INTERNET PRICING MODELS:
APPLICATION FOR UKRAINE

by

Volodymyr Vakhitov

A thesis submitted in partial
fulfillment of the requirements for the
degree of

Master of Arts

National University of Kyiv-Mohyla
Academy

1999

Approved by ____
Chairperson of Supervisory Committee

____

____

____

Program Authorized
to Offer Degree ____

Date ____

National University of Kyiv-Mohyla Academy

Abstract

INTERNET PRICING MODELS:
APPLICATION FOR UKRAINE

by

Volodymyr Vakhitov

Chairperson of the Supervisory Committee: Professor ________________
Department of Economics

This work deals with pricing models of Internet services. The rapid growth of the Internet along with the development of new multimedia applications creates a problem of congestion that may threaten free transmission of the data worldwide. Although technical solutions of this problem are the most desirable, they are not always feasible. New area of Economics, namely, Internet Economics is called on to provide an economic response to the congestion problem to reallocate scarce network resources efficiently. Two basic pricing models of the Internet access are proposed, usage-sensitive pricing and flat-rate pricing. The first class of models is based on priority pricing or peak-load pricing. They are efficient in solving congestion in cases of small networks or intranets, but for larger networks they involve complicated billing mechanisms. The second class of models is easier to implement in practice. They allow to plan the users' expenditures on the Internet use, but their major drawback is insensitivity to congestion. The aim of the work is to show that currently flat rate pricing models are more appropriate for Ukraine due to the lack of the proper infrastructure, first of all in the field of the financial sector. Nonetheless usage-sensitive pricing is more efficient and will possibly become the long-run solution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF NETWORK ACRONYMS

ISP – Internet Service Provider

IT – Infromation Technology

LAN – Local Area Network

QoS – Qality of Service

SOHO – «Small Office, Home Office»

TCP/IP – Internet Protocol

ACKNOWLEDGMENTS

The author wishes to Gene Ellis, EERC Research Director, for he has inspired the choice of the topic and helped to find initial bibligraphy. My special thanks are to my advisor, Klaus Wallner, Assistant Professor with Stockholm School of Economics, for his valuable guidelines throughout the course of the work and his constant support of the process. I would also like to thank to Lance Brannman, EERC Visiting Faculty, who became my another local advisor and provided me with many useful comments. Since the work involves a lot of specific technical information, I often asked Ihor Soroka and Vyacheslav Shmatukha, EERC System Administrators, to help me understand technical details of data networks, and I am very grateful to them for their assistance. When the thesis was almost ready, Roy Gardner, the University of Indiana, and Janusz Szyrmer, HIID, gave me a lot of good advice on how to make it look and sound better, which deserves a special thankful word. Since the works referes to the Internet in Ukraine, I had a great pleasure to talk with local businessmen who dealt with IT market. Among them are Boris Veligosh, ISP "Access", Volodymyr Storchak, ISP "Infocom", Gerard de Koning, "Lucent Technologies", Vlad Usenko, "Ukrsat", Volodymyr Posdniakov, "Computer Review" ("Komp'iuternoye Obozrenie"), Valery Moiseev, "Business", and a lot of other people from ISP "ElVisti", ISP "Relcom", ISP "Navigator On-line", who unfortunately did not provide their names, but instead helped me better understand the Ukrainian IT market in all its depth.

CHAPTER 1. INTRODUCTION

The information technology (IT) sector is by far the fastest growing market in the world economy. Capitalization of the IT market throughout the world, as well as the number of its participants and end users roughly doubles each year.[1] Computers and supporting services become more abundant and less expensive, and at the same time more user-friendly, which leads to a greater attractiveness and more widespread usage. However, technological improvements in this domain do not automatically imply the efficient use of new technologies. The treatment of congestion on the Internet is one of examples to support this issue.

One of the most used technologies on the Internet is the packet switching networking. When the number of incoming packets is too large for the available network capacity they are queued in a buffer. If the storage capacity of the buffer is insufficient they are permanently dropped out and must be resent. The technical solution to the problem is to increase the available capacity. However, despite the link capacity is enough or idle during some periods of the day, congestion may still occur, because the demands for the bandwidth are generally unpredictable. Thus, even overinvestment in capacity may not be sufficient for the permanent solution of the congestion problem.

Economics as a science of reallocation of scarce resources may provide an economic solution for congested networks. By revealing the true value of a congested link and charging users congestion tolls it is possible to reallocate the load on the circuit more smoothly, thereby decreasing the congestion. Theoretical fundamentals of resolving the problem of congestion are rooted in transportation economics. However, the Internet has particular differences from other public networks, which the Internet Economics is called on to solve.

The work starts with a technical description of the Internet infrastructure

---

[1] "The economics of the Internet," The Economist, October 19, 1996, pp. 23-7.

which is needed to help a reader to comprehend the technical side of the problem. In the next section, I consider the economic environment of the Internet, its growth as an economic system, the issues of network externalities and public goods, and finally, the structure of the Internet industry. This is needed to justify the shape of cost curves involved into the pricing models. I argue that the Internet is a club good when there is no congestion, which is reflected by diminishing average and marginal costs. However, when congestion occurs, the incremental cost of resending packets increases rapidly. Since congestion imposes significant delay costs on other users of the network, it may be treated as the negative network externality; thus the marginal social cost increases even more than the marginal private cost of Internet service providers.

The final section is devoted to pricing models of Internet congestion. First, I present current models employed in the business of providing the Internet services, basically flat rate and usage based schemes. Then I propose a benchmark theoretical model of internalizing congestion, that follows by the extensions to the model which reveal the difficulties of application it in the real life. Next, the literature survey of some models dealing with congestion is presented. All the theoretical approaches may be determined as either dynamic or static models. Finally I argue that although dynamic models are more optimal for alleviating the congestion problem, the static models are more appropriate for implementation in Ukraine. First, there are certain infrastructural restrictions of the Ukrainian economy. Secondly, although dynamic models look theoretically appealing, their implementation is still not elaborated in the modern world.

CHAPTER 2. TECHNICAL INFRASTRUCTURE OF THE INTERNET

The Internet is frequently referred to as "The Network of networks." For our purpose we will present it as a Black Box:: there is some system which connects different computers worldwide into one large network. The service of the Internet is rendered by special businesses – Internet Service Providers (ISP's). They connect end users, different data bases, and whole networks of lower levels in the network hierarchy into one global network. The conceptual scheme of the Internet is presented on the Figure 1.

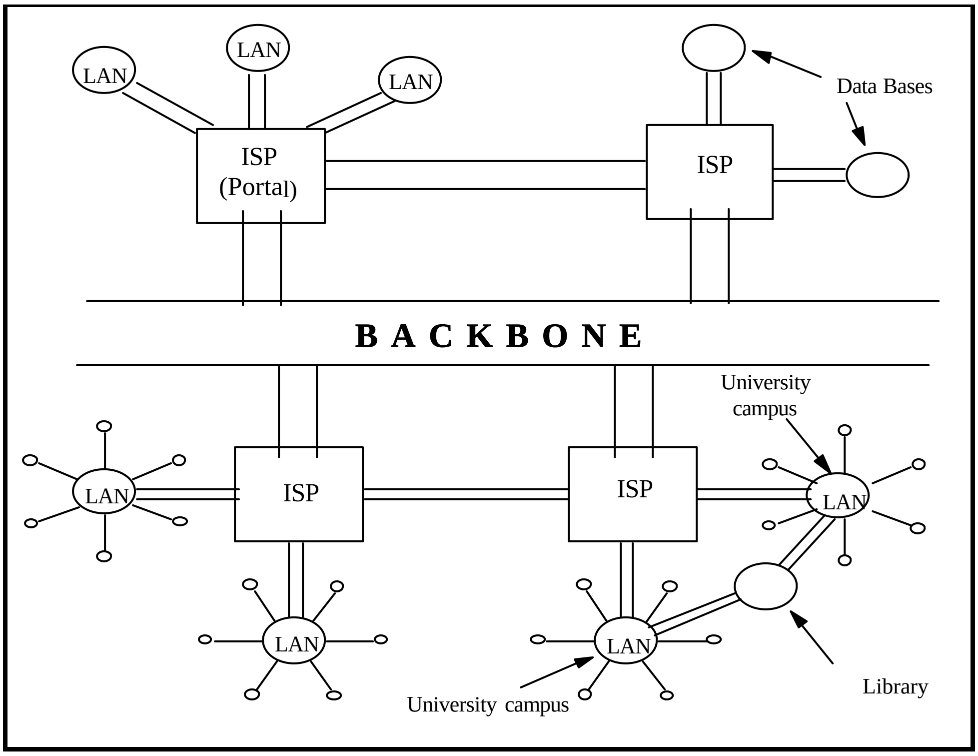

Figure 1. Conceptual model of the Internet

There are two methods of a connection to Internet: via dial-up connection, or through a leased line. The first method uses telephone service to connect to a provider, whereas in the latter case a direct link between a user and a provider is established.

Since data travel through usual cables that are also used for telephone calls, Internet service looks very much like a telephone service. However, the way of data transmission is different. The modern telephone networks are based on establishing a persistent connection between calling parts, reserving line capacity until they hang up. Such connection provides a guaranteed *quality of service* (QoS), but in a very inefficient way, since this requires significant capacity which is idle in some periods of conversation (since the human speech is not so continuous).

On the Internet, no such permanent connection exists. Data sets (jobs) are broken up into packets of a particular length (about 200 bytes, but this parameter may vary greatly), and then are directed to their destination through a series of servers, or routers, using a special transfer protocol. Several major routers constitute a backbone system of significant bandwidth, and serve as an "information superhighway" between smaller networks. Routing computers take the packets on a first-come, first-served basis. When each packet is treated equally, and its routing time depends on the whole available bandwidth, this is called "the best effort service". Since all incoming packets are routed independently, and there is more than one unique way of transmitting data between computers, their natural order may be violated, which requires additional computing at the destination point. The packet-switching technology (as opposed to the circuit-switching of telephone calls) makes heavy use of the routing computers but conserves on line capacity, since several jobs can share a line (Hazlett, 1996, p.3).

At the present time such technology cannot provide a guaranteed quality of service because the number, size and timing of jobs are extremely uncertain.

If a bandwidth of a backbone is considered sufficient for most incoming jobs, different ends of the network having much narrower channels may apply for jobs that in sum far exceed their total capacity at a certain moment. For example, if a big organization, such as a university campus, has a leased line, several users simultaneously downloading a big file (like a class of students who learn how to search for entries in distant databases) may completely exhaust the capacity of the conduit. In such a case, every additional user of this local network (LAN), as well as outside users who want to retrieve data from the LAN's servers will be unable to use it for a while. The situation when a network is temporarily unavailable due to overload is called *congestion*. During congestion periods, significant delays and losses of packets occur, that creates additional work for routers which have to resend data again and again until all packets are sent.

Most congestion occurs at the level of data pipelines, thus the latter become a bottle-neck on the way of information. When congestion happens because of insufficient bandwidth at some end of a network, specialists call this phenomenon "the last mile problem." With the development of new software, such as video conferences or real-time audio, the volume of information to be transmitted grows exponentially, as well as consumers' demands for these services. Thus, congestion may threaten the quality of Internet service.

Unfortunately, investment in new capacities, such as improving protocols of packets transmissions, better algorithms for data compression, increasing bandwidth, and improving quality of channels (through implementing fiber-optical digital lines) cannot alleviate the problem of congestion completely. In most cases, congestion is a temporary event, although its timing during a day is not always predictable. Expanding capacity requires significant financial resources, and it is not always feasible (and needed) to ensure a broad bandwidth all day. Thus, different parts of the network may remain effectively underloaded during some times of the day and be congested at other times (see figure 2). Current technologies of transmitting data[2] do not allow to dynamically adjust the bandwidth for congestion, therefore, congestion may occur even on those datalines where capacity is overinvested.

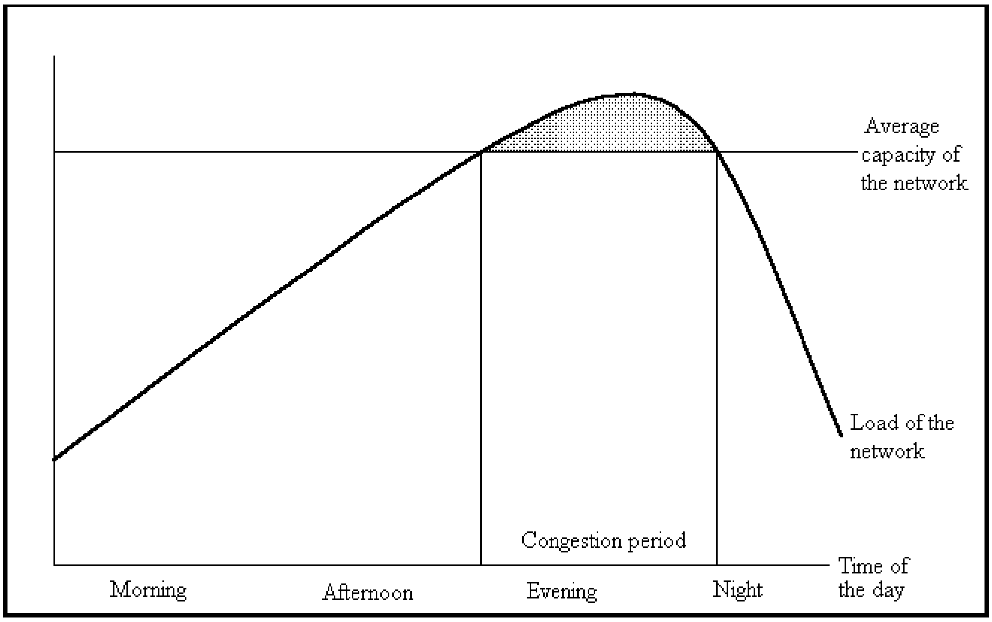

Figure 2. Possible congestion times.

Since it is not entirely possible to solve the congestion problem by investing in data pipelines, it should be addressed at the level of transmission protocols. One way to solve the problem is to assign each packet a priority class which is carried at the head of each packet. When a router receives incoming packets it puts them into different queues on the basis of these priority tags.[3] Packets with higher priorities are given better service (served first). This is not “best effort” service anymore, because the routing time for each packet depends upon its priority, and not only on the available bandwidth. When packets prioritizing is given a status of the internationally accepted standard, the same

---

[2] I do not mention in my research to such sophisticated protocols, as ATM or Frame Relay. Instead, the main working Internet protocol is TCP/IP as the most widespread today.

[3] A new version of one of Internet protocols, IPv6, allows this procedure.

routing algorithms will be used across the whole Network worldwide.

Due to reallocation of the network capacities on the basis of the priority classes the congestion problem may be reduced for some packets. The problem is how to assign priorities. Intuitively, higher priorities should be assigned to packets which require more bandwidth, like multimedia applications. The problem, however, is how to assign the priority classes in the most efficient way. It is not economically viable to set the higher classes just to those applications which technically require more bandwidth (like multimedia), because such reallocation of the resources does not reveal the true economic value of each packet. Choi *et al.* (1997) provides an example when a teenager with idle time can download tetrabytes of entertaining video clips, blocking a cardiac surgeon from receiving vital X-ray data from a distant hospital in time to save a patient. Hence, a certain allocation mechanism is needed. An economic solution to the problem of congestion involves different pricing schemes. Assigning a price to each packet will reveal its true value in a free-market interaction. Therefore, users will decide themselves which priority should be given by "dollar voting." The issues of assigning prices to packets will be considered in next chapters.

CHAPTER 3. ECONOMIC FEATURES OF THE INTERNET

Growth of the Internet

The Internet has developed very quickly over recent years. There are several possible reasons for such rapid growth. Commercialization of the Internet, availability and simplification of computers, the network nature of Internet, broad access to various information, as well as to scientific and entertainment resources should be mentioned as most important.

At the very beginning, public Internet was developed as a means of information sharing between scientific centers and universities in the Untied States. At that time, Internet services were provided by the National Scientific Foundation, which supported the NSFNET backbone. The funding of services was provided by U.S. government, and more or less free access was guaranteed to all users of that profile. However, private users were also allowed to use this service. Later, when the number of private users of the Internet increased dramatically, commercial firms started to consider the Internet as another advertising tool and also came online. In 1995, the NSFNET backbone was officially closed, and since that time, the services began to be provided on the private basis. Commercial providers began to emerge, meeting increasing demand.

In the mid-nineties, costs in computer production started to fall exponentially, which along with increasing competition between major producers has led to rapid decline in computers prices[4]. On the other hand, the capacity of computers increased, software was improving and becoming more user-friendly, that attracted households into computer market. As a result, the SOHO market (Small Office, Home Office) was growing, and the novel category of home computers has emerged. It is possible to assert that the

---

[4] Actually, as Prof. Gardner argues, this happened yet in fifties. However, the dramatic fall in costs and prices has occurred only during the recent decade.

growth rates of the computer market (both SOHO and corporate) and the Internet market are highly correlated, since at the same time, the Internet enjoyed very fast growth (a number of Internet users roughly doubles each year).

Such high demand for Internet services was immediately met by supply. A very large range of firms rendering information services on the Internet also emerged. Search engines, on-line news, libraries, archives of different information became widespread. A lot of companies trying to attract various types of consumers started to build so-called portals on the basis on their corporate sites. Portals propose the whole array of services "just on one click", making their services more wide-spread. Besides information services, improvement of encryption algorithms and highly developed system of banking payments allowed the use of the Internet as a shopping tool. After the U.S. Government granted the Internet commerce a tax-free status, an increasing number of firms began to propose a lot of their supply on-line,[5] and shopping through Internet has become more attractive both for households and corporate users. Capitalization of firms involved into Internet commerce roughly doubled each year, which was an important factor contributing to Internet growth.

Externalities

Besides institutional reasons contributing to Internet growth, it is necessary to mention a network factor. By its very nature, the Internet is a network good. Thus it possesses all the features of such kind of goods, and particularly, significant externality effects. In general, "networks exhibit consumption and production externalities" (Economides, 1995).

A *positive* consumption externality signifies the fact that the value of a unit of the good increases with the number of units sold. From the economic point

---

[5] Among the most striking are examples of rapid growth of Dell Computer and Amazon.com.

of view, to escape ambiguity about the slope of the demand curve (which looks upward sloping according to that definition), it is possible to restate, that "the value of a unit of the good increases with the expected number of units to be sold." Thus, demand slopes downward but shifts upward with increases in the number of units expected to be sold (Economides, 1995). In our case, increasing number of computers in use, as well as increasing number of connections to the Internet, creates additional stimulus for others to connect as well. The underlying logic is as follows: the more computers are united into one network, the more use of the whole system in terms of information it is possible to retrieve.

On the other hand, the negative network externally is also present. The more users are engaged into consumption of a specific network good (that is, the Internet), the more load they impose on the network capacity, thereby diminishing the speed of the transmission and thus decreasing the utility of other consumers. Theoretically, some point may be achieved after which the whole capacity of a particular part of a network is used up (since the network is a heterogeneous in terms of the bandwidth), and congestion occurs. Thus, congestion is the example of the negative network externality.

The Internet as a Club Good.

The aim if this sub-section is to introduce the notion of club goods, and to demonstrate that the Internet shares the features of club goods. This discussion is needed to justify the form of the cost function for Internet services. However, the questions of optimal provision of club goods will not be addressed.

First, it is useful to remind the definition of public goods, and examine if the Internet has attributes of such goods. According to the definition, the *public good* is one that possesses two features: it is nonrival and nonexclusive. Goods that are *nonrival* can be made available to everyone without affection any individual's opportunity for consuming them (Pindyk and Rubinfeld, 1998, p. 673). Nonrivalty means zero marginal cost at any level of production. This is exactly the feature of information good. It is possible to assert that when there is no congestion, and once a network is built, software is installed and connection is established, a cost of providing Internet services to an incremental consumer is virtually zero. A good is *nonexclusive* if people cannot be excluded from consuming it. As a consequence, it is difficult or impossible to charge people for using nonexclusive goods. Both public television and information services are examples of nonexclusive goods, since it is impossible for producer of information to restrict an access to it once it is already on the air. Thus, one can conclude that Internet is public good, according to the given definition.

However, it is very important to distinguish between the information good and the Internet service as a good, since the first is a public good, and the latter is not. In my work I will not consider the information content as a good, but rather Internet service as a good, which is in turn a tool to consume information "goods." In this definition, the Internet does not always posses the features of non-excludability and non-rivalry. It is easy for an ISP to technically prevent any given consumer from obtaining the service. The same is true for Telecom company which can easily prevent a provider from a connection with a backbone. And finally, an Internet Service Provider may effectively charge any consumer for using the service. In this respect, Internet service is excludable. The feature of non-rivalry is more complicated, although. Unless there is congestion, all the available bandwidth capacity is open for any incoming packets, and incremental packets can share this capacity without reducing the transmission characteristics for other packets. After some point, when congestion occurs, new incoming packets began to compete for the capacity, whereas the marginal cost of sending and re-sending each packet increases sharply. Hence, the Internet service is non-rivalry only until the congestion point.

Goods which are excludable but non-rival below capacity limits are called *club goods*. The theory of club goods was developed by C. Buchanan who established the conditions for their optimum output and membership. Classical examples of club goods are swimming pools or social clubs. The Internet is becoming the new example of a club good.

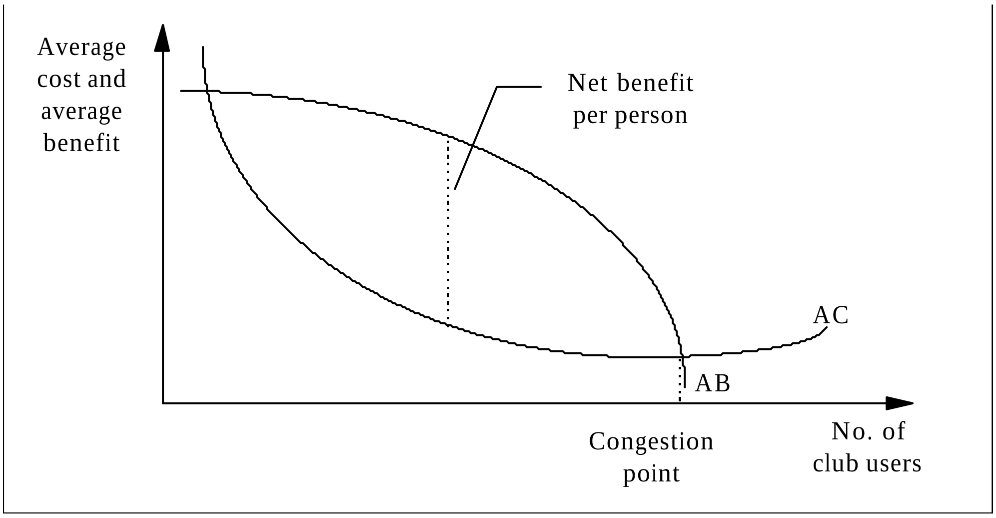

Figure 3. Internet as a club good.

As shown in Cullis and Jones (1998, p.57), the average cost per member of the club (a user of the Internet account) falls. The average cost of producing a given quantity of the good will fall as more people join the club and share the costs (see Figure 3). The benefit per person varies as the number of club members increases. Initially it may rise (as a result of the positive consumption externality, for example), but after a particular number have joined the club (get connected to the Internet), congestion may be experienced and benefit per person will fall. At the same time, both the marginal and the average costs will rise due to the need to resend the packets lost as a result of the congestion. Therefore, cost functions of the Internet service are U-shaped (see 3).

Internet Industry Structure

As already mentioned, the distinctive feature of the Internet is its cost structure as a club good. The largest share in costs of ISP's is constituted by sunk (irrecoverable) costs and fixed (independent on sales volume) costs. They include most equipment expenses and payments for leased telecommunications channels. The incremental cost which is represented by rendering a service to a marginal consumer, in this case is low, or even zero. Such cost structure assumes a declining average cost curve. The latter is true, however, only for non-congested periods, because congestion leads to a rapid growth of both marginal and average costs (see Figure 4.).

Such a cost structure is not compatible with a competitive market. Competition implies a standard economic recommendation of "set a price equal to the marginal cost." However, due to extremely low incremental costs, "such pricing does not recover high fixed costs, and thus is not economically viable" (Varian, 1996). Moreover, the competitive equilibrium price would attract the largest quantity of consumers, in accordance with the law of downward-sloping demand. Intuitively, such a large number of end-users may lead to a significant congestion problem. With an assumption of heterogeneous consumers' preferences it is possible to show (Clark, 1995)

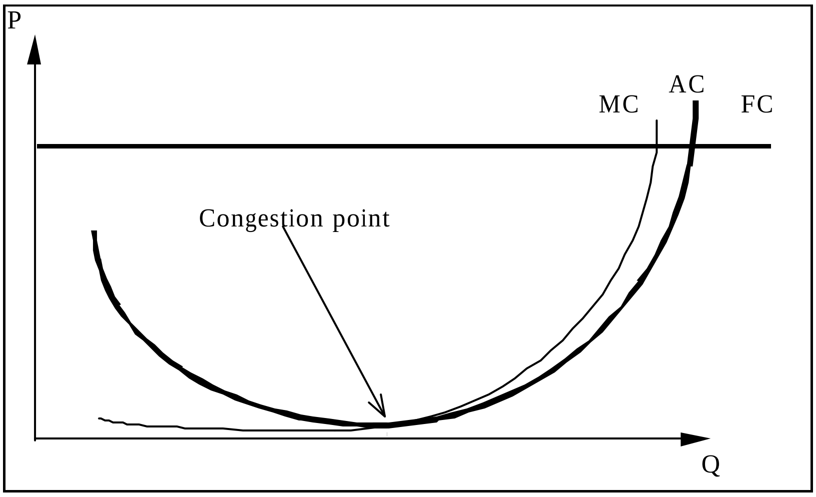

Figure 4. Cost structure of the Internet service

that delays and packets losses may cause either minor dissatisfaction (for example, when sending the electronic mail), or a huge loss of utility, like in a case of multimedia real-time applications. Thus, the competitive model of pricing Internet services is not socially optimal, because it leads to the loss of utility.

In practice the market is far less competitive. Although there are many ISP's nationwide, high charges for long-distance calls prevent users from using services of providers physically located in other areas (there are some minor exemptions however). Similarly, even if some users are ready to pay for a leased line, but local Telecom company is not able to provide free channels for lease, they face only two opportunities: either to finance the construction of an additional line (which is prohibitively expensive for small users), or to use the dial-up method. Thus, intransparency of the market due to absence of the efficient telecommunications infrastructure tends to reduce its competitiveness.

Besides that, the services of the Internet are not homogeneous: they differ in quality of connection, timing and pricing schemes, the range of additional services rendered by an ISP, etc. Services of different providers are substitutable, but not perfectly, that allows to assume product differentiation and segmentation of the market. Given the residual demand curve, each provider has some market power and can earn non-zero profit. In the long run this profit attracts more producers into the sector. Despite initial sunk and fixed costs are rather big, they are not so high as in automobile or aircraft industry to prevent the entry significantly. In addition, institutional barriers to entry are also absent. Thus, the number of producers may increase, their residual demands would fall, and as a result they would earn zero economic profit. Such conditions describe the case of monopolistic competition. It is plausible to impose this industry structure onto the market of Internet services.

Internet Market in Ukraine.

Although the Internet market in Ukraine passes through the same stages as in other countries, there still are some distinguishing features.

First of all, the technical feature of Ukrainian Internet is the absence of any significant backbone structure. Creation of backbones abroad was financed by governments or big businesses. In Ukraine, the Internet has no support from the government, and the private business is not so well developed to get involved in creation of the national backbone. Most major providers lease channels from foreign providers, either through a satellite or through a leased line, and only some of them have direct connections between themselves. The connection is established using telephone lines, both leased or dial-up. Since only one telephone company, Ukrtelecom, provides a majority of channels, there is a monopoly in this sector of the market, thereby increasing the cost of Internet service. Digital fiber-optical lines exist only between major cities, which prevents most consumers from obtaining a high-quality service even if they are able to pay for it. As a result, the total bandwidth of all Ukrainian Internet channels is comparatively narrow. Consequently, the quality of service is lower, and the probability of congestion is higher.

Another important feature of Ukrainian market is the fact that the size of the market is not so significant as in Europe or USA The main reason for that is the above-mentioned correlation with the computer market. The total number of computers in Ukraine hardly exceeds one million. Consequently, the number of Internet users, according to Internet Marketing on-line magazine, is estimated to be not more than 800 thousand (Internet Marketing, 1999). Although a SOHO sector has also appeared in Ukraine, the total demand for the Internet from this sector is low because of the relatively high price for the service.

Since the banking system is not sufficiently developed, the Internet is rarely used for on-line shopping. Only a few companies are engaged into Internet commerce, and only a few consumers have credit cards. On the other hand, there are no incentives to trade goods or services on-line, because the Ukrainian government does not provide any tax-breaks for such activities. Therefore, the only application of the Internet in Ukraine is as a source of information. Although major network factors are true for Ukraine also, they contribute significantly less than in western countries. As a result, the Internet market in Ukraine is rather thin, but rapidly growing.

CHAPTER 4. PRICING MODELS

Business Proposals for Pricing Internet Services.

Currently, two major pricing schemes exist for the Internet access: flat-rate pricing and the usage-based pricing. Both methods are used in modern business, and although they do not account for congestion, their modifications may become a starting point for congestion-sensitive models.

Under *flat-rate pricing*, end users pay a fee for the initial connection and some fixed fees in certain periods, say, monthly. They do not pay for each bit of information they send or receive. The major advantage of the flat-rate pricing is avoiding administrative costs of tracking, allocating and billing for usage, which may be very high. In the telecommunication market, that is supposed to be highly correlated with the market of Internet services, administrative costs may sum up 50 per cent of the bill (Clark, 1995). Moreover, since the flat rate is fixed and set beforehand, it encourages heavier use of the Internet.

However, flat-rate pricing method has a number of pitfalls. Since such pricing scheme allows unlimited use of the network resource, there is a high probability of congestion due to overuse. If ISP's wish to provide a certain minimum quality of service, they need to invest part of their proceedings into new capacity to meet the growing demand. Since there is no signal about true economic value of the resource, this need may lead to the overinvestment or improper investment. Thus, flat-rate pricing is not the most efficient for Internet access providers.

Firdman (1997) argues that under flat-rate pricing consumers do not benefit either: "Cost-conscious users who use the Internet infrequently feel that they pay too much, whereas quality-conscious customers don't believe they get good enough service for their money and would actually pay more for better service if it were available." He concludes that "the only thing we get from flat-rate pricing is equality, but only at the lowest common denominator level."

On the other side of the spectrum is *usage-based pricing*. It assumes separating fees into the fixed part paid for the initial connection and the variable part that bills for each bit consumed. During congestion periods this models pays out pretty well. However its major drawback is that it imposes usage costs whether the network congested or not. Moreover, as Clark asserts (1995), "there is a worry that usage-based model can lead to the collapse of the whole revenue model." Under this model, large users may be driven away, leaving only small users, who will contribute only small fees. This will require the provider to raise the fees even higher in an attempt to recover the fixed costs, and that will start a downward spiral of fleeing users and rising prices.

Crawford (1997) has shown that when a price is only dependent on usage (or, effectively, on congestion), it is in the interest of a producer to create congestion artificially, narrowing the bandwidth and thus increasing the price and profits. Such a monopolist's decision leads to a loss of the social welfare and thus is not socially optimal. Since the result of congestion is a loss of packets and delays, it may lead to a significant utility loss of consumers.

Both drawbacks of the usage-based pricing described above are valid if an ISP has enough market power to charge a monopoly price. Nevertheless, as I discussed at the previous section, the market of the Internet has the structure of monopolistic competition. In such market each provider of Internet services may take into consideration only his residual demand, but not the whole market demand. Since the services of different providers are close substitutes, the loss of utility may distract users from such a provider, thereby shrinking his residual demand and reducing his revenues.

Hence, there is a need to work out a mechanism which would effectively prevent congestion and at the same time would not diminish producers' profits or consumers' utility. The challenge for a pricing structure, then, is to avoid the problems of usage based fees, while addressing some of the concerns that are not captured in a simple access based flat-rate fees.

Benchmark Model of the Optimal Congestion Pricing.

The economic foundations for optimal congestion pricing go back at least to Pigou (1928) and Vickrey (1969). For the Internet, first models dealing with congestion are dated back to the beginning of nineties. Congestion is treated as a negative network externality, and the models represent different approaches towards internalizing it. Prior to discussing models I would like to propose a simple ideal benchmark case. Other models are just different proposals to implement the benchmark model within a set of given restrictions and assumptions.

As I mentioned at the previous section, the Internet possesses some features of a public good. This implies that such way of internalizing the externality, as the marginal cost pricing is not optimal. First, due to the cost structure of the Internet services the marginal cost is negligible during uncongested time, thereby reducing the price as well as revenues of the service almost to zero. Second, at the time of congestion the marginal cost of transmission of an incremental packet of data rapidly increases. Besides, congestion imposes additional costs upon other members of the society, which exceed the private costs of rendering the service. In this case marginal cost pricing leads to deadweight losses as a result of overprovision of the scarce network resource (see Figure 5).

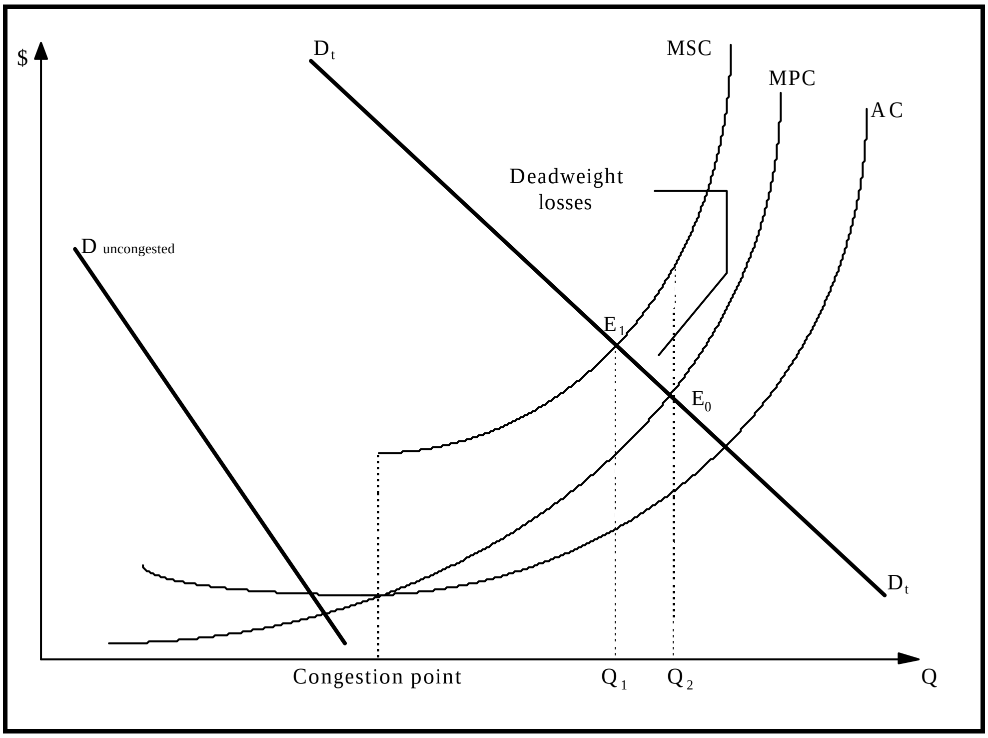

Figure 5. Congestion Pricing

This phenomenon is called "the tragedy of commons." To avoid it, as Choi *et al.* (1997) argues, "prices should exceed the marginal cost of production by the marginal social cost of congestion." On the Figure 5, this corresponds to the equilibrium price at the point E_1 instead of E_0. Thus, congestion will be sufficiently internalized if the price at a time of congestion is set at the level of marginal social cost.

In terms of this model, setting a priority class is equivalent to a guarantee of a certain qulity of the service during congestion time, which may be represented by some pre-determined level of Q_t. A higher priority class will be reflected in a greater Q_t which is guaranteed.

Similar to pricing other congested networks resources, this rather simplistic model requires further extensions to be more applicable for pricing Internet services.

First, the unit of measurement Q in the model is the number of packets transmitted through the available capacity per some period of time t. It is assumed that the bandwidth of a channel is constant in the short run. Thus, the demand function D_t represents the demand for a bandwidth at time t. In this respect, congestion is treated as a sudden increase in demand. Timing, length, and significance of congestion (shift of demand) follow unpredictable patterns. Optimal pricing should be responsive to such changes thus price P_t should be determined for each period of time t. Due to the high speed of data transmission, such periods are very small (measured in nanoseconds). Clearly, in real life it is not very likely that the price will be determined on such short intervals; thus optimality is hardly feasible.

The second extension to the model concerns administrative expenses. If we turn to the example with congestion on roads, the marginal cost of tracing each packet is much lower than tracing and billing each vehicle on a congested highway. According to Choi *et al.* (1997), "toll booths add considerable delay costs, whereas cost of setting up remote sensors in cars and roadway check points are also substantial." However, even negligible administrative costs may be not enough for setting the optimal congestion toll. The problem is to properly identify the owner of a particular packet, or in other words, to find a user who must pay for its transmission. This difficulty may be illustrated on the following example. A data set requested by a student from a public archive electronically looks the same as a spam e-mail message. In the first case, a receiver of the archive should pay for it, whereas in the latter case the sender has to. The problem is worsened by the fact that some packets may be transmitted anonymously, and it is not possible even to identify a server which had sent them into the Internet. Thus, unless more appropriate identification mechanism is developed, optimal pricing will be hard.

The third extension deals with the global character of the Internet. The philosophy of this network society rejects any regulatory body. Thus there is actually no any authority which would set prices or collect congestion tolls. To some extent this role is performed by Internet service providers. However, they can act only locally, controlling an access to a particular network by technical means and charging for it (at most on the national level). In a global sense, there is a lack of a universal unit of account, which would identically measure an economic value of a congested link, or of the congestion toll at another part of the world. This makes it difficult to estimate congestion costs properly, therefore some proxies are needed.

Survey of the Congestion Pricing Models.

Despite these restrictions, a number of congestion pricing models has been developed. Some of them are discussed at this section.

Gupta, Stahl, and Whinston (1997) present a priority pricing approach. When a user requests delivery of a service, he specifies one of several priority classes for the job. The optimal congestion toll depends on the traffic at the site [of possible congestion], the priority class and the social cost of time (delay time) a user imposes on others. Since the user has some expectations about congestion tolls and costs of the delay (for example, based on econometrical estimations), he will request the service within a chosen priority class, if his expected benefits exceed these expected costs. If he values the service less, he waits for a less congested time, thus reallocating the load on the network better over time. Expected time of the delay and the traffic through congested link instantaneously change, as well as user's expectations. Since the user accounts for the delay when making expectations about the price, he pays a marginal social cost, thereby internalizing the congestion externality. Gupta *et al.* (1997) argue that to achieve the optimal allocation of the network resources, it is necessary to set such optimal prices at each possible site of congestion. However, this model does not account for multiple congestion, since in this case the expected social price of the delay will rise with the number of congested sites.

Mackie-Mason and Varian (1995) propose another pricing mechanism that may be called “the smart market approach.” Instead of assigning priority classes explicitly, users submit a bid price for each requested job. After ordering, packets are queued according to the bid price, highest bids being transmitted first. The price paid by every processed job is the bid of the first job not processed during a particular time interval. If all jobs are processed, the price is zero. However, if congestion occurs, the price of transmission will be equal to the last unproceeded bid. The authors argue that such an application of Vickrey second-price bidding scheme encourages users to reveal the true preferences about the value of jobs. Thus users pay the marginal social cost of their Internet use, again internalizing congestion.

Hazlett (1997) proposes to set a priority not on jobs but rather on organizations submitting the jobs. Each organization would have a priority number. Every packet its members send out on the Internet has the organization's priority number in the header. Also, for each job, users choose a subpriority level. Packets are routed by their priority number, with the highest going first. For packets with the same priority number, higher subpriority levels go first. The priority numbers are determined by the dollar value of contributions to an organization's priority, divided by the bandwidth of the organization's line to the network. Anyone in the world may make a contribution to an organization's priority, including the administration of the organization, the individuals within the organization, and even users outside the organization. Contributions are normally made only during congestion periods. Thus, contributors, if behave rationally, will tend to pay for rasing the (sub)priority of jobs no more than the social value of the delay, thereby sufficiently internalizing congestion.

The static priority pricing model proposed by Cocchi, Shenker, Estrin, and Zhang (1993) deals with "maximizing time-averaged user benefits." The pricing is independent of the congestion costs imposed on others and seeks to reallocate network resources in favor of higher-valued jobs in times of congestion. To determine the value of each job, jobs are assigned to specific service classes (quite similar to priority classes discussed above). The authors present a simulation model by which they demonstrate that it is possible to derive optimal priority prices. This significantly increases benefits over single priority pricing.

Bohn, Braun, Claffy, and Wolf (1994) suggest establishing priority classes with non-price incentives. Users are asked to choose the appropriate classification voluntarily. The restrictive mechanism in this scheme is the quota on the weighted sum of packets each organization sends over some period of time. The weights are determined by the selected priority class and increase with the priority. The quota should be somehow divided among the members of the organization, and incentive to choose low priority should be implemented. However, the authors do not present either a mechanism to control the proper use of the quota by each organization, or a principle of dividing the quota among the members of the organization.

Evaluation of theoretical models.

All these suggestions to alleviate congestion fall into one of two classes. The first category consists of approaches by Gupta et al., Mackie-Mason and Varian, and Hazlett. These are dynamic pricing models. The second category presents the static pricing approach and involves models by Cocchi et al. and Bohn at al.

Dynamic models are close to the ideal benchmark case described above (see Figure 5.). Moreover, they tend to internalize congestion, thus presenting the most optimal solution (E_1). Setting the price equal to the marginal social cost dynamically is a very appealing idea. However, due to the restrictions I presented for the benchmark model, the practical implementation of such models is difficult. First, as Choi *et al.* (1997) argues, “in the presence of the negative externality the private market outcome will not be socially optimal, whereas the socially optimal outcome cannot be achieved by private markets.” Since in the current antiregulatory climate of the Internet community it seems unlikely that the government will be allowed to introduce measures to protect the public interest, the “tragedy of the commons” may persist. Second, dynamic pricing models assume that some responsive mechanism is needed for consumers to form their purchase decisions and expectations. This will require a complicated billing scheme that would adjust instantly with congestion. Since the models assume much less complex Internet, accounting costs from usage pricing in the real world may absorb the revenue generated (Mackie-Mason and Varian, 1995).

Proposals of Mackie-Mason and Varian and Hazlett require a mechanism that would allow consumers to submit their bids or contributions at the time of congestion. Such a mechanism may be implemented through a banking system by billing the account of the client. Besides, bidding or making contributions may be done only in some fixed periods of time. This may lead to a situation when a packet from a particular data set which arrived several nanoseconds later than the rest may be put into another bidding auction. It thus may achieve a different priority and be significantly delayed or even dropped. If this packet contains crucial information for the whole data set, it can make the data worthless. To achieve optimality, dynamic bidding process should be implemented in which each packet could communicate with others to coordinate their bids. Such bidding is not possible in complicated networks, and thus is impractical.

Finally, all dynamic pricing models regard the Internet as one economy without barriers and at least with a common currency. Even if simulations of the models may work pretty well, in the real business they may be appropriate at most in large corporative networks (intranets). On the current stage of the development of the Internet, these models can hardly be implemented worldwide.

General drawbacks of dynamic pricing models as well as the absence of a developed banking infrastructure needed to bill users directly at the periods of congestion prevent implementing these models in Ukraine.

I would argue that static models are currently more appropriate for use in real business. The idea behind static models is to make higher-priority requests more expensive regardless of the social cost of congestion. The priority tag may be put into the head of each packet by the local Internet service provider. The price for the priority is determined either in money terms (as in Cocchi *et al.* model) or in terms of the organizational quota.

I propose a simplified model of static priority pricing as represented on the Figure 6.

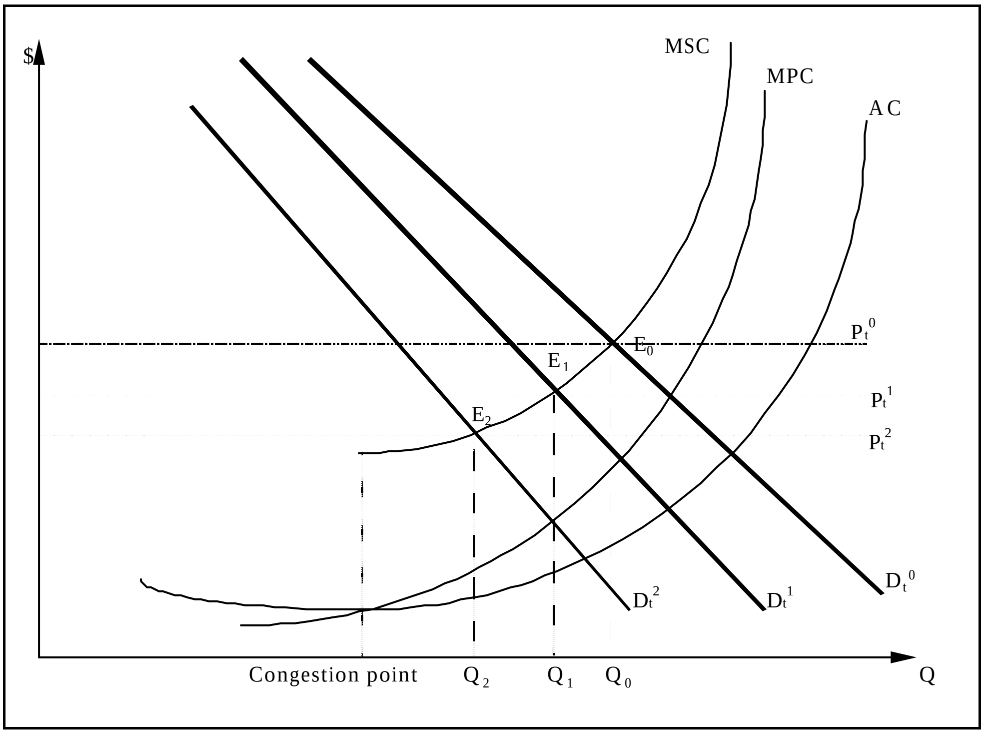

Figure 6. Static Pricing

Each user may purchase a certain package of services which would involve a minimum delay value, or a minimum guaranteed quality of service, or a minimum speed of transmission of packets of a particular size, etc. Providers sell these packets of services to the public. During uncongested period all packets are treated equally on the current best effort basis. In congestion times, packets receive the minimum quality service that is prepaid.

Consider the following example. On the Figure 6, I present three different levels of congestion. For example, the current congestion is on the level D_1. In this case all the packets of types Q_1 and Q_0 will be treated preferentially, Q_0 having the higher priority. Packets of type Q_2 are treated on the best effort basis as usually. It is impossible to raise the priority during congestion, but it is possible to set a service class either for all packets transmitted between an ISP and the user, or for some particular data sets beforehand.

This model may be more appealing for Ukraine, since it provides a rather simple treatment of congestion. It does not require a complicated payment system, as dynamic models do. This model is a modification of the flat rate pricing currently in use. There may be a possible critique that charging for service classes may be not desirable during non-congested period. However, the quality of channels in Ukraine is insufficient most of the time; thus congestion is a real problem. On the other hand, the pitfall of the model is that it is difficult to estimate true congestion cited costs, because static pricing does not reveal true preferences about each congestion point. This may lead to the overprovision trap as discussed earlier.

CONCLUSIONS

The issues raised by Internet Economics are not well developed yet and still wait for their researchers. Economists have already proposed many different solutions to the problem of congestion, one of the most important issues in recent years. However, although congestion in data networks has much in common with congestion on roads, guidelines of the transportation economics usually are not fully applicable for the Internet. The global character of the Network of networks, very high speed of data transmission, lack of any regulatory body and the problem of tragedy of the commons are among the most obvious difficulties for implementation of congestion-sensitive models.

The modern IT business has chosen an extensive way of reducing congestion, i.e. by investing in new technical capacities. However, without pricing that would account for congestion, it becomes more difficult to reveal the true social value of a congested link, which in turn may lead to misallocation of the capital and overinvestment.

The work presents several model dealing with congestion in the Internet. The models may be divided in two main types: dynamic and static ones. Dynamic models are more appropriate for dealing with congestion externality by sufficiently internalizing it either through optimal pricing, or various bidding schemes. All these models, however, consider the Internet as one economy, with a common currency and a relatively small number of congested nodes. On the contrary, in the real life the things are not so well - defined. Even using such models, a user cannot affect congestion in another country (or at least in another currency zone), unless Internet service providers still to the same model. Since implementing such models requires significant administrative and billing costs, and there is no unique authority that would enforce all providers to use the same model, it is unlikely that optimal solution may be achieved.

Instead, businesses are more inclined to use static pricing schemes, such as flat-rate pricing. Probably, this is the solution that is the closest to the optimal on the current stage of development of the Internet, because it is dominated on the market. Some extensions of current pricing models may be appropriate for partial solving the congestion problem (such as introducing the priority classes for each particular user or a user's job). However, again, decentralization of the network and necessity to incur some additional administrative expenses may play a destructive role for such scenarios.

It is necessary to stress that the unique feature of the Internet is its global character. Thus, it is not viable to talk about "American Internet" or "Ukrainian Internet" as separate parts, since they all are the components of the same network, nor it is possible to provide a specific model for each particular country. In the modern world when technologies quickly become available to all countries almost simultaneously, the technical structure of networks is more or less uniform. The only difference is in the capacity in terms of the available bandwidth, quality of telecommunication channels, and national infrastructure. Therefore, a possible solution to the problem may be found in the decentralization issues, that is, in situations, when all ISP's and users behave in their own interests. Such dynamic "general equilibrium" models may provide a sufficient solution to the problem of congestion.

Besides, I would like to mention that currently, researchers do not consider the informational content of the Internet traffic as a binding constraint for their models. This may probably be a result of the fact that an underlying "environment" for the Internet pricing models was circuit-switching network, initially designed for voice traffic. But being unbinding constraint on the voice line, the informational content becomes crucial for digital data lines. For example, servers with important information, such as portals, may experience relatively greater number of requests, thus influencing the congestion level in their local networks. Therefore, the information content should be another important variable when constructing future models dealing with congestion.

WORKS SITED

1. Bohn, R. H. Braun, K. Claffy, and S. Wolff (1994). "*Mitigating the coming Internet crunch: multiple service levels via precedence*." Technical Report, University of California-San Diego, San Diego Supercomputer Center, and NSF
2. Soon-Yong Choi, Dale O. Stahl, and Andrew B. Whinston. (1997) *The Economics of Electronic Commerce*. Macmillan Technical Publishing. Also available on-line: http://merchant.superlibrary.com/products/15787/1578700140/1578700140s.html
3. Clark D. D. (1997) *Internet cost allocation and pricing.* In L. W. McKnight & J. P. Bailey, (Eds.), *Internet Economics* (pp. 215-252). Cambridge, MA: The MIT Press. Also available on-line: http://www.press.umich.edu/jep/works/ClarkModel.html
4. Cocchi, R., S. Shenker, D. Estrin, and L. Zhang. (1993). "*Pricing in computer networks: motivation, formulation, and example*." Technical report, Xerox Corporation.
5. Crawford, D. W. (1997). *Internet services: A market for bandwidth or communication?* In L. W. McKnight & J. P. Bailey, (Eds.), *Internet Economics* (pp. 379-400). Cambridge, MA: The MIT Press.
6. Economides, N. (1996). *The Economics of Networks.* International Journal of Industrial Organization, 16(4), 673-699. Also available on-line: http://raven.stern.nyu.edu/networks/top.html
7. Firdman, E. (1997) *A Rational Business Model for ISP's.* (InFocus, March 31) Available on-line: http://www.internettelephony.com/archive/featurearchive/3_31_97.html
8. Gupta, A., Stahl, D. O., & Whinston, A. B. (1995) *Pricing of services on the Internet* Available: http://cism.bus.utexas.edu/alok/pricing.html
9. Hazlett, D. (1997), *An Interim Economic Solution to Internet Congestion*.. Social

Science Computer Review. Vol. 15:2, pp. 181-189

10. MacKie-Mason, J. K. & Varian, H. R. (1997). *Economic FAQs about the Internet.* In L. W. McKnight & J. P. Bailey, (Eds.), *Internet Economics* (pp. 27-62). Cambridge, MA: The MIT Press.
11. MacKie-Mason, J.K., and H. Varian, (1994) "*Pricing congestible network resources.*" Technical Report, University of Michigan. Also available on-line: ftp://gopher.econ.lsa.imich.edu/pub/Papers/pricingcongestible.ps.Z
12. McKnight, L. W. & Bailey, J. P. (Eds.), (1997a). *Internet Economics*. Cambridge, MA: The MIT Press.
13. McKnight, L. W. & Bailey, J. P. (1997b). *An introduction to Internet economics.* In L. W. McKnight & J. P. Bailey, (Eds.), *Internet Economics* (pp. 3-26). Cambridge, MA: The MIT Press.
14. Pindyk R.S. and Rubinfeld D.L. (1998) *Microeconomics.* Prentice Hall, New Jersey.
15. Varian H.R. (1996). *Differential Pricing and Efficiency.* Available on-line: http://www.firstmonday.dk/issues/issue2/different/

APPENDIX: GLOSSARY OF INTERNET TERMS

Backbone: high capacity channel that is situated on the highest level of the network hierarchy and is being used by all (or most) users of the network. Through a backbone, the major network traffic is transmitted. The speed of transmission is usually mush higher than in any local network.

Bandwidth: capacity of a network link usually is measured in number of megabytes or kilobytes that may be transmitted in a second.

Circuit-switching network: usual voice telephone network on which the persistent connection between calling parties is established through a series of switches (circuits) connected by phsical channels. Such connection is established untill one of the calling parties hangs up.

Congestion: the state of the link when the amount of data to be transmitted exceeds (or is very close to) the maximum banwidth of the link. This situation leads to loss of packets or delays in data transmission.

Internet Service Provider (ISP): a local company that renders a service of connection end-users to the Internet.

Job: a network task performed by a specific router, for example, transmission of packets.

Local Area Network (LAN): a network which consists of computers located not very far from each other and interconnected either directly or through a central server. Usually LAN's may be installed within ornaizations or on campuses.

Packet: For transmission, data set is broken up in separate uinits called packets. A packet is group of bytes that has exactly predetermined structure: a header, a data part, and a trailer. A header brings such important technical information as the address of the destination, or priority of the packet within a particular network.

Packet-switching network: a network on which every link may be shared by multiple users at the same time. Data transmission may be performed through any route (chain of servers) that is available at the momnet. Since different packets from the same data set may be transmitted through various ways, at their destination point, a data processing may be required to restore the original order of a data flow.

Router: Special computer that provides intereconnection between different parts or levels of the network.

SOHO: Acronym which stands for "Small Office, Home Office", and means IT market segment of small businesses and households.

Transfer Protocol: a set of rules which determines interrelation between different parts of a network and serves for transmission of packets and data processing at the destination point.