An Analysis of Incentives for Network Infrastructure Investment Under Different Pricing Strategies

Alok Gupta  
Dept. of Information and Decision Sciences  
3-365 Carlson School of Management  
University of Minnesota  
Minneapolis, MN 55455  
agupta@csom.umn.edu

Boris Jukic  
School of Management  
Enterprise Hall, MS 5F4  
George Mason University  
Fairfax, Virginia 22030-4444  
bjukic@som.gmu.edu

Prof. Dale O. Stahl  
Department of Economics  
University of Texas at Austin  
Austin, TX 78712  
stahl@eco.utexas.edu

Prof. Andrew B. Whinston  
MSIS Dept., CBA 5.202  
The University of Texas at Austin  
Austin, TX 78712-1175  
abw@uts.cc.utexas.edu

(Under Revision, April 2001)

* This research was funded in part by the National Science Foundation, # IRI-9509914, but does not necessarily reflect the views of the NSF. First author's research is supported in part by NSF CAREER grant # IIS-0092780.
An Analysis of Incentives for Network Infrastructure Investment Under Different Pricing Strategies

Abstract

In this paper we present an analysis of investment incentives for network infrastructure owners under two different pricing strategies: congestion based negative externality pricing and the prevalent flat rate pricing. The challenge in a networked environment is to increase capacity of the different components of the network appropriately based on demand for these different components. We develop a theoretical approach to guide the capacity appropriation through the network when the dynamic congestion pricing strategy presented in Gupta et al. (1996) is used. The theoretical model is used to construct a heuristic for capacity appropriation when flat-rate pricing is used. We then use a simulation model to compare the impact of dynamic congestion based pricing with flat rate pricing on the choice of capacity level by the infrastructure provider. We investigate the impact of variety of factors including the per unit cost of capacity of a network resource, average value of the users’ requests, average level of users' tolerance for delay, and the level of exogenous demand for services on the network. Our results indicate that relationships between these factors are crucial in determining which of the two pricing schemes results in a higher level of socially optimal network capacity. The results also indicate that regardless of how these factors are related to each other, the average stream of the net benefits realized under congestion based pricing is always higher than the average net benefits realized under the flat-rate pricing. We also investigate the impact of profit maximization and value flow maximization on the optimal capacity levels.
1. Introduction

“There will be abundant bandwidth, but it's all dark silicon. It's just so far from fruition. What is going to take it to fruition? We need an economic model for that ... Everyone wants it [the Internet] to be free or flat rate or as close to free as they can get. So they won't move to fix the Internet ... What I'm saying is there should be a marketplace, and it should be rational. There should be pricing -- pricing in markets that have choice and competition. It's not that I think only one particular pricing structure will work. I'm saying the current structure is inadequate because it's either free or flat. It does not approximate cost or value or competition anywhere near close enough to be a viable economic model ... I think the marketplace will tolerate a wide variety of pricing mechanisms. My prediction is -- and my advice is -- that we should welcome experiments which attempt to approximate cost and value in the prices ...” -- Bob Metcalfe (The inventor of Ethernet) Interview in IEEE Internet Computing.

The belief that network congestion is not a long-term issue is founded on the notion that a so called "bandwidth bonanza" is bound to happen within the next two to three years in the US because of the availability of a large availability of existing fiber-optic lines\(^1\). However, as the above quote indicates, there is a significant issue of incentives of infrastructure providers to invest in the network capacity to sustain services that require significantly higher bandwidth than today’s applications. While significant increase in backbone bandwidth has taken place in last 5 years we still are not able to provide enough capacity for interactive audio/video services.

\(^1\) See George Gilder’s interview in IEEE Internet Computing, 1997.
While believers in unlimited bandwidth at zero prices contend that pricing mechanisms that manage network congestion are irrelevant (since they deal with a problem that will soon will not exist) and that such pricing policies would only provide a disincentive for capital investment since they discourage usage. Such criticism further claims that a lower level of capital investment, paired with congestion based pricing, will result in a segmentation of user base with a high level of service given to those who can afford it and very poor or no service at all to those with less means. This paper examines the issue of capital investment incentives and capacity expansion via a theoretical model of optimal network capacity investment with congestion based pricing and its simulation implementation. The simulation experiments compare the capacity investment with congestion based pricing to those with flat rate pricing. The results of these experiments show that the ability of a pricing scheme in providing incentives for additional capital investment depends greatly on the relationship between the per unit cost of the network capacity and the average value users have for the services provided on the network. Other factors such as the average users’ cost of time and the level of availability of services\(^2\) have some impact as well. Our results lead to conclusion that the claims of congestion based pricing being not “investment friendly” are not generally correct.

The simulation model investigates capacity expansion incentives under two different overall objectives – socially optimal capacity levels and profit-maximizing capacity levels. We show that the ability of a pricing scheme to motivate higher level of socially optimal investment depends on the nature of the demand for services on the network of interest as well as on certain characteristics of the network itself. We also show that under the externality based social welfare maximizing pricing, enough profits are generated at the optimal capacity level to cover almost entire cost. However, the level of capacity at which profits are maximized falls short of the benefits optimizing capacity level. Therefore, treating the computer network as a profit

\(^2\) As expressed through total number of servers on the network offering a subset of services.
maximizing resource as opposed to a resource designed to deliver the maximum flow of net user benefits may result in a lower level of network capacity.

The rest of this paper is organized as follows. In section 2 we present a brief background on development of pricing methods for computer and network services. In section 3 we present an analytical approach for expanding system-wide capacity of a network form an arbitrary level to socially optimal level with externality based pricing. Section 4 presents the simulation implementation of the theoretical approach in section 3 and outlines all the experiments conducted using the simulation model. Section 5 presents the results from the simulation study outlining implications of different pricing schemes and variety of other factors for network capacity expansion. Finally, we conclude in section 6 with directions of future research.

2. Background

The pricing of computing services has been addressed extensively and with increasing intensity during the last ten to fifteen years, both in computer science literature and in management science literature. Moreover, there exists a large body of literature form the 60's, 70's and 80's covering a variety of approaches to controlling the access to a telecommunication, transportation, manufacturing or computer system. For a good review of earlier models we recommend Stidham (1985). Currently considered approaches vary from the use of dynamic auctions (MacKie-Mason and Varian, 1994) to extensions of the general equilibrium theory. Congestion based pricing was initially proposed by Naor (1969) as a way to optimize the use of one computing resource. Mendelson and Whang (1990) presented an optimal pricing scheme in a closed form for a system in which the individual user possesses private information on delay cost and expected service time and is oriented toward self-interest rather than the overall system optimization. Stahl and Whinston (1994) independently developed a similar theoretical model for distributed computing with additional modification enabling it to handle non-homogenous sized

---

3 See equation 2 for the expression defining net user benefits.
requests. Further extension of that work resulted in an optimal pricing model of a networked computing system by Gupta, Stahl and Whinston (1996). We use and expand that theoretical model to study the network capacity issues. The implementation of that pricing scheme in combination with newly developed demand estimation techniques by Gupta et al. (1997) is incentive compatible, i.e., users are provided no incentive for behavior that may exploit the process of information extraction and price setting\textsuperscript{4}. The solution provided by that pricing scheme is nearly optimal, since optimal prices are computationally approximated and adjusted periodically.

Recently, more attention has been given towards the long-term problem, in which a firm can, besides allocating the usage of the existing capacity, modify the levels of available capacity in order to reach its objectives. Mendelson and Dewan (1990), Stidham (1992) and Dewan (1996) have investigated the issue of capacity expansion for a system consisting of one facility under a variety of pricing and control structures. In this paper we investigate the impact of different pricing policies on computer networks consisting of multiple servers with possibility of changing the existing levels of capacity in unequal amounts. We develop an analytical expression for optimal expansion of network’s capacity for a system using congestion based pricing. This expression uses only the parameters that can be monitored and measured using existing network management technology and, as such, can lead to practical guidelines for network capacity expansion. In the next section we outline the basic approach, based on an analytical model of network pricing, that we implement in the simulation model to evaluate the incentives for expansion in network capacity expansion.

3. Theoretical Background and Approach to Evaluate the Capacity Expansion Problem

In this section we present the theoretical background for network capacity expansion with the objective of maximizing social welfare. We focus on results that are utilized in the simulation

\textsuperscript{4} On the incentive compatibility of our pricing mechanism, see Gupta, Jukic, Stahl and Whinston [1997].
model to implement the capacity expansion. The strategy to discover the optimal capacity given a certain level of demand is based on a combination of: (i) the analytical expressions that provide a vector of directions for capacity expansion for servers that make up a given network; and (ii) intuitively cautious approach of not trying to take too big a step in a given direction. While the idea of a socially motivated monopolist with a goal to maximize system benefits may be considered unrealistic, we use this method to create a benchmark for performance of more market-oriented models.

The objective function in our social welfare model is the sum of the system-wide net benefits realized on a network. The prices derived from using this objective function result in a “stochastic equilibrium,” where the choice of network services by consumers is optimal (i.e., there is no excess demand) given the information about waiting times are true in a stochastic sense.

Our model of the network is characterized by the set M of servers, the set S of all services available on the network, and the set I of users. Let $V_{is}$ denote the instantaneous value to user $i \in I$ of service $s \in S$, and let $\delta_{is}$ denote user $i$’s cost of waiting per unit time for service $s$. Also, let $\tau$ denote the expected throughput time; it consists of service time (which depends on service size $q_s$) and time spent waiting in a server’s queue ($w_m$). Finally let $x_{ism}$ represent the average flow of requests by user $i$ and for service $s$ at server $m$. The social welfare objective function aggregates all of the benefits realized by the users of the network through services obtained, minus the irrecoverable dead-weight losses (waiting costs suffered by the users). These aggregate system-wide benefits do not include the monetary payments from users to the service provider because these payments are merely transfers from one element of the system (user) to another (server).
Therefore, the system wide benefit function is the sum of non-monetary user benefits can be written as:

\[ W(x,K) = \sum_m \sum_i \sum_s [V_{is} - \delta_{is} \tau_m(q_{is}, w_m(x, K))] \ x_{ism} \ s \in S, \ m \in M, \ i \in I \]  \hspace{1cm} (1)

where demand flow vector \( x = \{ x_{ism}: i \in I, s \in S, m \in M \} \), \( K \) is a vector of system capacity \( K \in R^n \) and \( n = \#M \) (\#M represents cardinality of M). Capacity is defined as the ability of the individual server to handle the flow of job requests. It is expressed in the same measurement units as the flow of job requests multiplied by the measurement of the job’s size (number of data size units per time unit: typically bits per second).

Gupta, et al. (1996) prove that prices using this objective function make sure that demand flows are welfare optimizing demand flows, \( x^*(K) \), given the capacity. However, if the level of capacity of network components changes then the allocation and the resulting net demand will change as well. Therefore, Our first goal is to develop an approach that maximizes net benefits with respect to the capacity of the network, i.e., expressions that provide direction for optimal capacity expansion. We used the envelope theorem to find the optimal direction of movement for a set of network components. Proposition 1 below presents the optimal direction of capacity increase.

**Proposition 1:** The optimal direction of capacity increase can be expressed as:

\[ \frac{dW[x^*(K), K]}{dK_m} = (\partial W / \partial K_m)_{x=x^*(K)} \]  \hspace{1cm} (2)

**Proof:**

Note that \( dW/dK_m = \partial W / \partial K_m + \nabla_x W (\partial x / \partial K_m) \ \forall \ m \) from equation 1.

If demand flow \( x \) is optimized with respect to system-wide capacity \( K: x^*(K) \), then \( (\nabla_x W)_{x=x^*} = 0 \) resulting in the expression presented as equation (2). Q.E.D.
The envelope theorem, used for proof of proposition 1, results in the application of the chain rule and the first order condition for unconstrained maximization. The term ‘envelope’ is motivated by the fact that the value function \( B(K) = W[x^*(K),K] \) is given by the upper envelope of \( W[x,K] \).

We now construct the problem of maximizing the benefits with respect to the capacity. The problem can be represented in the following general form:

\[
\max_K B(K) - cK_1; \quad K_1 = (1,\ldots,1) \in \mathbb{R}^n, \quad n=\#M. \tag{3}
\]

where \( B(K) = W(x^*(K),K) \) represents system-wide benefits as a function of a vector \( K \) of system capacity, and \( c \) represents the per-unit cost of capacity at the optimal flow level \( x^*(K) \).

Let the initial vector of system capacity is represented by \( K_0 \). Now suppose that we want to allocate \( b \) units of total capacity among the \( n \) available servers such that \( K \) represents the expanded capacity vector. Then \( K \) can be represented by the following equation:

\[
K = K_0 + b\theta; \quad \theta \in \mathbb{R}^n, \quad \|\theta\| = 1 \tag{4}
\]

where \( \theta \) represents a normalized direction of increase in system capacity. Our maximization problem in equation (3) can then be rewritten as:

\[
\max_{\theta} B(K_0 + b\theta) - cb(\theta 1) \quad \text{s.t.} \quad \|\theta\| = 1 \tag{5}
\]

Proposition 2 provides the optimal value of direction vector \( \theta \).

**Proposition 2:** Given the capacity maximization problem as stated in equation (5), the direction vector of optimal expansion vector can be expressed by the following equation:

\[
\theta^* = (\nabla B - c1)/\|\nabla B - c1\| \tag{6}
\]

\(^5\) It is assumed that unit cost of capacity remains the same regardless of a particular server and its existing capacity level.
Proof: Provided in the Appendix.

We can write the individual components of optimal expansion direction vector as follows:

\[ \theta_m^* = \frac{\partial B/\partial K_m - c}{\|\nabla B - c\|} \quad (7) \]

If capacity is increased in small amounts then to reach the optimal system capacity we can repetitively solve \( \theta^* \) at each new capacity level for as long as the increase in system wide benefits exceeds the cost of the increase in the system’s capacity, or mathematically as long as the following holds true:

\[ \nabla B \theta^* - c \geq 0 \quad (8) \]

We would now derive an expression for equation (6) in terms of observable parameter values. The basic task is of finding an expression for the dependency between the increase in system-wide welfare and the increase in servers capacity \( \nabla B \). Proposition 3 provides the expression for \( \nabla B \) in terms of observable parameters.

**Proposition 3:** Each individual component of \( \nabla B \) in equation (7) can be expressed in the following form:

\[ \frac{\partial B}{\partial K_m} = \left( \sum_i \sum_s \delta_{is} x_{ism} q_s^k / K_m^2 - \sum_i \sum_s \delta_{is} x_{ism} (\partial \Omega_m(X_m, K_m) / \partial K_m) \right) \quad (9) \]

Proof:

Taking the first derivative of the benefit function:

\[ B(K) = \sum_m \sum_i \sum_s [V_{is} - \delta_{is} \tau_m(q_s, w_m(x, K))] x_{ism} \quad s \in S, m \in M, i \in I, \]
with respect to each individual server’s capacity ($K_m$) we obtain the equations expressing the relationship between the increases in the system-wide welfare and the individual server’s capacity:

$$\frac{\partial B}{\partial K_m} = -\sum_i \sum_s \delta_{is} x_{ism}(\partial \tau_m(q_s,w_m)/\partial K_m)$$  \hspace{1cm} (10)

We can represent the throughput, $\tau_m$, as follows:

$$\tau_m = q_s/K_m + w_m$$  \hspace{1cm} (11)

where the first term represents the service time on the server $m$ for a service $s$, while the second term represents the accumulated waiting time on server $m$. In general, waiting time can be expressed as a function of the set of all flow rates ($X_m = \{x_{ism}, i \in I, s \in S\}$) to server $m$ and the capacity of that server. Let $\Omega$ represent a generic function representing the waiting time as follows:

$$w_m = \Omega(X_m,K_m)$$  \hspace{1cm} (12)

Substituting the values from equations (11) and (12) into equation (10), and taking first derivative with respect to $K$ results in the expression presented in the proposition 3.

We can rewrite the expression (9) as follows to further emphasize the fact that its value can be obtained in the terms of performance parameters measured at each server:

$$\frac{\partial B}{\partial K_m} = E_{mc}(\delta) \frac{y_m}{K_m^2} - E_m(\delta) x_m \left(\frac{\partial \Omega_m(X_m,K_m)}{\partial K_m}\right)$$  \hspace{1cm} (13)

where:

$$E_m(\delta) = \frac{\sum_i \sum_s \delta_{is} x_{ism}}{\sum_i \sum_s x_{ism}}$$ is the expected delay cost on server $m$;

$$E_{mc}(\delta) = \frac{\sum_i \sum_s \delta_{is} x_{ism} q_s}{\sum_i \sum_s x_{ism} q_s}$$ is the weighted delay cost on server $m$;

---

6 From this point on we will write $x_{ism}$, assuming that it is clear that these are the welfare maximizing flows $x^*_{ism}$
\[ x_m = \sum_i \sum_s x_{ism} \text{ is the aggregate job flow at server } m; \]

\[ y_m = \sum_i \sum_s x_{ism} q_s \text{ is the aggregate computation cycle flow to server } m, \text{ which is an average demand for that server's capacity per unit time.} \]

The only part of the expression (13) that cannot be obtained through measurement is \((\partial \Omega_m(X_m, K_m)/\partial K_m)\). For the purpose of expanding the capacity of our simulated network, we approximate queues at all the servers with an M/G/1 queuing system. Even though the arrival process to each individual server in real computer networks is not expected to be Markovian, this approximation is appropriate for our high-level simulation model of the computer network as a system of parallel queues. In real-life implementation, capacity expansion model would use a computational approach to such as perturbation analysis to obtain the values of the derivative of waiting time function with respect to the capacity \((\partial w_m/\partial K_m)\) for each \(m\).

By approximating queues at all the servers with an M/G/1 queuing system we can express the individual component of \(\nabla B\) entirely in terms of measurable values as presented in the following proposition.

**Proposition 4.** In an M/G/1 queuing system, each individual component of \(\nabla B\) in equation (7) can be expressed in the following form:

\[
\frac{\partial B}{\partial K_m} = E_m c(\delta) y_m/K_m^2 + E_m(\delta) x_m w^2 \left(4K_m - \sum_i \sum_s x_{ism} q_s\right) / \sum_i \sum_s x_{ism} q_s^2
\]

(14)

**Proof:**

We start with a Pollaczek-Kinchun formula for average waiting time in M/G/1 system (Kleinrock 1976)

\[
W = (2(1-p))^{-1} \times E(q^2)
\]

(15)
where \( x \) is the aggregate flow of submitted requests into the system and \( \rho \) is utilization ratio, which can be re-written in our terms as: \( \sum_i \sum_s x_{ism} q_s / K_m \). The average delay at each individual machine can then be expressed as:

\[
 w_m = \Omega(X_m, K_m) = (2-2\sum_i \sum_s x_{ism} q_s / K_m)^{-1} \sum_i \sum_s x_{ism} q_s^2 / K_m^2
\]  

(16)

Rearranging the terms we get:

\[
 w_m = \Omega(X_m, K_m) = (2K_m^{-2} - 2K_m \sum_i \sum_s x_{ism} q_s)^{-1} \sum_i \sum_s x_{ism} q_s^2
\]  

(17)

Then taking the derivative with respect to \( K_m \) we get:

\[
 \partial \Omega(X_m, K_m) / \partial K_m = - (2K_m^{-2} - 2K_m \sum_i \sum_s x_{ism} q_s)^{-2} (4K_m^{-2} - 2 \sum_i \sum_s x_{ism} q_s) \sum_i \sum_s x_{ism} q_s^2
\]  

(18)

which can be expressed in terms of \( w_m \) as:

\[
 \partial \Omega(X_m, K_m) / \partial K_m = - w_m^2 (4K_m^{-2} - 2 \sum_i \sum_s x_{ism} q_s) / \sum_i \sum_s x_{ism} q_s^2
\]  

(19)

Substituting (19) into (13) we get the equation (14).

In the simulation model, we use the equation (14) to evaluate and expand the network capacities. In order to discover the optimal capacity for a given exogenous load we use the following approach:

(i) move a small fixed length in that direction to a new level of capacity for each component of the network;

(ii) rerun the simulation at this new level of network capacity, until the stochastic equilibrium is reached again with the new capacity levels;

(iii) find the new direction of optimal increase in the capacity, go back to step (i).

This process is repeated until expanding capacities do no yield any additional benefit. We do not compute the optimal step length each time the direction of capacity expansion is calculated. The reasons for not calculating the optimal step-length are twofold:
• First, computation of optimal step length would involve running additional iterations of simulation model imposing higher computational burden.
• Moreover, we can only derive the optimal capacity expansion vector for the pricing model with the objective of maximizing social welfare. None the less, the expression representing this vector provide important insights regarding the dominating factors that influence the direction for capacity expansion using any pricing approach. However, if we used optimal step length with social welfare pricing, the comparisons of net benefits delivered under different pricing schemes at identical levels of capacity at each step would be impossible -- one of the main purposes of this chapter.

In the next section we will describe the conceptual model of the Internet on which this theoretical model is based.

4. Simulation Model and Description of Experiments

In this section we outline the set of simulation experiments designed to determine impact of a pricing scheme on network capacity investment. All the simulation experiments described here are based on the conceptual model of the Internet (Figure 1) developed by Gupta, Stahl and Whinston (1995). This model treats the Internet infrastructure as a "black box", where the total delay is modeled in such a manner that it appears that the delay is only suffered at the server7. The users are connected to the Internet through access providers (which we can consider as a service in itself). The access providers and the service providers (e.g., news, movies, video-conferencing, databases, etc.) are "directly" connected to the Internet through a data-pipeline of a

---

7 The delay within the backbone can be easily modeled as in Gupta, Stahl and Whinston (1995a).
certain capacity. In this model, the capacity of the data-pipeline is essentially the bottleneck for the service providers\(^8\).

This assumption is consistent with the present situation in which the servers have bandwidth limitations, with a majority of the Internet servers connecting to the external world using T1 lines and some larger servers being linked with fractional T3 links at 45 Mbits/sec (Wong, 1997). Several researchers such as Wong (1997) and Tittel (1996) have documented that

![Figure 1: Conceptual Model of the Internet](image)

the most common bottleneck on the Internet is connection saturation at the servers. In the absence of any usage based pricing mechanism, as more users demand services, the quality of the

\(^8\) From the users' perspective, in reality, the bottleneck is either the server's pipeline or the slowest data communication link in their path to the server.
service (in terms of data transfer rates) suffers. Our model assumes that the network service providers are able to monitor the loads at different servers and the access price for a server is a function of the load imposed on that server. When these prices are based on the congestion-pricing model of Gupta, Stahl and Whinston (1995) then at the equilibrium flow of submitted request to each server is optimized.

\[\text{Figure 2: Flow Diagram of the Simulation Model}\]

\[\text{Start Simulation at the initial capacity level } K_0\]

\[\text{Start generating arrivals to the system with exogenous arrival rate } X_0\]

\[-\text{Generate service choice } s, \text{ service value } V_i, \text{ delay } \delta_i \text{ for each request by each user } i.\]

\[-\text{Find lowest total cost server and calculate total cost}\]

\[\text{Is service value } V_i > \text{ total cost } ?\]

\[\text{No}\]

\[\text{User’s request is not submitted}\]

\[\text{Yes}\]

\[\text{Submit and process the request}\]

\[\text{Store predicted waiting times and rental prices on each server}\]

\[\text{T time units since last update}\]

\[\text{Update rental prices } p_s \text{ and expected waiting time estimates } w_s \text{ for all servers } s\]

\[\text{Sufficient number of updates completed } ?\]

\[\text{Yes}\]

\[\text{Collect the performance measures at each server and use them to compute the expansion vector.}\]

\[\text{Increase the capacity of the system by a fixed amount in the optimal direction}\]

\[\text{Desired region of capacity expansion covered } ?\]

\[\text{No}\]

\[\text{Terminate simulation Collect all results}\]

\[\text{Yes}\]

\[\text{Terms of use}\]

\[\text{Note that some users might decide not to get the service because of excessive delays, however, users with negligible delay costs factor } \delta_i \text{ will try to obtain the service regardless of the delays. Thus, with no pricing mechanism, only the users with the lowest value of time can potentially access the services.}\]
Figure 2 provides a flow diagram of the simulation model. The arrival rates to the system are price/cost sensitive. We can interpret $X_0$ as the arrival rate to the system that would occur if there were free access and zero expected waiting times (i.e., the hypothetical non-congested arrival rate or the demand for network services). Note that realized arrival rate into the system, being price and delay-sensitive, is always less than $X_0$.

In the simulation model, a service is characterized by the load it imposes on a server (data pipeline). Upon the arrival of a potential service request the type of service is identified. Then, the current prices and predicted average waiting times are obtained for all the servers that offer the particular service. Prices and expected waiting times are updated after fixed intervals of time. The user values and delay costs are generated from normal distributions such that the mean delay costs are less than 1% of the mean job value. The user then evaluates the total expected cost of this service in terms of her delay cost and the service cost against her value of the service. If the total cost of the service is higher than her value for the service the user quits the system; otherwise, she submits the request to the system.\[10\]

A user's request is sent to the server that was chosen as the least cost server. If the server queue is empty, the request is immediately processed. However, if there are other job requests in the server queue, then the requests are handled in a FIFO manner.\[11\]

In each successive simulation run the capacity of the whole system is increased by a small, fixed, amount.\[12\] In the optimal, congestion-based pricing case, the portion of the increase

---

10 Realistically, this work would be done by a smart agent executing on the users machine.
11 In reality, servers as defined in this chapter execute arriving requests simultaneously, processing units of different service requests in some variation of round-robin scheme. However, the total throughput time is still the sum of the service time and the time spent waiting for the service to be executed. For the jobs arriving early FIFO assumption will result in under-reported waiting times while for jobs arriving later this same assumption will result in over-reported waiting times. However, for the average waiting times, this assumptions will not result in significant bias.
given to each individual server was determined by using a weight factor $\theta_m^*$ as explained in the previous section (using equation 14). In other words, using the expression for the optimal incremental capacity expansion, we find the direction of optimal increase of the capacity of individual network components. The increase in the capacity of individual network components is determined by using the weight factor $\theta_m^*$ so that the total network-wide capacity is increased by a predetermined amount. After increasing the capacity the simulation is run again until the job request flows reach the level that optimizes the sum of net benefits generated on the network for the new level of network capacity. The entire capacity expansion path is constructed by repeating the procedure until no additional benefits are derived from increasing the capacity.

The results of the capacity expansion experiments under congestion-based pricing are compared with results of the capacity expansion experiments under zero pricing. In this set of the experiments, users do not face any monetary charges for submission of their service requests, but they still decide whether to submit a request based on the minimum expected waiting time (and associated cost of waiting). Zero-pricing scheme is a representative of all flat-rate pricing schemes where one can interpret set of zero-pricing experiments as experiments describing the situation where the exogenous demand $X_0$ comes form the population of users that has incurred an up-front fixed fee. For the zero pricing case, $\theta^*$ vector would not necessarily expand the system capacity in the optimal direction. Since we cannot determine the optimal direction of the capacity increase from our observables, we have used a simple heuristic for expanding capacity.

---

12 A small fixed amount does not correspond to the vector of fixed length in n-dimensional space as outlined in the description of the theoretical model. However, we carefully monitored corresponding vector lengths in each step, and they have exhibited sufficiently low fluctuation.

13 This expression is based on observable performance measures at each component of the network and the present levels of components' capacities.

14 The envelope theorem cannot be applied to this case, since the flows are not welfare optimizing under zero pricing. The direction of the optimal increase cannot be expressed solely in terms of the measurable values based on the information we are able to collect on each server.
under the zero pricing policy. We have apportioned a fixed capacity increment to all of the
servers based on the square of the waiting times experienced at each server, recognizing that the
dependency between increases in benefits and increases in capacity on each server is still
approximately proportionate to that measure as is evident from equation 14.

The results presented here are based on a model with 50 servers and 100 services. A
server can provide several of the 100 services, and a service can be provided on up to 25 servers.
In the original simulation model by Gupta et al (1997b) a service "directory" was determined
randomly and fixed throughout the simulation run. The capacity of the data pipelines on the
servers were generated through a random process to be among the following: (i) 128 kbps
(kilobits per second), (ii) 256 kbps, (iii) 384 kbps, (iv) 1.544 Mbps (megabits per second), (v) 4.0
Mbps, and (vi) 10.0 Mbps. The first three choices here represent 2, 4, or 6 multiplexed ISDN
(Integrated Services Digital Network) data channels respectively, the fourth choice is the capacity
of a T1 line, and the fifth and sixth choices are typical of those achieved via Frame-relay or
SMDS (Switched Multi-Megabit Data Services) connections. However, for the purpose of the
capacity expansion experiment, the model was modified so that the initial capacity of the whole
system was set to be very low, at 6.4 Mbps, where each of the 50 servers received an identical
amount of capacity, 0.128 Mbps. The size of each service is also randomly generated to be in the
range of 10 Kb -15 Mb (or, 1.22 kilobytes - 1.8 megabytes). This distribution was chosen so that
there is a higher number of smaller services to simulate a more realistic service request
distribution. The mean size of service is 2.4 Mb. The service directory and the network
configuration, in terms of service sizes and server capacities, were kept constant for all the results
reported here.
5. Simulation Results and Interpretation

We have conducted simulation runs at exogenous arrival rates of $X_0 = 50, 100, 200$ and $500$ requests per second for the system under

(i) the zero pricing policy, and

(ii) the socially optimal pricing policy.

At each capacity level we executed multiple series of simulation runs using differently seeded random values for all four exogenous arrival rates. In our results we report the average value of benefits in the steady state. The variance of the benefits generated was extremely low at each capacity level and at each arrival rate. The average value of the optimal direction of capacity increase was computed and used to apportion capacity increase across the servers. Our results proved to be very robust with both profit maximizing and net benefits maximizing levels of capacity experiencing no fluctuation under all four exogenous job arrival rates. Discussion in this section will be divided into two parts. The first parts will discuss the performance with respect to net-benefits realized in the system and the second part will discuss profits.

5.1. Benefits Generated through Congestion Based Pricing and Zero Pricing as the Network Expands.

As a measure of the network performance, we used the increase in average amount of benefits (per time unit) realized on the network as a result of the new amount of capacity added to the system. Figure 3 shows the results of the simulation runs for the arrival rate of 100 requests per second.

---

15 Word “server” as used in this section refers to the data-pipeline bottleneck, and its capacity is the bandwidth of that bottleneck.

16 Average benefits were measured once network has reached steady state, characterized by very low fluctuation in prices, flow rates and expected waiting times at all servers.

17 The results at other arrival rates are qualitatively similar.
Also included in these graphs are two straight cost lines, reflecting the assumption of constant per-unit capacity cost in our model. The steeper line illustrates the case where the per-unit cost of capacity equals the average value of the users’ request for service. The cost line representing half of the original per-unit capacity cost is also provided for comparison purposes. Solid dark line represents growth of realized system-wide benefit under congestion based pricing as we increase the system wide capacity in optimal fashion. Benefits are growing in a concave fashion reaching a point (a certain amount of capacity) after which additional expansion costs more than the additional amount of benefits it results in. These points are marked as A and A’ for the two per-unit cost lines, respectively. On the other hand, the zero pricing case is characterized by a convex growth function. Notice the impact of the per-unit cost of capacity on the optimal level of system-wide capacity in zero pricing case. Optimal capacity is either zero (point B) when the per-unit cost of capacity equals the average value of the users’ request, or at the level slightly exceeding the entire average exogenous demand2 when the per-unit cost of capacity is twice as low (point B’). This result shows that impact of a pricing scheme on the optimal level of capacity expansion depends on the cost of the network infrastructure and it is summarized in the observation 1 below.

---

18 Based on the relevant data regarding costs of T-lines: The current rental cost of a T1 line is about $1500 per month, which implies that the cost of a 2.45 megabit/second capacity (the average capacity of servers in our simulation) is about $2000 per month, or $0.00077 per second. The average job size in our simulation program was 2.4 Mb, so an average server would handle one job/sec. Our original cost line assumes that the mean value of a job is at least the cost of processing: i.e. $0.00077. In our simulation, the mean cost of delay was set to 0.008 times the mean value of a job (or only $0.022 per hour).

19 100 arrivals/sec * 2.4 Mb = 240 Mbps
Figure 3: Benefits Generated ($/second) under Both Pricing Schemes for the Exogenous Arrival Rate of 100 Requests/Second (capacity is expressed in Mbits/second)

**Observation 1.** Flat rate pricing (represented by zero-pricing) is superior (resulting in a higher level of optimal capacity) to the benchmark congestion-based pricing strategy only when per-unit price of capacity is low compared to user's valuation of services.

Alternatively, congestion based pricing may make possible deployment of systems in an earlier phase of the underlying technology development, characterized by higher price/performance ratio.
Figure 3 also compares the performances of both pricing schemes with respect to the amounts of net benefits delivered, which are the differences between system-wide benefits and system wide capacity costs and are represented by vertical dashed lines at optimal capacity points. The system under optimal pricing performs better, i.e., has a higher sum of net benefits than the system under the zero pricing policy. Better performance under optimal pricing is not only noticeable at the optimal capacity point, but also at every other level of system-wide network capacity where simulation runs were conducted. These findings can be summarized in the following observation.

Observation 2.: Congestion based pricing will always deliver equal or higher system-wide amount of net benefits than the zero pricing at any given system capacity level.

It is interesting to note that the simulation results comparing the performance of congestion based pricing (with congestion fee) with the existing pricing scheme (without congestion fee) at several different capacity and traffic levels have been reported in the context of the airport hub bottleneck model (Daniel, 1995). The average social cost results in that simulation model were significantly higher when congestion fee was not imposed for all five combinations of the system’s capacity and the traffic rate.

Let us also make few comments on the fact that a huge gap between benefits delivered under the two pricing schemes exists for a large range of system wide capacities (between 40 Mbps and 220 Mbps in Figure 3, for example). This gap is a consequence of a convex growth in benefits with respect to capacity under zero pricing, with a very small marginal increase in

---

Also, see other three figures in Appendix 2
benefits, as long as systems capacity is below the average exogenous demand. By analyzing submission rates, queue lengths and average delays in the zero priced network for those lower levels of capacity and comparing them with the same measures from the congestion-priced networks, we noted a large difference in all three of those measures. For a given level of capacity (lower than the average exogenous demand), submission rates and corresponding queue lengths and waiting times experienced by requests were much higher under zero based pricing. Arrival rate of submitted jobs under zero pricing was always only slightly below the current capacity of the system, resulting in very long queues. This can be explained by the fact that our chosen average delay cost factor $\delta$ was very low (see footnote 15), implying high tolerance for even very large delays by majority of users. In order to investigate importance of the average delay cost factor, we conducted a series of experiments with a much higher average delay cost factor (average $\delta$ was increased by a factor of 10). Our results show that the increase in average cost delay has caused change in the shape of expansion curve under zero-based pricing. While still convex, curve is approaching a linear shape whereby the marginal benefits realized through capacity increase at an earlier stage have become bigger. Submission rates and corresponding queue lengths sand average delays dropped significantly, resulting in higher benefits realized. These findings can be summarized in the following observation:

**Observation 3.** At a given demand/capacity ratio the size of performance gap (difference in the level of delivered system benefits) between the two pricing schemes depends on the users' average tolerance for delay.
We would like to note that even if tolerance for delay is very low, quite a large performance gap remains between realized benefits under zero pricing and congestion pricing. This can be explained by the fact that congestion based pricing reflects the idea that efficient pricing of network resources should reflect the cost of externality user's traffic imposes on other network users. In other words, a user should submit a request only if a value of her request exceeds an increased cost of delay imposed to others. Gupta, Stahl, and Whinston (1996) implemented that idea in their pricing scheme used in these experiments which results in higher prices on more congested servers and for larger requests. Overall, the results show that the system implementing the congestion based pricing is much more successful in balancing the load among servers and ensuring that the aggregate capacity of the system is serving the requests with higher value to their users.

Figure 3 and Observation 1 show that the ability of two pricing schemes to encourage a given level of capacity depends on the ratio of the per-unit capacity cost and average value of submitted request. Figure 4 shows the dependence between the optimal level of capacity and the ratio between the per-unit cost of capacity and the average value of users’ request for both pricing schemes for a large range of such ratios. It was derived by gradually changing the per-unit cost of the capacity from a very high value to a low one and observing the optimal capacity level at each cost level.

Under the optimal pricing the optimal capacity levels monotonously increases as per unit cost of capacity decreases relative to the average value of users’ service request. This figure also shows that under flat rate pricing, level of socially optimal capacity is zero as long as per-unit cost of capacity exceeds the average value of users’ service request. When per unit cost of capacity drops below the average value of users’ request, optimal level of system-wide capacity
jumps to the level that can satisfy entire average potential demand, exceeding socially optimal level under congestion-based pricing.

Figure 4: Levels of Optimal Capacities (Mbps) under Both Pricing Schemes as Per Unit Cost of Capacity Drops

The results in figure 4 clearly demonstrate a key point of this chapter. In a network of diverse servers offering a variety of services, the impact of congestion based pricing on the incentives for network capacity investment will depend on the relationship between the per-unit cost and the average value of users’ service request. If that ratio is high, congestion based pricing will enable the existence of the network that simply would not be built with flat-rate pricing. However, if that ratio is sufficiently low, the implementation of zero pricing will result in a higher level of capacity. Finally, as the per-unit cost of capacity approaches levels negligible in comparison to the average value it delivers, the choice of the pricing scheme becomes irrelevant. Under this scenario, the amount of capacity that could be built at a very cost is so large that the
congestion can be eliminated resulting in zero prices. These findings can be summarized in the following observation.

**Observation 4:** Optimal level of system-wide capacity is higher under flat-rate pricing when per-unit cost of the capacity is lower than the average value of user's request. When that price/performance threshold is reached, building of a large system will be the optimal strategy under the zero based pricing. Congestion based pricing will result in an earlier deployment and a more gradual expansion of a system. In other words, until the users start putting high value for a service relative to the cost of its delivery, congestion based pricing is necessary for providing investment incentives. However, once a high valuation is attached to the network usage (as compared to capacity costs) fixed price access will provide sufficient incentives to expand the capacities as required.

A note of caution is in order regarding this observation. One might conclude that price/performance ratio of telecommunication technology is such that it corresponds to the right hand side of figure 4, with flat-rate pricing providing more incentive to expand. This conclusion is valid if we assume that while the per-unit cost of the network capacity is dropping, consumer's valuation of transmissions per unit of time remains the same. That assumption may be challenged from two different perspectives. One is to take into account user's diminishing marginal value of quality and frequency of completed transmissions. Another is to emphasize the fact that users are generating their requests and receiving services through computer applications which are consuming more and more capacity without essential changes in the tasks they are designed to accomplish.

Even if we disregard those factors and assume that current level of per-unit cost of the capacity compared to average value of user's request is low (possibly very low), resulting in
higher level of capacity under flat-rate pricing, performance question still remains. As stated in our observation 2 congestion based pricing will always deliver higher amount of system-wide net benefits at any given system capacity level including of course, the optimal capacity level. Figure 5 shows how the difference between the net benefits delivered changes with the changing ratio of per-unit capacity cost and average request value. It shows the levels of net benefits delivered at the optimal capacity levels under both pricing schemes as we gradually decrease the ratio between the cost of capacity and the average value of users’ request.

Contrasting this figure with Figure 4, we can see that while congestion based pricing is not consistently more investment friendly when compared with zero pricing, it always outperforms the zero-pricing with respect to the amount of net-benefits delivered through the system. As the per-unit cost of capacity decreases relative to the average value of users’ service request that difference becomes smaller. That observation indicates that significant reduction in the cost of network resources renders choice of access pricing scheme less relevant. However, we emphasize again that while the absolute per unit of capacity cost of network equipment and infrastructure is dropping at a fast rate, it is less clear whether such cost relative to the users’ perceived value of services such network delivers is following the same trend. Let us assume that a system exists where it can indeed be proven that average per-unit cost of delivery of services is much lower than the users' average per-unit valuation of service. Then our results seem to indicate that for such a system, the choice of a socially optimal pricing scheme is not really an issue. However, even for such a system the issue of profit maximizing pricing and investment strategies still remains.
5.2. Profits Generated through Congestion Based Pricing as the Network Expands.

One might consider that there exists a discrepancy in our results and the present state of the private and public networks, if one assumes that at present per-unit cost of network-capacity is such that the ratio between it and the average value of user's request is low. According to our results, regardless of the pricing scheme, optimal strategy is to expand networks to the level where there is very little congestion and majority of requests is served with a small amount of
However, there is no proof that many networks, private or public, exist in which there is very little congestion. Nonetheless, we do see a major effort by telecommunication companies to expand the reach of broadband access. Further, we would like to point out that it is not very realistic to assume that maximizing social-net benefits are an objective for a majority of network service providers (excluding private corporate networks designed to serve internal population of user's to the maximum benefit of the company\textsuperscript{21}). Therefore the comparison, presented above, of socially optimal pricing with flat rate pricing does not necessarily capture the real-world environment since the performance metric used is system-wide benefits. From the perspective of a more real-world applicability, we investigated the issue of capacity expansion with profit maximization as the objective. Figure 6 shows the net profits generated from experiments using profit maximization as a goal.

Regardless of the level of network traffic, two observations can be made. First, at the net benefit maximizing level of capacity net profits are close to zero\textsuperscript{22}. This indicates that if a benevolent monopolist were to expand the network to the optimal level, most of the expansion cost could be covered by the profits generated through congestion based pricing. Interestingly, this result also suggests that claim of computational optimality of Gupta et al. (1996) pricing method has merit by being in agreement with the basic findings of welfare economics that socially optimal pricing and production strategies result in zero profits.

\textsuperscript{21} Even in those cases, providers of network and other IT services are often required to behave as profit centers.

\textsuperscript{22} Specifically, for our four arrival rates 97.7%, 80.5%, 88.3% and 92% of capacity cost were covered by profits respectively
Figure 6: Net Benefits and Net Profits ($/second) under Congestion Based Pricing and Original Per Unit Capacity Cost for the Exogenous Arrival Rate of 100 Requests/Second. (capacity is expressed in Mbits/second)

Second, it is apparent that profits reach their maximum at a much lower level of capacity, which points to the fact that a profit-maximizing monopolist using this pricing scheme and capacity expansion policy will under-invest. This simulation result is consistent with the analytical result by Mendelson (1985) where it was shown that the optimal capacity of one computing resource would be lower if it is treated as a profit center as opposed to a net-value maximizing resource. In the case of the multiple computing resources, we cannot fully claim the generality of our simulation results, since the capacity expansion path (i.e., the distribution of capacities at each expansion stage) was the one motivated by the net-benefits maximization.
However, while it may well be that the profit-maximizing capacity distribution at each stage is different from the one obtained in our simulation, our experiments with the variety of expansion heuristics show that the shape of the expansion path does not change much as we change the capacity distribution methods. We summarize our findings in the following observation.

**Observation 5.** A profit maximizing network pricing scheme will result in a choice of system-wide capacity that is below the capacity chosen by socially optimal pricing scheme.

Naturally, there were no profits generated under the zero pricing policy, so we do not provide a comparison between the two policies. However, Stahl, Whinston, and Zhang (1998) show that even the best flat rate pricing policies will generate a lower level of profits compared to congestion based pricing. In future work, we will focus on the impact of profit-maximizing flat-rate pricing strategy on capacity expansion.

6. Conclusions and Recommendations for Future Research

Comparing two pricing policies and studying their impact on level of capital investment in a simulated network economy provides a performance benchmark for the resource distribution on large public networks such as the Internet. As such, the model, presented here, can be used to compare the investment policies under different pricing schemes and investment criteria, such as profit maximization. Through simulation results, we showed that congestion based pricing policy can result in much more robust capacity investment than the investment under flat-rate pricing, if maximization of net benefits is accepted as investment criterion and relative per-unit cost of
capacity is high. This finding counters opinion presently held by many in the field of information technology, which is that implementation of congestion based pricing is detrimental for future investment regardless of the demand characteristics of network users and their valuation of network services.

We derived analytical expression for optimal expansion of network resources using measurable performance parameters. We believe that the rational approach to the management and expansion of a company’s computing resources would be far more beneficial than the presently common bureaucratic approach. Unlike investment in other capital equipment, this approach towards investment in computing resources rarely involves formal and precise cost-benefit analysis. It is based on prompting users to evaluate their need for more resources, and as such is inherently incentive incompatible. Our approach, based on congestion based pricing, ensures that users will reveal their true values for the services they request and provides the decision makers with a much better view of the company-wide need for computational facilities. However, believe that significantly more research has to be performed before guidelines for general applicability can be developed.

We plan to investigate several research issues in the future. First, we would investigate a much richer realm of gradient estimation techniques to compute optimality direction of capacity expansion. We will also test analytical expression under different network specifications, e.g., by varying the number and availability of services and monitoring how changing these specifications affects the assumptions we made in order to obtain the analytical solutions. Finally, this area of research will not be complete until the analytical and simulation based work is expanded to include the investigation of the impact of competitive pricing.
Appendix

A.1 Direction of Optimal Network Capacity Expansion

Here, we will provide the proof of the proposition 2, stating the expression for the normalized vector of the optimal direction of increase of the capacity of our theoretical network. We start with our maximization problem as expressed in equation (4) of the section 3:

\[
\text{max}_\theta B(K_0 + b\theta) - cb(\theta^\perp) \quad \text{s.t.} \quad \|\theta\|=1
\]  

(A1)

Our Lagrangian is:

\[
L(\theta,\lambda) = B(K_0 + b\theta) - cb(\theta^\perp) - \lambda(\theta\theta)^{1/2}
\]

(A2)

Taking first derivative with respect to each \(\theta_m\) \((m=1,..,n)\) and setting it to equal zero, we will get the optimal direction of capacity adjustment:

\[
\frac{\partial L}{\partial \theta_m} = (\frac{\partial B}{\partial K_m})b - bc - \lambda \theta_m = 0
\]

(A3)

\[
\theta_m^* = (\frac{b}{\lambda}) ((\frac{\partial B}{\partial K_m}) - c)
\]

(A4)

The entire optimal direction vector becomes:

\[
\theta^* = (\frac{b}{\lambda}) (\nabla B - c^\perp), \text{ where } \nabla B \equiv (\frac{\partial B}{\partial K_m}, m\in M, m=1,..,n)
\]

(A5)

The value of a scalar \(b/\lambda\) can be retrieved using the condition from equation (A1):

\[
\|\theta^*\| = (\frac{b}{\lambda}) \|\nabla B - c^\perp\| = 1
\]

(A6)

\[
\frac{b}{\lambda} = 1/\|\nabla B - c^\perp\|
\]

(A7)

Therefore our direction vector of optimal expansion can be written as:

\[
\theta^* = (\nabla B - c^\perp)/\|\nabla B - c^\perp\|
\]

(A8)
References


