibility to regulate the network management practices of an Internet service provider. United States Court of Appeals for the District of Columbia Circuit, decided April 6, 2010, No. 08-1291, Comcast Cooperation, Petitioner v. Federal Communications Commission and United States of America, Respondents, NBC Universal, et al., Intervenors, On Petition for Review of an Order of the Federal Communications Commission.

"intelligent" Internet architecture is necessary. Different technical solutions may be chosen to implement quality differentiation of transport of data packages.

3. Congestion pricing in broadband Internet

The basic goal of congestion based per packet pricing is to charge the user relative to the amount of congestion in the network. Congestion increases with the number of packets. When the network is uncongested, the cost of transporting an additional packet is minimal; when the network is congested, the cost of transporting an additional packet grows with the degree of congestion. The model of congestion pricing, applied to the problem of Internet traffic by MacKie-Mason and Varian (1995), is extended to the multi-channel case where the network is partitioned into separate channels. Since there is no priority implied among channels, from the perspective of the users only the degree of congestion within the different channels is relevant. It is shown that under competition price differentiation among channels with equal congestion cannot be stable. Thus, congestion pricing without traffic prioritization results in a homogeneous traffic quality.

3.1 Single-channel congestion pricing

The starting point is the well known concept of congestion pricing based on transportation economics which has been introduced to the field of Internet traffic by MacKie-Mason and Varian (1995, 288 ff.). All data packets pay a uniform congestion fee and are served without priority. Thus, socially optimal congestion pricing in a single channel network without quality of service differentiation has been derived. For each chosen capacity holds: when demand is low und congestion is low the packet price is low; when demand is high and congestion is high the packet price is high. If a network is strongly congested, the opportunity costs of an additional data packet are high and thereby optimal congestion prices are high. Thus, optimal prices reflect the level of congestion. Users not prepared to pay the congestion fee are excluded from data packet transmission. Optimal congestion fees (short run problem) and the optimal choice of capacity (long run problem) are derived simultaneously, supposing that several competing firms provide traffic services.

The allocation of traffic flows (short run) can take place over time scales ranging from seconds to minutes to days. Provisioning of network resources (in particular bandwidth) takes place over intervals of weeks and months (Gibbens et al., 2000, 2165).[3] Most congestion pricing theories assume an ex ante fixed price schedule for each (short run) time period, and thus a regular pattern of demand. MacKie-Mason and Varian (1995, 292) considered a kind of smart pricing where users do not pay the price that they actually bid, but rather pay for the packets at the market clearing price – reflecting the maximum bid amount of all packets that are not served – which will be lower than bids of all admitted packets. Thus, a uniform price for the transmission of data packets results within each short run time period.

3.2 Multi-channel congestion pricing

Network capacity (bandwidth, buffer space) is allocated to each channel separately. In particular, separate channels are not allowed to use spare resources from other channels. There are no quantifiable requirements with respect to delay or jitter associated with the forwarding of data packets. Thus, short term congestion may become relevant (Bouras, Sevasti, 2004, 1878). In the context of Internet traffic allocation problems it has been suggested (Cheng, Zhang, 2004, 375) that there is no priority implied among channels, although the resources allocated to each channel might be different. Only the degree of congestion is relevant. If two channels are equally congested, the users are indifferent. Channel numbering does not indicate any quality of service hierarchy.[4] From the cost side there are countervailing effects of channel separation. Since congestion increases with the number of packets, congestion costs are reduced by separating traffic into several channels. On the other hand, multiplexing advantages of an integrated network (with only one channel) are lost. Moreover, economies of scale with respect to bandwidth expansion may exist.[5]

Under competition on the markets for Internet traffic each traffic service provider makes his autonomous decisions. For each chosen number of channels the bandwidths and usage dependent prices for the different channels are derived simultaneously. There is no social planner to globally optimize capacities and prices for all traffic service networks. In the

[3] In contrast, investment decisions in transportation sectors (e.g. roads) have a much larger time horizon.
[4] In contrast, intermodal approaches to congestion fees for various transportation infrastructures examine alternative traffic modes (rail, road, etc.). Demand functions for different modes of traffic differ systematically and cross elasticities between different traffic modes are relevant (Braeutigam, 1979).
[5] This may be relevant in access service networks rather than backbone service networks.

following the allocation problems of an arbitrarily chosen traffic service provider under free entry are analyzed.

It is assumed that the network of a typical traffic service provider consists of *n* logically separated channels. Let $P_{it}(Q_{it})$ denote the inverse demand function for packet transmission in channel *i* in period *t* with traffic flow Q_{it}. We assume that demand is independent across time periods, because we do not aim to analyze intertemporal demand interdependencies. There is no reshifting of capacity from one time period to another between channels. It is reasonable to consider a single scalar that summarizes the resource requirement of any given channel (Paschalidis, 2000, 172). Capacity costs of the channel with bandwidth w_i are denoted $\rho_i(w_i)$.

Let $k_{it}(Q_{it}, w_i)$be the private (average) variable costs of a packet transmission within channel *i* with capacity w_i.

$\frac{\partial k_{it}}{\partial Q_{it}} > 0$ if capacity w_i remains constant, additional traffic within channel *i* will slow down every packet within this channel, thereby raising intra-channel externalities $\frac{\partial k_{it}}{\partial Q_{it}} Q_{it}$

$\frac{\partial k_{it}}{\partial w_i} < 0$ if traffic remains constant, additional bandwidth capacity of channel *i* will allow to speed up every packet in this channel.

Under competition each traffic service provider chooses channel capacities and packet prices in each channel in such a way as to maximize profit. The profit maximization problem is defined by:

$$\Pi(Q_{1t},...,Q_{nt}, w_1,..., w_n) = \sum_{t=1}^{T}\left[P_{1t}(Q_{1t})Q_{1t} + ... + P_{nt}(Q_{nt})Q_{nt}\right] - \sum_{t=1}^{T}\left(\sum_{i=1}^{n} k_{it}(Q_{it}, w_1)Q_{it}\right) - \sum_{i=1}^{n}\rho_i(w_i) \quad (1)$$

Necessary conditions for the maximum are derived by differentiating (1) with respect to $Q_{1t},...,Q_{nt}$ for each t=1,…,T and with respect to $w_1,...,w_n$ and setting each derivative to zero.[6]

The optimal pricing rule concerning the congestion fee for a packet transmission on channel i is given by:

$$(2) \qquad \tau_{it} = P_{it} - k_{it} = \frac{\partial k_{it}(\cdot, w_i)}{\partial Q_{it}} \cdot Q_{it} \quad t=1,\ldots,T; \qquad i=1,\ldots n$$

Increasing congestion results in higher packet charges.

The first best optimal rule for (bandwidth) capacity in channel i is given by:

$$(3) \qquad \frac{\partial \rho_i(w_i)}{\partial w_i} = \sum_{t=1}^{T} Q_{it} \cdot \frac{\partial k_{it}(Q_{it}, w_i)}{\partial w_i} \quad i=1,\ldots,n$$

Simultaneous solutions of equation (2) and (3) provide first-best allocation of traffic flows $Q_{1t}^*,...,Q_{nt}^* \qquad t=1,\ldots,T$ between the different channels as well as first-best capacity dimensions for each individual channel $w_1^*,...,w_n^*$

The following conclusions can be drawn:

$$(2') \qquad \tau_{it} = \frac{\partial k_{it}(\cdot, w_i^*)}{\partial Q_{it}} \cdot Q_{it}^*$$

Congestion externalities in channel i are increasing with the number of packets transmitted in this channel. Due to the bandwidth reservation for each channel, only externalities within separate channels are relevant, whereas externalities between channels (inter-channel externalities) do not occur. Optimal congestion fees increase with the level of congestion. Congestion fees within each channel i τ_{it} may vary over time, if demand varies over time.

$$(3') \qquad \frac{\partial \rho_i(w_i)}{\partial w_i}\Big|_{w_i = w_i^*} = \sum_{t=1}^{T} Q_{it} \cdot \frac{\partial k_{it}(Q_{it}, w_i^*)}{\partial w_i}$$

Capacity (bandwidth) in each channel i should be extended to the point where the marginal cost of an extra unit of capacity is equal to its marginal benefits of reduced congestion (for the

[6] The optimal pricing and investment rule under competition equals the socially optimal pricing and investment rule under maximisation of social welfare. For the case of one channel without quality of service differentiation see also McKie-Mason, Varian, 1995, 301 ff.

packet flow) within channel *i*. In particular, the extension of capacity until all congestion disappears (maximum quality) would not be optimal and lead to over-capacity.

Since there is no priority implied among channels, from the perspective of the users only the degree of congestion within the different channels is relevant. If congestion is different on different channels, the price must reflect this difference in congestion externalities. Since more congested channels are more expensive, incentives occur to switch to less congested and cheaper channels. Thus, in equilibrium traffic may split symmetrically to different channels with identical bandwidth and identical optimal congestion fees. Alternatively if bandwidth capacity varies among channels, the number of packets must also be adapted such that the level of congestion and the congestion prices remain identical.

Consider for example the case of two channels with $w_2<w_1$.

$$\tau_{1t} = P_{1t} - k_{1t} = \frac{\partial k_{1t}(\cdot, w_1)}{\partial Q_{1t}} \cdot Q_{1t} \qquad t=1,\ldots,T;$$

$$\tau_{2t} = P_{2t} - k_{2t} = \frac{\partial k_{2t}(\cdot, w_2)}{\partial Q_{2t}} \cdot Q_{2t} \qquad t=1,\ldots,T;$$

$$Q_{2t}^* < Q_{1t}^* \text{ is chosen s.t. } \tau_{1t} = \frac{\partial k_{1t}(\cdot, w_1)}{\partial Q_{1t}^*} \cdot Q_{1t}^* = \frac{\partial k_2(\cdot, w_2)}{\partial Q_{2t}^*} \cdot Q_{2t}^* = \tau_{2t}\ t=1,\ldots,T;$$

Optimal congestion prices of the smaller and the larger channel are identical. Optimal multi-channel congestion pricing and capacity choice within the network of each network service provider result in a complete absence of quality of service differentiation.

Under competition price differences between service providers can only be stable if they are completely caused by different congestion levels. Otherwise, users would switch to alternative traffic service providers offering the same degree of congestion at a lower price. However, specialised single quality networks with different congestion levels cannot survive under competition. Users would immediately switch from the more expensive and more congested network (with under-capacity) to a cheaper and less congested network (with over-capacity). In equilibrium optimal capacity and optimal congestion prices are identical. Thus, even the extension to multi-channel congestion pricing does not lead to quality of service differentiation with quality guarantees.

4. Quality of service differentiation based on interclass externality pricing

In the following quality of service differentiation within the Internet architecture of differentiated services (DiffServ) networks will be analyzed. Within DiffServ architecture data packets are classified into an exogenously determined number of classes at the network edge. Only the edge routers (ingress or egress edge routers) perform packet classification based on the priority information in the packet header, whereas core routers inside each DiffServ domain only deal with aggregated traffic for given service classes (Chen, Zhang, 2004, 370 ff.).

Within each DiffServ-enabled domain, servicing of packets by routers is performed according to traffic classes, not according to the flow to which they belong. Thus, within a DiffServ domain, all packets belonging to a given quality of service class receive the same treatment; in particular, within one service class no priority rule is applied. Due to its scalability the DiffServ framework is considered particularly suitable for larger packet transmission networks (Bouras, Sevasti, 2004, 1868 f.).

From the perspective of price differentiation of different service qualities, traffic prices should fulfil their incentive function in such a way that users with high preference for quality (low congestion) have the possibility to get premium traffic quality. Traffic prices should be monotonic with respect to decreasing quality of service classes, such that premium class packets have to pay the highest price. Whereas in congestion pricing models prices are high if congestion is high, quality of service price differentiation requires premium class users to pay high prices to enjoy absence of congestion. The purpose is to combine both approaches to develop incentive compatible quality of service differentiation within the networks of Internet traffic service providers under competition.

Data packets are transmitted on one channel only and no network partitioning is applied. The DiffServ scheduler router offers a predefined number of traffic classes using strict priority scheduling. A packet is inserted into the transmission buffer behind previous packets of the same traffic class but ahead of packets of a lower traffic class. The scheduler transmits the packets which are at the head of the buffer; packets at the tail of the buffer are dropped as soon as the buffer is full. Traffic quality can be measured by mean packet delay and packet loss. Applying the strict priority scheduler, traffic classes are monotone with respect to traffic

quality. Packets within a higher traffic class will be transported with lower delay and lower loss than packets within lower traffic classes (Jin, Jordan, 2005, 842).

Depending on the demand for high quality, medium quality and low quality traffic, quality of service in different classes results endogenously. The carrier provides a quality of service guarantee for the data packet transmission within a quality of service class – irrespective of the forwarding rate of lower classes (Borella et al., 1999, 279; Chen, Zhang, 2004, 374 f.) – defining a maximum allowable delay and packet loss.

4.1 Intraclass externalities versus interclass externalities

Consider the network of a typical traffic service provider with packet transmission in quality of service classes within one channel. Application of strict priority scheduling provides a structure for congestion externalities. It is important to differentiate between congestion externalities within a traffic class (intraclass externalities) and congestion externalities between traffic classes (interclass externalities). Intraclass externalities reflect the delays which an additional data packet causes for all other data packets of the same class.[7] Interclass externalities reflect the delays which an additional data packet imposes on the data packets in the other quality classes.

Consider the congestion pricing framework introduced in section 3 adapted to one channel only. Packets are classified and grouped into *n* different traffic classes. Let $P_{it}(Q_{it})$ denote the inverse demand for aggregated traffic in traffic class *i* in period *t* with traffic flow Q_{it}. $\rho(w)$ denotes the capacity costs of the channel with bandwidth *w*.

Let $k_{it}(Q_{1t},...,Q_{nt},w), \quad i=1,...,n$ be the private (average) variable costs of a packet transmission within traffic class *i*, which also depend on the flows of packets in other traffic classes.

$\frac{\partial k_{it}}{\partial w}<0, \quad i=1,...,n$ if traffic remains constant, additional bandwidth capacity will allow speeding up every packet.

[7] Within DiffServ architecture all data packets within the same class are treated equally, thus only average delay within a quality class is considered but not the individual delay of a packet depending on the position of the data packets within the queue at the router.

$\frac{\partial k_{jt}}{\partial Q_{it}} > 0$ if capacity w remains constant, additional traffic within traffic class *i* will slow down packets in its own class as well as in other service classes, thereby raising externality costs.

Externality costs caused by a given packet for all other packets may encompass other classes $j \neq i$ as well as its own class *i*. Externality costs $\sum_{j=1}^{n} \frac{\partial k_{jt}}{\partial Q_{it}} \cdot Q_{jt}$ consist of intraclass externalities $\frac{\partial k_{it}(\cdot, w)}{\partial Q_{it}} \cdot Q_{it}$ as well as interclass externalities: $\sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\partial k_{jt}(\cdot, w)}{\partial Q_{it}} \cdot Q_{jt}$

Due to the strict priority rule, interclass externalities are top-down / one sided. Only upper traffic classes cause interclass externalities for lower classes, but not vice versa. Interclass externality of class *i* to subsequent classes $i + 1,\ldots,n$ (externality to all subsequent traffic classes) is given by

$$\sum_{j=i+1}^{n} \frac{\partial k_{jt}(\cdot, w)}{\partial Q_{it}} \cdot Q_{jt} \qquad t = 1,\ldots,T$$

4.2 Price and quality differentiation based on interclass externalities pricing

The basic idea is to define a hierarchy of service classes, such that the highest quality class is the most expensive and the least congested. The prices are monotone decreasing with the number of the quality classes. Packets within the lowest quality class (with the highest congestion) should be charged the lowest price.

The starting point for the development of such price differentiation strategies are the opportunity costs of network usage due to congestion externalities. Under competition each traffic service provider chooses the channel capacity and packet prices in each traffic class.

The profit maximisation problem is defined by:

$$(4) \qquad \max_{(Q_{it},\ldots,Q_{nt},w)} \Pi = \sum_{t=1}^{T} \left[\sum_{i=1}^{n} P_{it}(Q_{it}) Q_{it} - \sum_{i=1}^{n} k_{it}(Q_{1t},\ldots Q_{nt}, w) Q_{it} \right] - \rho(w)$$

Necessary conditions for the maximum are derived by differentiating (4) with respect to $Q_{1t},\ldots,Q_{nt}$ for each $t=1,\ldots,T$ and with respect to *w*.

The optimal pricing rules concerning the congestion fee for a packet transmission within traffic class *i* is given by:

$$(5) \qquad \tau_{it} = P_{it} - k_{it} = \frac{k_{it}(\cdot, w)}{\partial Q_{it}} \cdot Q_{it} + \sum_{j=i+1}^{n} \frac{\partial k_j(\cdot, w)}{\partial Q_{it}} \cdot Q_{jt} \qquad t=1,\ldots,T; \qquad i=1,\ldots n$$

The optimal rule for bandwidth capacity *w* is given by:

$$(6) \qquad \rho'(w) = -\sum_{t=1}^{T} \sum_{i=1}^{n} \frac{\partial k_{it}(Q_{1t},...,Q_{nt}, w)}{\partial w} \cdot Q_{it}$$

Simultaneous solutions of equation (5) and (6) provide optimal allocation of traffic flows $Q_{1t}^{*},...,Q_{nt}^{*}$ $t=1,\ldots,T$ as well as optimal capacity dimension w^{*} .

Due to heterogeneous demand for different traffic qualities it is neither economically efficient nor incentive compatible to extend capacity, in such a way that for all users the highest quality class would be provided.[8] Instead, extension of capacity is beneficial until the marginal cost of an additional capacity unit is equal to the sum of marginal benefits of reduced opportunity costs of capacity usage in each quality class.

It is important to differentiate between the quality of traffic in a given class, which is determined by intraclass externalities, and the opportunity costs caused to the subsequent classes, which is determined by interclass externalities. Quality of service based price differentiation can be developed by focussing on the opportunity costs which the transmission of packets in high quality classes causes to the packets in subsequent lower classes. Even if the traffic in the premium class is low (and intraclass externality prices would be zero), the delay imposed by high priority traffic to the traffic of subsequent classes may be substantial. Opportunity costs of the transmission of data packets under strict priority scheduling are strongly determined by interclass externalities, the increasing delay of lower class packets due to the transmission of premium class packets. In contrast, intraclass externalities in upper classes are of less importance, given the quality standard is defined high enough, such that transportation quality is sufficient for all relevant applications independent of the traffic load in this class.

Congestion fees based on interclass externalities are monotone.

---

[8] The extra capacity required in order to transmit all data packets at premium quality has been analyzed in Yuksel et al., 2007.

$$\tau_1(t) > \tau_2(t) >,..., \tau_{n-1}(t) > 0$$

and the lowest traffic class with the highest intraclass externalities has a data packet transmission price of zero.

$$\tau_1(t) = \sum_{j=2}^{n} \frac{\partial k_{jt}(\cdot, w)}{\partial Q_{1t}} \cdot Q_{jt}$$

$$\tau_i(t) = \sum_{j=i+1}^{n} \frac{\partial k_{jt}(\cdot, w)}{\partial Q_{it}} \cdot Q_{jt}$$

$$\vdots$$

$$\tau_{n-1}(t) = \frac{\partial k_{nt}(\cdot, w)}{\partial Q_{n-1}} \cdot Q_{nt}$$

$$\tau_n(t) = 0$$

Quality of service based price differentiation according to interclass externality pricing provides incentive compatible prices. Users with higher preference for priority traffic services have the possibility to choose a higher service class and thereby to pay a higher price for high quality (less congested) traffic services. Moreover, its advantage is that important elements of congestion pricing are included. High priority users have to compensate for additional traffic delay imposed on lower classes. These opportunity costs can hardly be ignored when establishing competitive pricing strategies and should therefore survive under competition and free entry in Internet traffic service markets.

Congestion fees based on interclass externalities result in a price of zero for the lowest quality traffic class *n*. Since in the lowest traffic class *n* no quality guarantee is provided, intraclass externality pricing may be applied in class *n* to solve the allocation problem of packet dropping and socially inefficient delay. The intraclass externality price in class *n* is always lower than the interclass externality price of class *n*-1. Due to top priority scheduling an additional package in class *n*-1 causes a larger delay on the packages in class *n* than an additional packet in class *n*. Thus, monotony of traffic class prices is guaranteed.

In contrast to the claimed ad hoc allocation of capacity between high quality and low quality classes in the above mentioned “dirt road” fallacy, capacity is allocated endogenously between the different quality classes according to the degree of heterogeneity between the different consumers. Since the capacity is chosen endogenously, an increase in the demand for

high quality transmission with subsequent high opportunity costs of additional high quality traffic will lead to an incentive compatible capacity extension.[9]

Price differentiation based on interclass externalities should allow viability of the traffic service providers under competition. Since intraclass externalities are neglected in the packet prices, and economies of scale with respect to bandwidth expansion may occur (in particular in access service networks), a competitive search for cost covering tariff structures becomes relevant.[10] One possibility is to apply mark-ups on interclass externalities based on price elasticities of demand for the transmission in different service classes (endogenous Ramsey pricing).[11] Under the assumption that price elasticities for the demand in higher traffic classes are lower, the monotony of packet charges is still guaranteed. As an alternative, in particular in access service networks two-part tariffs can be applied with a fixed connection charge and a variable data packet transmission fee based on interclass externalities.

4.3 Quality of service based interconnection agreements based on interclass externalities

Increasing congestion and asymmetric traffic flows and increasing demand for traffic quality differentiation result in a need for more complex interconnection contracts among different network carriers. New forms of interconnection arrangements, such as partial transit, paid peering, secondary peering, have arisen. By means of secondary peering arrangements the participating networks directly exchange traffic destined for each other's customers bypassing the universal connectivity providing core networks (Besen et al., 2001, 292f; Laffont et al., 2001, 288). Paid peering arrangements reflect the increasing asymmetry of interconnection traffic by allowing side payments between peering partners. In contrast to the traditional full transit arrangements, partial transit arrangements only guarantee interconnection to a subset of Internet users. As a consequence, Internet interconnection agreements are becoming more complex (Faratin et al., 2007). These innovative interconnection solutions are still based on average transportation quality, ignoring the potentials of quality of service differentiation

[9] For a detailed criticism of the "dirt road" fallacy see Sidak, Teece, 2010, 56 ff.
[10] It is well known from transportation economics that for the case of constant returns to scale with respect to capacity expansion, optimal congestion fees cover the capacity costs. However, if there are economies of scale with respect to capacity expansion, optimal congestion fees result in a deficit (Mohring, Harwitz, 1962, 81-86).
[11] For the concept of Ramsey pricing for competitive services (endogenous Ramsey pricing), see e.g. Baumol, Willig, 1983, 36 ff.

based on traffic classes. As a consequence, there is an increasing need to bargain on quality of service based interconnections arrangements.

Global service level agreements based on universal quality of service have been considered as one possible solution. A global market would be created for all networks in order to provide a quality of service guaranteed interconnection service (Li et al., 2004, 93). An alternative proposal is to search for an agreement on one global quality of service standard scheme (Borella et al., 1999, 287). However, due to the large number of possible priority schemes and the large number of participants involved, achieving a global bargaining solution seems unrealistic.

Instead, an evolutionary search for bilateral and multilateral quality of service based service level agreements should be initiated and should not be disturbed by government regulations (e.g. prescribing specific quality of service standards). Thus, bilateral or multilateral interconnection agreements among Internet traffic service providers taking into account different quality of service classes can develop. Quality differentiated service level agreements by means of interclass externality pricing provide compensation of the opportunity costs for offering premium services. Interconnection charges according to interclass externalities are incentive compatible, because compensation of the marginal congestion costs of delay, imposed by premium class traffic on lower quality traffic, is provided.

References

Bauer, S., Clark, D., Lehr, W. (2009), The Evolution of Internet Congestion, Massachusetts Institute of Technology, 37th Research Conference on Communication, Information and Internet Policy (www.tprcweb.com), Arlington, VA, September 2009

Baumol W.J., Willig, R.D. (1983), Pricing Issues in the Deregulation of Railroad Rates, in: J. Finsinger (ed.), Economic Analysis of Regulated Markets, McMillan, London, 11-47

Besen, S., Milgrom, P., Mitchell, B., Srinagesh, P. (2001), Advances in Routing Technologies and Internet Peering Agreements, American Economic Review, 91/2, Papers and Proceedings, 292-296

Borella, M.S., Upadhyay, V., Sidhu, I. (1999), Pricing Framework for a Differential Service Internet, European Transactions on Telecommunications, 10/3, 275-288

Bouras, C., Sevasti, A. (2004), SLA-based QoS pricing in DiffServ networks, Computer Communications 27, 1868-1880

Braeutigam, R.R. (1979), Optimal Pricing in the Intermodal Competition, American Economic Review, 69, 38-49

Briscoe, B. (2008), Internet – Fairer is Faster, in Proceedings: Qualität im Internet; 41. Freiburger Verkehrsseminar (Quality on the Internet, 41st Freiburger Traffic Seminar), Sep. 2008, 23-68

Cerf, V.G., Kahn, R.E. (1974), A Protocol for Packet Network Intercommunication, IEEE Trans on Comms, Com-22/5

Chen, J.-C., Zhang, T. (2004), IP-Based Next-Generation Wireless Networks: Systems, Architectures, and Protocols, John Wiley & Sons

Cremer, J., Rey, P., Tirole, J. (2000), Connectivity in the commercial Internet, Journal of Industrial Economics 48, 433-472

Economides, N. (2008), "Net Neutrality," Non-Discrimination and Digital Distribution of Content Through the Internet, I/S: A Journal of Law and Policy for the Information Society 4/2, 209-233.

Faratin, P., Clark, D., Gilmore, P., Bauer, S., Berger, A., Lehr, W. (2007), Complexity of Internet Interconnections: Technology, Incentives and Implications for Policy. Paper prepared for 35th Annual Telecommunications Policy Research Conference, George Mason University, September 2007

Federal Communications Commission (FCC) (2007), Formal Complaint of Free Press and Public Knowledge against Comcast Cooperation for secretly degrading peer-to-peer applications, November 1, 2007

Federal Communications Commission (FCC) (2008), In the Matter of Formal Complaint of Free Press and Public Knowledge against Comcast Corporation for Secretly Degrading Peer-to-Peer Applications, File No. EB-08-IH-1518 (FCC 08-183), Memorandum Opinion and Order, Adopted: August 1, 2008

Federal Communications Commission (FCC) (2009), In the matter of preserving the open Internet; Broadband industry practices, Notice of proposed rulemaking, GN Docket No. 09-191, WC Docket No.07-52, FCC 09-93, Adopted: October 22, 2009

Gibbens, R.J., Sargood, S.K., Kelly, F.P., Azmoodeh, H., Macfadyen, R., Macfadyen, N. (2000), An approach to service level agreements for IP networks with differentiated services, Philosophical Transactions of The Royal Society A (Phil. Trans. R. Soc. Lond. A), 358, 2165-2182

Jacobson, V. (1988), Congestion avoidance and control, in Proceedings of SIGCOMM '88 (Standford, CA, Aug. 1988) ACM.

Jin, N., Jordan, S. (2005), Information exchange in diffServ pricing, IEEE Globecom 2005 proceeding, 841-846

Knieps, G., Zenhäusern, P. (2008), The fallacies of network neutrality regulation, Competition and Regulation in Network Industries, 9/2, 119-134

Laffont, J.-J., Marcus, S., Rey, P., Tirole, J. (2001) Internet Peering, American Economic Review 91/2, Papers and Proceedings, 287-291

Lehr, W., McKnight, L.W. (2002), Show me the money: contracts and agents in service level agreement markets, info, 4/1, 24-36

Li, T., Iraqi, Y., Boutaba, R. (2004), Pricing and admission control for QoS-enabled Internet, Computer Networks, 46/1, 87-110

MacKie-Mason, J.K., Varian, H.R. (1995), Pricing the Internet, in W. Sichel and D.L. Alexander, editors, Public Access to the Internet. MIT Press, Cambridge, 1995, 269-314

Mohring, H., Harwitz, M. (1962), Highway Benefits: An Analytical Framework, Northwestern University Press, Evanston, Il.

OECD (2006), Internet Neutrality: A Policy Overview, OECD Working Party on Telecommunication and Information Services Policies, OECD Report, DSTI/ICCP/TISP(2006)4/FINAL, May, 29-30, Dublin

Paschalidis, I.Ch. (2000), Congestion-Dependent Pricing on Network Services, EIII/ACM Transactions on Networking, 8/2, 171-184

Schwartz, M., Weiser, P.J. (2009), Introduction to a Special Issue on Network Neutrality, Review of Network Economics, 8/1, 1-12

Sidak, J.G., Teece, D. (2010), Innovation Spillovers and the “Dirt Road” Fallacy: The intellectual Bankruptcy of Banning Optional Transactions for Enhanced Delivery over the Internet, Journal of Competition Law & Economics, 6, 1-64

Wu, T. (2003), Network Neutrality, Broadband Discrimination, Journal on Telecommunication and High Technology Law, 2, 141-179

Yuksel, M., Ramakrishnan, K.K., Kalyanaraman, S., Houle, J.D., Sadhvani, R. (2007), Value of Supporting Class-of-Service in IP Backbones, IEEE, 109-112