DOI: 10.1007/s10272-008-0242-2

Jörn Kruse*

Network Neutrality and Quality of Service

One of the most important factors for the tremendous worldwide success of the internet is that all the different services are transformed into homogeneous data packets for the transport over the IP networks. They are handled by universal protocols (TCP, IP) and sent (by routers as switching devices) over universal network infrastructures.

It used to be the common procedure that all data packets, whatever service or content they might belong to, would be treated as equal at the different routers on their way to their destination. Thus, if complications such as traffic congestion occur each data packet has the same likelihood of going through, being withheld, or even dismissed. This is called the network neutrality principle.

Some people in the internet community, especially in the USA, regard network neutrality as a basic element of a "democratic internet" with equal access for everybody. Legislation has been proposed that would make any deviation from network neutrality by internet service providers (ISPs) or other network operators unlawful. This started a controversial debate,[1] with political, economic and almost ideological arguments, and significantly supported by the economic interests of users, network operators and service and content providers respectively.

The network neutrality regulation problem contains basically two different and separable issues: (1) discrimination and (2) quality of service.

(1) The proponents of network neutrality regulation argue that network operators and ISPs might use their control over routers and transmission networks to slow down or block certain data packets in order to discriminate competing services. If, for example, telecommunication network operators blocked data packets of Voice-over-IP services that might substitute their own telephone services, this would not only discriminate against specific firms, but also reduce competition and economic welfare. Technically, this would not be a problem. Although data packets are homogeneous with respect to switching and transmission treatment, type, source, and destination can be revealed and data packets be handled differently if a network operator prefers to do so.

Under the conditions of competition between networks, as is common in European countries (in contrast to the USA)[2] a network operator would not have an economic incentive to do so, because he would drive himself out of the market. Such network behaviour seems to be transparent not only to the service and content providers but also to the internet user community. Market reaction would follow promptly if network operators discriminated against specific services.

Beyond that, the above mentioned discrimination would be an offence against European competition law and would certainly be prosecuted if it occured. The discrimination issue will not be discussed in the following.

(2) Quality of service. A network neutrality regulation would not allow differentiation between data packets according to their economic value in the case of congestion. When not every incoming data packet can be conveyed instantly, some kind of rationing has to be applied. If rationing is done by chance (as under a network neutrality regime) instead of following the economic value of congestion free transmission, the results will be inefficient. This problem and its potential remedies will be addressed below.

Internet Congestion and Quality

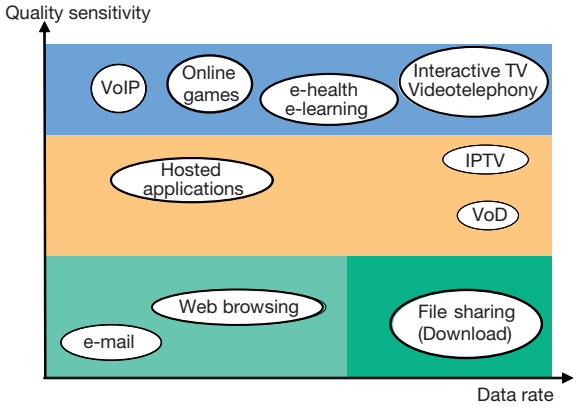
Internet traffic is increasing dramatically due to additional users and, especially, to high-data-rate applications such as peer-to-peer (P2P) file sharing etc. Although network operators constantly increase their router and transmission capacities, congestion occurs regularly. In economic terminology, congestion is characterised by partial rivalry, which is defined by the fact that although additional users do not exclude oth-

* Helmut Schmidt University – University of the Federal Armed Forces Hamburg, Germany.

[1] Cf. J. Gregory Sidak: A Consumer-Welfare Approach to Network Neutrality Regulation of the Internet, in: Journal of Competition Law and Economics, Vol. 2, No. 3, September 2006, pp. 349-474; Barbara van Schewick: Towards an Economic Framework for Network Neutrality Regulation, in: Journal on Telecommunications and High Technology Law, Vol. 5, No. 2, 2007, pp. 329-392; R. Litan, H. Singer: Unintended Consequences of Net Neutrality Regulation, in: Journal on Telecommunications and High Technology Law, Vol. 5, No. 3, 2007, pp. 533-572.

[2] Cf. J. Scott Marcus: Network Neutrality: The Roots of the Debate in the United States (in this volume).

Figure 1
Quality Sensitivity and Data Rate of Selected Services



ers, congestion affects all users negatively by reducing their transport service quality.

When the number of data packets exceeds router capacity, additional packets will be intermediately stored and, with more traffic coming in, will finally be dropped altogether. Congestion is leading to increased delay, jitter and packet-loss, which may significantly reduce the quality of certain applications. Among these are interactive services like VoIP, online gaming etc. and other on time services like internet television.

Although data packets are homogenuous at the transmission and switching level, they are not at all homogeneous at the service level, but differ dramatically with respect to at least three relevant parameters: (1) data rate, (2) quality sensitivity, and (3) economic value.

(1) Individual services have very different data rates which are defined by the number of data packets per time unit. Certain services send extraordinarily large numbers of data packets over the internet and thus play a particularly significant role in the reduction in quality for all services. Such traffic often has to do with P2P file sharing platforms. It includes (frequently high volume) downloads (and uploads) of software, music and videos, a large percentage of which is basically illegal because of copyright violations. Other services such as e-mail, web browsing etc. involve comparatively small numbers of data packets.

If we look at total internet traffic, P2P file sharing is responsible for a large proportion of the internet workload. In Germany, peer-to-peer traffic accounts for 69.25%, web browsing for 10.05%, media streaming (including YouTube etc.) for 7.75%, VoIP for 0.92%, email for 0.37%.[3]

(2) Quality sensitivity stands for the congestion effect on the quality of a specific service from the viewpoint of the consumers and their willingness to pay. The reductions in quality due to congestion (delay, jitter, packet loss) differ extremely according to the service involved.

Some services will not be affected at all, or only by extremely large network failures. These include elastic services where lost packets will be reordered from the source, such as email, web browsing, downloads and filesharing.

The qualities of other services are severely affected in the case of congestion. These include interactive services (e.g. voice over IP, online gaming etc.) as well as many business applications and internet television.

A selection of internet services with respect to data rate and quality sensitivity is shown in Figure 1.

(3) The economic value of a specific service is based on the users' willingness to pay per data packet. In economic terms, the value of a service is represented by its welfare measured by the sum of consumers' and producers' surplus. From the business viewpoint, it might be measured by the total revenues derived from a specific service.

Economists are familiar with the "tragedy of the commons" in connection with commonly owned pasture. The tragedy is the inefficient outcome as a result of poorly defined property rights. Today's internet is moving into a similar situation, although the culprits and the victims are different entities. For illustration, let us look at the partial rivalry between two services, IS_1 and IS_2, which both use the common resource "internet capacity". IS_1 is a high data rate service with very low quality sensitivity. It has very little economic value. P2P file sharing is the most relevant example. IS_2 is a highly quality sensitive and valuable service. Examples include interactive applications, such as VoIP, online games, and a number of business applications (credit card authorisation).

As mentioned above, P2P accounts for more than two thirds of internet traffic. It is important to note that

[3] The P2P file sharing percentage in internet traffic is 83.5% for Eastern Europe, 63.9% for Southern Europe, 49% for the Middle East and 57.2% for Australia. For more details cf. Hendrik Schulze, K. Mochalski: The Impact of P2P File Sharing, Voice over IP, Skype, Joost, Instant Messaging, One-Click Hosting and Media Streamimg such as YouTube on the Internet, Ipoque Internet Study 2007, p. 2.

the marginal costs for users are zero because of flat rates, and the growth rates of these services are still remarkable. With further development, IS_2 services will suffer more and more from IS_1 growth because of quality problems and, as a result, declining demand.[4] There is an obvious tendency for certain valuable, quality sensitive services to be driven out of the market by filesharing traffic which has low economic value.

In general, considering the specific incentive structures resulting from flatrates and the volumes of high data rate services which themselves are quality insensitive, it is realistic that high value services will be crowded out. This results in economic inefficiency due to decreased consumer and producer surplus as well as limited business revenues in this market – and in others where internet traffic is an important input.

Additionally, innovative services requiring high quality standards may not be developed at all even if they would have high economic value. These consequences are detrimental to economic development and will have a negative impact on growth and employment etc. which can be traced back to both net neutrality and flat rates.

Internet Pricing

As mentioned above, the flat rates for internet end users are part of the problem. Thus, implementing volume-based internet transmission prices (per data packet) will be part of the solution. If we apply a standard congestion pricing model to the internet, the welfare maximising price and volume are determined by the point of intersection of the quality adjusted demand function with the function of the marginal congestion externalities.[5]

Although this would not yet be the optimal solution to our problem (see below), it would certainly be a large step towards efficiency. The price per data packet reflects the opportunity costs and differentiates between high value and low value services using the consumers' actual willingness to pay as an appropriate criterion.

However, even this theoretically quite simple solution would be hard to implement in reality. Since data volume fluctuates significantly over time, so do externality and demand functions and, therefore, the appropriate peak load prices. In order to be effective for data volume and the resulting service qualities, these functions would have to be anticipated to come up with the "right congestion prices" (which will be zero most of the time). These could solve the quality problem, assuming that the senders of the data packets would indeed adequately respond to these prices.

Both assumptions (efficient *ex ante* prices and adequate impact on volume) are unrealistic if we take the extremely short-run usage patterns in the internet into account. But even a rather crude peak load pricing scheme (with significant prices at prime times and lower or zero prices at other times) would certainly be more efficient than flat rates.

However, even if it were possible to actually install a peak load pricing scheme which would always meet the above mentioned "externality-equals-demand" condition (and volume would adjust), this would still not be efficient under the specific conditions of internet technology and usage. This will be outlined further below.

Over-provisioning and Network Separation

In principle, it is possible to avoid a majority of congestion problems if appropriate investments are made for higher capacities of routers and transmission lines.[6] Capacity is defined as the maximum quantity of data packets for a very small time slot that can be handled without any delay, jitter or packet loss. It could be worth considering building large reserve capacities and network redundancies so that all data packets can be forwarded immediately at any time, even in the event of extremely short-run peak loads. This over-capacity strategy is called over-provisioning.

Sizing the capacities for a potential maximum peak load requires high reserve capacities and causes correspondingly high costs for the network operators. This raises the question, firstly, whether such capacities are economically efficient and, secondly, whether the network operators have appropriate economic incentives to make the required investments.

[4] For details cf. Jörn Kruse: Crowding-Out bei Überlast im Internet, Helmut-Schmidt-University, Economic Discussion Papers 72, November 2007, download http://www.hsu-hh.de/kruse/index.

[5] The quality adjusted demand function represents the users' willingness to pay for the actual internet service, including the possibly reduced quality. Beyond capacity, each additional data packet causes lower quality for other users. These negative effects on others, however, are not taken into account in individual usage decisions, which is why they are known as "congestion externalities". The function of the marginal congestion externalities covers these negative impacts on all other internet users. Cf. Jörn Kruse, op. cit.

[6] Nevertheless, congestion and reductions in quality may also occur due to capacity-induced overload resulting from unexpected network failures as a result of network breakdowns, earthquakes or other disasters, when the workload of the failing capacities has to be additionally managed by other routers and lines, if there are any.

What is the optimum capacity, taking congestion-induced quality reductions into account? The smaller the capacity, the more likely is it that impairments will occur at peak times, and the more severe they will be for a given number of data packets. The optimum solution can be derived by a long-term analysis with internet capacity as the relevant variable. Let us assume that an allocatively efficient uniform volume-based internet usage price is generally applied. A specific long-term utility function represents the relationship between capacity and total utility, allowing the derivation of the long-term marginal utility curve.[7] Its point of intersection with the long-term marginal cost curve determines the optimum capacity.

Since it can be assumed that long-term marginal utility is continuously decreasing towards zero and the long-term marginal costs of expanding capacity are positive throughout, the socially optimal capacity is always smaller than the congestion-free capacity. Thus over-provisioning internet capacity is economically inefficient and would be a waste of resources, even under the assumption of efficient prices.

The individual network operators would generally have no incentive to invest in additional infrastructure if the foreseeable capacity is larger than the optimal one, since their outlays could not be amortised.

Things get even worse if we assume that end-user flat rates prevail. Under these conditions, even more high data rate, quality insensitive, low-value applications and content (high-definition videos etc.) will be developed and used by even more consumers. Thus, under these circumstances, striving for over-provisioning would be a bottomless barrel even in the medium-term future, and it would be economically irrational. This does not seem to be far away from our present situation (with over-provisioning, flat rates and network neutrality).

Let us assume that flat rates and network neutrality continue to prevail and congestion is appearing more often and more heavily, such that quality-sensitive services suffer and are ultimately driven out of the market or are unable to develop and prosper. It could then be expected that large providers of economically high-value and quality-sensitive services consider building their own IP networks in order to be independent of the low-quality universal internet and able to adequately market their services and contents. Also, they might contract with existing network operators to implement separate and exclusive infrastructure solutions for quality-sensitive services. Individual service providers could reserve a specific proprietary capacity which is always available to them. The different services or providers would therefore be treated differently according to their willingness to pay.

This would, however, mean that a large proportion of the capacities would not be used most of the time and the required overall capacity (and hence also the investments and costs) would be higher than otherwise. This would lead to higher average prices. Such a solution would be technically inefficient. Moreover, the internet in its present form would be considerably changed and would cease to be a universal network.

To put this differently, if government opted for network neutrality regulation and was unable to come up with adequate quality solutions (see below), the market forces would. The network operators would have strong incentives to look for solutions that would make it possible for high-value service and content providers to market their products via IP networks.

As a result, all IP networks as a whole would not be "neutral" at all, economically inefficient, and detrimental to competition.

Priority Pricing

The conventional congestion models suggest prices which theoretically seem to solve the partial rivalry rationing problem. They are uniform prices in the sense that each data packet in a given time slot pays the same price. However, this does not take the specific internet technology and congestion procedures into account, which are based on the different substitutability of time slots among individual services.

In the internet, the congestion periods are often extremely short. They may last for seconds, while after that time router and line capacities may be available again. If data packets of quality-insensitive services are withheld during those short intervals, there will be no quality reductions.[8] If these packets wait until router capacity is available again, they will not cause any congestion externalities and their specific short-run marginal cost will be zero. Under these conditions it would be economically inefficient to exclude these packets from transport over the internet by a price which includes congestion externality markups.

Taking this into account, the internet congestion problem can be seen as being merely a problem of

[7] Cf. Jörn Kruse, op. cit.

[8] The same holds for packet losses when the protocols take care of the lost packets by reordering them from the source. This is the case not only for emails, but also for downloads and P2P file sharing services.

Table 1
Quality Class Concept with Four Quality Classes

Quality class	Typical services	Technical QoS parameters	
Interactive	Voice telephony/conferencing Video telephony/conferencing Online gaming Interactive TV feedback	Bandwidth: Delay (one way): Jitter: Packet Loss:	16 - 500 Kbps 100 - 200 ms < 30 ms < 1 %
Multimedia	Broadcast TV Video on demand Streaming audio Internet radio Voice messaging	Bandwidth: Delay (one way): Jitter: Packet Loss:	384 Kbps - 14 Mbps 400 - 1000 ms < 1000 ms < 0.1 %
Critical	Business Applications e.g. SAP, eHealth	Bandwidth: Delay (one way): Jitter: Packet Loss:	16 Kbps - 16 Mbps 100 - 200 ms < 100 ms < 0.1 %
Best Effort	Email Web browsing P2P Internet downloads	Bandwidth: Delay (one way): Jitter: Packet Loss:	up to line rate < 2000 ms n. a. n. a.

Source: Walter Brenner, M. Dous, R. Zarnekow, J. Kruse: Quality in the Internet. Technical and economic development prospects, University of St. Gallen 2007 (German version download: http://www.hsu-hh.de/kruse/index).

adequate prioritisation of data packets at times of congestion. It requires priorities such that (a) data packets of quality sensitive, high-value services will be conveyed instantly, while (b) data packets of quality-insensitive, low-value services would possibly have to wait and only be forwarded with some delay or have to be replaced by the service protocol later on.

Technically, the internet infrastructure (routers) already provides for the introduction of packet prioritisation. The headers of the data packets may contain specific priority information which can be used by the routers for setting differentiated priorities.

The most appropriate method for assigning priorities to individual data packets is by an adequate pricing mechanism using willingness to pay. "Priority pricing" is characterised by such a specific pricing mechanism assigning the "right to be served with a certain priority". The price for transmission with that priority applies, no matter whether congestion actually occurs or not.

The service and content providers' willingness to pay for high priority will depend mainly on two factors: (1) the quality sensitivity of the individual services and (2) the willingness to pay on the part of the users of those services. Only providers of quality-sensitive services will have any reason whatsoever to pay for priority since only they will gain any advantage from it. The providers of quality-insensitive services (email, web browsing, downloads etc.) will be adequately served with best effort and will thus obtain cheap service. The providers of quality-sensitive services will only be willing to pay for priority of the data packets if the users of the services (or indirectly the advertisers) on their part are also willing to pay for the quality of these services. This means that generally only high-value services will choose a high priority.

In competitive network markets a specific quality of service (QoS) system is likely to emerge with specified quality classes and different prices in order to deal adequately with heterogenuous services with different sensitivities with regard to delay, jitter, and packet loss. An example for a system with four specified quality classes (interactive, multimedia, critical, and best effort) is outlined in Table 1, which is taken from a recent study.

Such a market driven quality class model will generally result in an economically efficient rationing of scarce router capacity according to the economic value of the congestion-free services, and thus avoid the above-mentioned crowding-out problems.

Should government decide in favour of network neutrality regulation, an economically efficient QoS-concept could not be implemented. In a quality of service system, all users with the same willingness to pay will be treated equally. Since a network neutrality regulation is economically inefficient, it should certainly not be implemented.

Even if QoS concepts are not used in the universal internet (possibly because of a network neutrality reg-

ulation), the network operators will be bound to look for solutions that would make it possible for high-value service and content providers to market their offerings via IP networks. If adequate solutions cannot be found through any kind of packet prioritisation, it can be expected that individual providers of economically high-value and quality-sensitive services will implement separate infrastructure solutions (of a proprietary nature) for quality-sensitive services. This would not only be technically inefficient. The conventional internet would cease to be a universal network. And network neutrality would not be achieved either, if the IP network as a whole is considered.

Sending Party's Network Pays

It is important for a quality-of-service concept to develop and implement a price model for the interconnection between different networks that is not only volume-based but also explicitly quality-based. This means that interconnection tariffs must depend on whether or not the network complies reliably with agreed quality parameters. Without quality elements of this kind in the interconnection pricing, certain service-specific quality requirements would not be possible beyond the network borders.

The starting point is the labelling of the data packets with the chosen QoS-class by the sender or by its ISP. The sending ISP (and any other network operator) must ensure that when traffic is handed over to the next network that operator treats the data packets in such a way that the quality parameters are met. It will therefore only pass on its quality traffic to networks that comply with these quality standards.

Since the permanent implementation of defined quality parameters causes higher costs than best-effort traffic, a network operator will only guarantee this quality if the forwarding network pays appropriate prices which are higher than those for best-effort. In other words, using the sending party's network pays principle (SPNP) is a precondition for successful implementation of a quality-of-service concept.

The individual "original ISP" will bill its customers accordingly. The commercial service and content providers will send the majority of all QoS data packets, so they will also bear the bulk of the costs. How they refinance this, is a question of their business model. Most of the other traffic (emails, web browsing, downloads etc.) will use the best-effort class, which will be cheap.