Information Systems Research

Articles in Advance, pp. 1–21
ISSN 1047-7047 | EISSN 1526-5536

informs®

DOI 10.1287/isre.1090.0253

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

Alok Gupta
Department of IDSC, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455, agupta@csom.umn.edu

Boris Jukic
Operations and Information Systems, Clarkson University, Potsdam, New York 13699, bjukic@clarkson.edu

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

Andrew B. Whinston
Information, Risk, and Operations Management Department, University of Texas at Austin, Austin, Texas 78712, abw@uts.cc.utexas.edu

The Internet is making a significant transition from primarily a network of desktop computers to a network variety of connected information devices such as personal digital assistants and global positioning system-based devices. On the other hand, new paradigms such as overlay networks are defining service-based logical architecture for the network services that make locating content and routing more efficient. Along with Internet2's proposed service-based routing, overlay networks will create a new set of challenges in the provision and management of content over the network. However, a lack of proper infrastructure investment incentive may lead to an environment where network growth may not keep pace with the service requirements. 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. We develop a theoretically motivated gradient-based heuristic to compute maximum capacity that a network provider will be willing to invest in under different pricing schemes. The heuristic appropriates different capacities to different network components based on demand for these components. 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. The simulation model implements the heuristic and ensures that near-optimal level of capacity is allocated to each network component by checking theoretical optimality conditions. We investigate the impact of a 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 simulation results provide a possible explanation for the evolution of the Internet pricing from time-based to flat-rate pricing. The results also indicate that regardless of how these factors are related, the average stream of the net benefits realized under congestion-based pricing tends to be higher than the average net benefits realized under flat-rate pricing. These central results point to the fallacy of the arguments presented by the supporters of net neutrality that do not consider the incentives for private investment in network capacity.

Key words: Internet pricing; infrastructure investment; simulation; investment incentives; net neutrality
History: Published online in *Articles in Advance*.

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.... 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.... 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 (1997).

The belief that network congestion is not a long-term issue is founded on the notion that there is a large surplus of existing fiberoptic lines. However, high-speed access that is becoming increasingly prevalent is generating greater demand for multimedia content, virtual reality, telepresence, multiplayer games, Internet Protocol (IP) voice, video conferencing, and other content-rich applications. Indeed, recent reports indicate that large backbone providers have, because of an increased international demand, "have depleted inventories of unsold circuits on many submarine cables and on some segments of terrestrial networks" (Telegeography 2006). Other reports are more moderate, indicating that thanks to an abundant unlit supply on existing networks, most suppliers can respond to demand increases by lighting wavelengths and fiber pairs on an as-needed basis, predicting that this incremental approach to regulating spare circuit inventories may gradually result in a balanced supply and demand of backbone capacity (Telegeography 2005). On the other end of the opinion spectrum, almost alarmist reports state that "Escalating demand for bandwidth-hungry services, such as HDTV and online gaming, is gradually leading to a critical lack of capacity in cable operators' networks." (ABI Research 2007).

Perhaps the most appropriate way to look at trends in supply and demand for Internet capacity infrastructure (local and backbone) is through the lens of a "circular model wherein a technology infra-structure presupposes a sustainable level of traffic to justify its investment; which depends on sufficient penetration of client-side facilities; which, in turn, depends on hardware and software pricing and availability, attractive tariffs, and traffic generators requiring infrastructure technologies to handle the traffic with reasonable latencies" (Israelsohn 2003). Nevertheless, as the opening quote indicates, little attention has been paid to the investment incentives of the infrastructure providers to sustain services that require a significantly higher amount of bandwidth than is available today.

Believers in unlimited bandwidth at near-zero prices contend that pricing mechanisms, which manage network congestion are irrelevant, because they deal with a problem that will soon cease to exist. Moreover, it is stated that such pricing policies would only provide a disincentive for capital investment because they discourage usage and are contrary to customers' desires for price simplicity (Odlyzko 2000). Critics of pricing-based network resource allocation further claim that a lower level of capital investment, paired with pricing, will result in a segmentation of the user base with a high level of service given to those who can afford the service and a very poor level of service or no service at all to those who cannot afford the service, or at least a certain level of service (e.g., Bailey et al. 1995). Many researchers have refuted the notion of unlimited bandwidth as overly optimistic (e.g., MacKie-Mason et al. 1995).

The issue of free access has been further complicated by a plethora of new services such as voice-over IP and Internet-based multimedia transmissions. The emergence of these services means that individuals having the same connection may use the networks quite differently. Internet2 protocol (see http://www.internet2.edu/) is attempting to include service-based routing to appropriately reserve bandwidth for different services to ensure service quality. However, this would mean that for "low bandwidth" services, such as HTTP, high transmission rates may not be available. On the network design side, the idea of a service-based network design has resulted in the conceptualization of overlay networks. The general idea of an overlay network is to define a network of resources that share a common set of services. An overlay network defines a logical network on top of actual network topography. Two nodes that are next to each other in an overlay network may be topologically far apart. Overlay networks are perhaps best understood in the context of peer-to-peer (P2P) Internet applications, which were popularized through file-sharing applications such as Napster (Napster protocol specification. http://opennap.sourceforge.net/napster.txt), Gnutella (The gnutella protocol specification. http://www.clip2.com), and Freenet (Clarke et al. 1999).

However, most of the current P2P designs are not scalable. For example, in Napster's case, a central server stored the index of all of the available files. A user had to query the central server by using a file name or other search criteria. The central server then obtained the IP number and connected the user to another user machine that had the requested

file. Thus, while the first generation of P2P systems used decentralized storage mechanisms, the process of locating the content was still very centralized. Several research groups have independently proposed a new generation of scalable P2P systems, such as Tapestry (Zhao et al. 2004), Pastry (Rowstron and Druschel 2001), Chord (Stoica et al. 2001), and CAN (Ratnasamy et al. 2001). These systems use a distributed hash table (DHT) functionality, where files are associated with a key (produced, for instance, by hashing the file name), and each node in the system is responsible for storing a certain range of keys. By using such a system, essentially, a new logical service network is generated on top of the actual network topology, which is usually referred to as an overlay network. DHT defines a virtual routing table for the purpose of a specific service, while actual routing uses IP routing at the infrastructure level. In other words, while overlay networks provide a mechanism to enable users to control their routes by relaying through overlay nodes, the route between two overlay nodes is still governed by the underlying routing protocol (see Savage et al. 1999).

While the context of P2P networks helps illustrate the ideas surrounding overlay networks, one of the key applications for the concept of overlay networks is in content distribution networks (CDNs) for distributed resource management and access. Overlay networks can allow companies to locate their content close to users. In other words, data can be made available at the edge of individual networks of a geographically distributed organization. For example, synchronized data can be made available in a distributed fashion instead of invariably loading it from a central server or by using explicit updating schemes that may make the data stale. In addition, the computational resources of an organization as a whole can be tapped to provide businesses with large-scale computer processing capabilities.

As the discussion above indicates, the distributed content management and provision is going to be a significant challenge in the near future. One of the key issues in provisioning content is the capacity management and infrastructure investment that infrastructure providers, such as Internet Service Providers (ISPs) or even CDNs, need to make. In this paper, we examine the issue of capital investment incentives and capacity expansion for an arbitrary computer network that provides a variety of services via a distributed network. We use an economic theory-based simulation approach to examine the issue of optimal network capacity investment with and without pricing. The simulation experiments compare the capacity investment process in a network that manages its usage through congestion-based pricing to the capacity investment process in a network that uses flat-rate pricing. Congestion-based pricing is usage-based pricing that is computed based on the negative externality imposed by a requested service (see Gupta et al. 1996, for the details of this pricing approach). The negative externality can be seen as the monetary value of delay that a given service imposes on the rest of the users in the system. An outcome of congestion-based pricing is that users are charged higher amounts for services when the network components are relatively busy (congested) and lower amounts for services when the network components are relatively less congested.

In this paper, our goal is twofold: to develop a methodology to determine a choice of optimal capacity by an infrastructure provider under different pricing schemes; and to perform a sensitivity analysis on this choice of capacity under varying market conditions, i.e., using different costs of capacity and users' value for services. The computation of the optimal distribution of capacity among different network components is a nontrivial task. We develop a theoretically motivated gradient-based heuristic that allows us to determine the maximum investment in the capacity of a network at a given external load. The heuristic we develop does not provide a guaranteed optimal distribution of capacity among various network components. Therefore we use an iterative approach to redistribute capacities until the theoretical conditions for optimality are met within computational limits. The benefit of our approach is that it can be used with both congestion-based and flat-rate pricing approaches.

We conduct simulation experiments with various external loads and market conditions. The results show that the ability of a pricing scheme to provide investment incentives depends on the relationship between the per unit cost of the network capacity and the average value users place on the services provided by the network. Other factors such as the

average users' cost of time, i.e., the average rate at which the value of a service decays as user is waiting for its completion, and the level of the availability of services[1] have some impact as well. Our results indicate that claims of congestion-based pricing being "investment unfriendly" are generally not correct. In particular, our results provide a rationale for the historical change in pricing approaches from usage-based pricing to flat-rate pricing and provide insights for future pricing strategies.

The simulation model investigates capacity expansion incentives under two different overall objectives: (i) when the infrastructure provider is maximizing social welfare, i.e., the aggregation of revenue and consumer surplus and (ii) when the infrastructure provider is maximizing its profits. We show that the ability of a pricing scheme to motivate higher levels 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 congestion-based pricing, enough profits are generated at a socially optimal capacity level to cover almost the entire investment in capacity. However, the level of capacity at which profits are maximized falls short of the socially optimal capacity level. Therefore, a key takeaway is that treating the computer network as a profit-maximizing resource may result in an overall lower level of network capacity.

The rest of this paper is organized as follows. In §2, we present a brief background on pricing methods for computer and network services. In §3, we present the heuristic and implementation details of our approach for expanding the capacity of a network from an arbitrary level to a socially optimal level with externality-based pricing. Section 4 presents the simulation implementation of the theoretical approach and outlines our experiments. Section 5 presents the results from the simulation study and the implications of different pricing schemes on the network expansion. We conclude with directions for future research.

2. Background

The pricing of computing services has been studied extensively during the last decade by researchers in diverse areas such as computer science, economics, operations research, and management. Moreover, there exists a large body of literature from the 60s, 70s, and 80s, which covers a variety of approaches to controlling access to a telecommunication, transportation, manufacturing, or computer system. For a good review of earlier models, we recommend Stidham (1985). More recently considered approaches vary from the use of dynamic auctions (MacKie-Mason and Varian 1995) to applications of general equilibrium theory (Gupta et al. 1997a). 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 closed form for a system in which the user possesses information on both delay cost and expected service time, and is oriented toward self-interest rather than an overall system optimization. Stahl and Whinston (1994) independently developed a similar theoretical model for distributed computing. Further extension of that work resulted in an optimal pricing model of a networked computing system by Gupta et al. (1996, 1997a). The common feature of these models is that the optimal congestion price for an incoming service request corresponds to the additional cost of waiting imposed on the job requests that are currently in the system. The service request will be submitted only if its utility exceeds this congestion-based price, and, in turn, the loss of utility that it imposes on the system. We use and expand this theoretical model to study network capacity issues.

The implementation of the congestion pricing scheme in combination with the newly developed demand estimation techniques by Gupta et al. (1997c, 2000) is incentive compatible, i.e., users are provided no incentive for behavior that may exploit the process of information extraction and price setting. The solution provided by this pricing scheme is *nearly* optimal (i.e., results in social welfare-maximizing allocation), because optimal prices are computationally approximated and adjusted periodically. In addition, in Gupta et al. (2000), it was shown that users' private delay cost can be estimated, using a nonparametric technique, from observing the users' choice behavior. The technique involves reverse optimizing users' decision model and generating optimal parameter

[1] As expressed through the total number of servers on the network offering a particular subset of services.

ranges. These ranges can then be aggregated using a quasi-Bayesian update algorithm which, is similar to the product limit estimate method described in Kaplan and Meier (1958), and uses some of the properties of the maximum likelihood method similar to Harris et al. (1950). Gupta et al. (2000) further show that such an estimate could be used in pricing without significant loss in efficiency in terms of realized benefits when comparing the deployment of the pricing scheme using estimated delay cost versus the deployment of the same pricing scheme that uses (in real-life unobservable) actual delay cost for each submission.

More recently, attention has been given to the long-term problem, in which a firm can modify the levels of available capacity to reach its objectives. Dewan and Mendelson (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 the possibility of changing the existing levels of capacity in unequal amounts using a gradient-based heuristic. This heuristic computes an expression for the direction of capacity expansion for a network using congestion-based pricing based only on the observable parameters using existing network management technology.

3. Theoretical Background and a Heuristic for Network Capacity Expansion

In this section, we present the theoretical background and develop a gradient-based heuristic for network capacity expansion. Specifically, this work extends the Gupta et al. (1997a), however, unlike Gupta et al. (1997a), we focus on the impact of both pricing schemes on the long-term management of network involving capacity provision. We derive optimal capacity expansion vector for a network at a given exogenous rate to achieve optimal network capacity by employing a theoretically grounded expansion method for congestion-based pricing. The heuristic is developed with the objective of maximizing social welfare, i.e., to maximize the sum of systemwide benefits as reflected by the sum of collected revenues by the infrastructure provider and users' surplus. As Gupta et al. (1997a) show, this social welfare-maximizing allocation can be supported by dynamic prices resulting in a "stochastic equilibrium." Under stochastic equilibrium, users myopically consume the socially optimal level of network services. Specifically, a stochastic equilibrium satisfies three conditions: (i) user service requests are optimal for each user given the prices and anticipated waiting times; (ii) the anticipated waiting times are the correct ex ante expected waiting times given the average flow rates; and (iii) the aggregate average flow rates are equal to the welfare-maximizing rates. While the idea of a socially motivated monopolist with a goal of system benefits maximization may be considered unrealistic, we use this method to create a benchmark for more market-oriented models.

The gradient-based heuristic provides an easy computational approach to increase the capacity of a network's individual component, while determining the optimal level of capacity at a given external load. The general strategy for determining the appropriate allocation of capacity, given a certain level of demand, is based on a two-step process consisting of: (i) using the analytical expressions that provide a vector of directions for capacity expansion for servers that make up a given network and (ii) moving in the direction provided by step (i) in a judicious fashion, keeping the total amount of capacity increment relatively small. Steps (i)–(ii) are repeated until a capacity level is reached where no further gains can be obtained by expanding the capacity. This gradient-based heuristic does not necessarily provide the optimal capacity expansion path. However, because the increments are relatively small and the number of steps is large, the final reallocation approaches the capacity allocation among servers that would be reached via the optimal path as well. We verified this claim numerically by developing a computational approach that reallocates the capacity among the different components of the network such that the resulting allocation of capacities across these components is near optimal at every level during the expansion. We

provide comparison results and a short discussion in the online Appendix A2.[2]

Our model of the network is described as follows:

(i) There are M servers in the network, each server m ($m \in \{1, \ldots, M\}$) offers a subset of a total of S services available on the network.

(ii) Each service s ($s \in \{1, \ldots, S\}$) is available on one or more servers.

(iii) The network serves I myopic users that maximize their net benefits.

(iv) Let V_{is} denote the instantaneous value to user i ($i \in \{1, \ldots, I\}$) of service s, i.e., the value of the service if the service was delivered without delay.

(v) Let δ_{is} denote user i's cost of waiting per unit time for service s.

(vi) Also, let τ_{sm} denote the expected throughput time for service s on server m; it consists of service time (which depends on service size q_s) and time spent waiting in a server's queue (w_m).

(vii) Finally, let x_{ism} represent the average flow (per unit of time) of requests by user i and for service s at server m.

Note that our generic model encompasses the framework of overlay networks and maps the structure of CDNs closely.

The social welfare objective function aggregates all of the benefits realized by the users of the network through services obtained, minus the irrecoverable deadweight losses (waiting costs suffered by the users).[3] It can be written as follows:

$$W(x, K) = \sum_m \sum_i \sum_s [V_{is} - \delta_{is}\tau_{sm}(q_s, w_m(X_m, K_m))]x_{ism},$$
$$1 \le i \le I;\ 1 \le s \le S;\ 1 \le m \le M, \quad (1)$$

where demand flow x is an element of R_+^{MxSxI}: $x = \{x_{ism:}: 1 \le i \le I, 1 \le s \le S, 1 \le m \le M\}$; K is a vector of system capacity: $K \in R_+^M$: $\{K_m: 1 \le m \le M\}$; and X_m represents all flow rates to server m: $X_m = \{x_{ism,} 1 \le i \le 1, 1 \le s \le S\}$.

Capacity is defined as the ability of the individual server to handle the flow of job requests. It is expressed as the number of data size units per time unit (e.g., bits per second). This corresponds to the flow of job requests multiplied by the measurement of the job's size.

Gupta et al. (1996, 1997a) show that congestion-based prices p_{sm} can be computed and regularly updated for each service s offered on server m, resulting in the maximization of the objective function (3). Consequently, our capacity expansion discussion and analysis will be based on the situation in which the pricing scheme makes sure that satisfied demand flows are welfare-optimizing demand flows, denoted $x^*(K)$, given the fixed level of capacity. When the aggregate level of capacity of network components *changes*, then the allocation of the overall network capacity to the individual server components and the resulting satisfied net demand for all the server/service combinations will change as well. Therefore our first goal is to find an approach that addresses the capacity allocation for the network as its aggregate capacity expands.

Note that, from Equation (1), $dW/dK_m = \partial W/\partial K_m + \nabla_x W(\partial x/\partial K_m)$ for all servers m. If demand flow $x(K)$ is optimized with respect to systemwide capacity K, then $(\nabla_x W)_{x=x^*} = 0$, and hence

$$\frac{dW[x^*(K), K]}{dK_m} = \left(\frac{\partial W}{\partial K_m}\right)_{x=x^*(K)}. \quad (2)$$

The envelope theorem, used above, is a result of 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 $B(K) = W[x^*(K), K]$ representing optimized system-wide benefits for any feasible vector K of system capacity, is given by the upper envelope of $W[x, K]$.

Having defined systemwide benefits $B(K)$ as the optimal benefits achieved (through the use of a pricing mechanism that guarantees optimal demand flows) for any level of capacity and allocation of system

[2] Additional information is contained in an online appendix to this papaer that is available on the *Information Systems Research* website (http://isr.pubs.informs.org/ecompanion.html).

[3] As stated by Hassin and Haviv (2003), among many others, "…when a system is considered from a social point of view, assuming social net welfare maximization, payment transferred between system constituents has a zero net effect on social welfare and therefore no effect on the system's optimization." Equivalently, in our model, that is based on the social welfare maximization assumption, the aggregate systemwide 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).

wide capacity K. The problem of finding the level and allocation of capacity $\bar{K}$ that will maximize the systemwide benefits $B(K)$ can be represented as follows:

$$\max_K B(K) - c(K\underline{e}), \tag{3}$$

where $\underline{e} = (1, \ldots, 1) \in R^M$, and c represents the per-unit cost of capacity[4] at the optimal flow level $x^*(K)$. We use a gradient search method, starting at an arbitrary initial vector of capacity K_0, and moving in the direction of the gradient of $[B(K) - c(K\underline{e})]$ in small steps, recomputing the gradient after each step.[5]

We can write the individual components of the normalized gradient vector as follows:

$$\theta_m^* = \frac{(\partial B/\partial K_m - c)}{\|\nabla B - c\underline{e}\|}, \tag{4[6]}$$

as long as

$$\nabla B\theta^* - c\underline{e} > 0. \tag{5}$$

We now derive an expression for Equation (4) in terms of observable parameter values. The basic task is that of finding an expression for the dependency between the increase in systemwide welfare and the increase in servers' capacity ∇B. Proposition 1 provides the expression for ∇B in terms of observable parameters.

Proposition 1. *Each individual component of* ∇B *in Equation* (4) *can be expressed as*

$$\partial B/\partial K_m = \frac{(\sum_i \sum_s \delta_{is} x_{ism} q_s)}{K_m^2} - \sum_i \sum_s \delta_{is} x_{ism} \frac{\partial \Omega_m(X_m K_m)}{K_m}, \tag{6}$$

where the queuing time at server m: w_m *is expressed as a generic function* Ω_m *of the set of all flow rates to server* m: $(X_m = \{x_{ism}, 1 \le i \le 1, 1 \le s \le S\})$ *and the capacity of server* m:

$$w_m = \Omega_m(X_m K_m), \quad 1 \le m \le M. \tag{7}$$

Proof. Taking the first derivative of the benefit function

$$B(K) = \sum_m \sum_i \sum_s [V_{is} - \delta_{is}\tau_{sm}(q_s, w_m(X_m, K_m))] x_{ism}^*, \tag{8}$$

with respect to each individual server's capacity (K_m), we obtain the equations expressing the relationship between the increases in the systemwide welfare and the individual server's capacity. Note that, from this point on, we will use the expression for a demand flow for each service s on each machine m, generated by each user i: x_{ism} without a superscript " *," assuming that it is clear that these are the welfare-maximizing flows x_{ism}^*:

$$\frac{\partial B}{\partial K_m} = -\sum_i \sum_s \delta_{is} x_{ism} \frac{\partial \tau_{sm}(q_s, w_m)}{\partial K_m}. \tag{9}$$

We can represent the total throughput time for service s at server m, τ_{sm}, as follows:

$$t_{sm} = \frac{q_s}{K_m} + w_m, \tag{10}$$

where the first term represents the service time on the server m for a service s, while the second term represents the accumulated queuing time on server m. Substituting the Equations (7) and (10) into Equation (9), and taking first derivative, with respect to K results in the expression presented in the Proposition 1. Q.E.D.

We can rewrite the expression (6) as follows to further emphasize the fact that its value can be obtained in the terms of performance parameters measured at each server:

$$\frac{\partial B}{\partial K_m} = E_{mC}(\delta)\frac{y_m}{K_m^2} - E_m(\delta) x_m \frac{\partial \Omega_m(X_m K_m)}{\partial K_m}, \tag{11}$$

where

$E_m(\delta) = \sum_i \sum_s \delta_{is} x_{ism} / \sum_i \sum_s x_{ism}$ is the "per service request" expected delay cost on server m;

$E_{mc}(\delta) = \sum_i \sum_s \delta_{is} x_{ism} q_s / \sum_i \sum_s x_{ism} q_s$ is the weighted "per capacity/service size unit" delay cost on server m;

$x_m = \sum_i \sum_s x_{ism}$ is the job flow at server m, aggregated across all users i and services s;

[4] It is assumed that unit cost of capacity remains the same throughout the expansion regardless of a particular server and its existing capacity level.

[5] An alternative would be to find the direction that maximizes the incremental net benefit/cost ratio; however, starting at an arbitrary K_0, the first step may require massive reallocations, which would be impractical given adjustment costs. In contrast, following the gradient minimizes adjustment costs. Both methods will find the optimal K^* in well-behaved problems. The results of this comparison are shown in Online Appendix A2.

[6] Further discussion of this expression is in the online Appendix A1.

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

The only part of the expression (11) 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 a real-life implementation, this capacity expansion model could use a computational approach 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 server m.

By approximating queues at all the servers with an $M/G/1$ queuing system, we can express the individual component of ∇B entirely in terms of measurable values as shown Proposition 2.

Proposition 2. *In an* M/G/1 *queuing system, each individual component of* ∇B *in Equation* (4) *can be expressed as follows*:

$$\frac{\partial B}{\partial K_m} = E_{mC}(\delta)\frac{y_m}{K_m^2} + E_m(\delta) x_m w_m^2 \frac{4K_m - \sum_i \sum_s x_{ism} q_s}{\sum_i \sum_s x_{ism} q_s^2}. \quad (12)$$

Proof. We start with a Pollaczek-Khinchin formula for average waiting time in an $M/G/1$ system (Kleinrock 1975)

$$W = (2(1-\rho))^{-1} \lambda E(T^2), \quad (13)$$

where λ is the aggregate flow of submitted requests into the system, $E(T)$ is the expected service time and ρ is the utilization ratio. For each individual machine m, the utilization ratio can be written in our notation as $\rho_m = \sum_i \sum_s x_{ism}(q_s/K_m)$. Similarly, the second component of the Pollaczek-Khinchin equation can be written as $\lambda_m E(T_m^2) = \sum_i \sum_s x_{ism}(\rho_s^2/K_m^2)$. The average delay at each individual machine can then be expressed as

$$w_m = \Omega_m(X_m K_m) = \left(2 - 2\sum_i \sum_s x_{ism}\frac{q_s}{K_m}\right)^{-1} \sum_i \sum_s x_{ism}\frac{q_s^2}{K_m^2}. \quad (14)$$

Rearranging the terms, we get

$$w_m = \Omega_m(X_m, K_m) = \frac{\sum_i \sum_s x_{ism} q_s^2}{2K_m^2 - 2K_m \sum_i \sum_s x_{ism} q_s}. \quad (15)$$

Taking the derivative with respect to K_m, we get

$$\frac{\partial \Omega_m(X_m, K_m)}{\partial K_m} = -\left(4K_m - 2\sum_i \sum_s x_{ism} q_s\right) \cdot \frac{\sum_i \sum_s x_{ism} q_s^2}{(2K_m^2 - 2K_m \sum_i \sum_s x_{ism} q_s)^2}, \quad (16)$$

which can be expressed in terms of w_m as

$$\frac{\partial \Omega_m(X_m, K_m)}{\partial K_m} = -\frac{w_m^2(4K_m - 2\sum_i \sum_s x_{ism} q_s)}{\sum_i \sum_s x_{ism} q_s^2}. \quad (17)$$

Substituting (17) into (11), we get Equation (12). Q.E.D.

In the simulation model, we use Equation (12) to evaluate and expand the network capacities.[7] To analyze the system performance across the feasible systemwide capacity range, we use the following two-step approach:

(i) Move a small fixed length in the direction computed using Equations (4) and (12) to a new level of capacity for each component of the network;

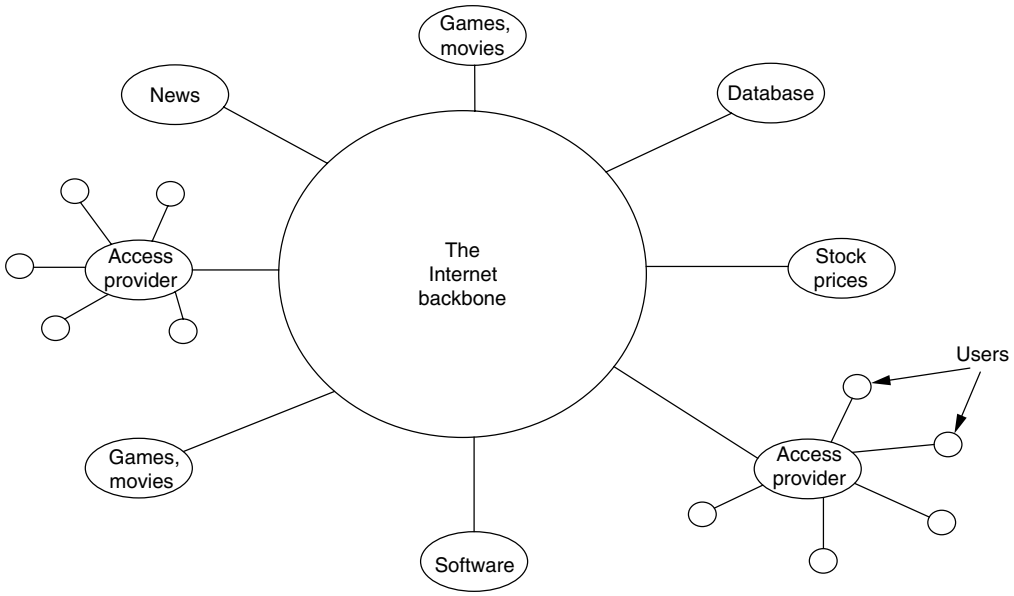
(ii) Rerun the simulation at this new level of network capacity, until the stochastic equilibrium is reached again with the new capacity levels.[8]

This process is repeated until expanding capacities do not yield any additional benefits, i.e., we continue our expansion until we reach a capacity saturation point at which all congestion is eliminated and adding additional level of capacity does not have any impact. Naturally, different objectives, such as social welfare maximization and profit maximization, will lead to

[7] Note that the effect of variance of the waiting time on the values $\partial B/\partial K_m$ is captured through a multiplier w_m^2 in the second term of the equation.

[8] For comparison purposes, we also use a capacity reallocation scheme to see the impact of nonoptimal expansion paths generated by our gradient method. To do so, at each level, we subtract $\varepsilon \underline{e}/n$ from $\partial B/\partial K$, where n is number of servers in the network and $\varepsilon = (\partial B/\partial K)\underline{e}$. This scheme computes the new reallocation at the same aggregate capacity level. We do this until no further benefits are generated by reallocation. This approach ensures that the capacity distribution among the network components is consistent with an implicit budget constraint.

Figure 1 Conceptual Model of the Internet



the different levels of capacity at which the expansion process will be stopped.[9]

In the next section, we will describe the main features of our simulation environment.

4. Simulation Model and Description of Experiments

In this section, we outline the set of simulation experiments designed to determine the 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 et al. (1997b). 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 server.[10] The users are connected to the Internet through access providers, which we can consider as a service in itself. The access and service providers, such as news, movies, and videoconferencing, are "directly" connected to the Internet through a datapipeline of a certain capacity, with this capacity being the bottleneck for the service providers. This assumption is consistent with the situation in which the servers have bandwidth limitations as is the case with CDNs.

4.1. Comparing System Expansion Under Congestion Pricing and Flat-Rate Pricing

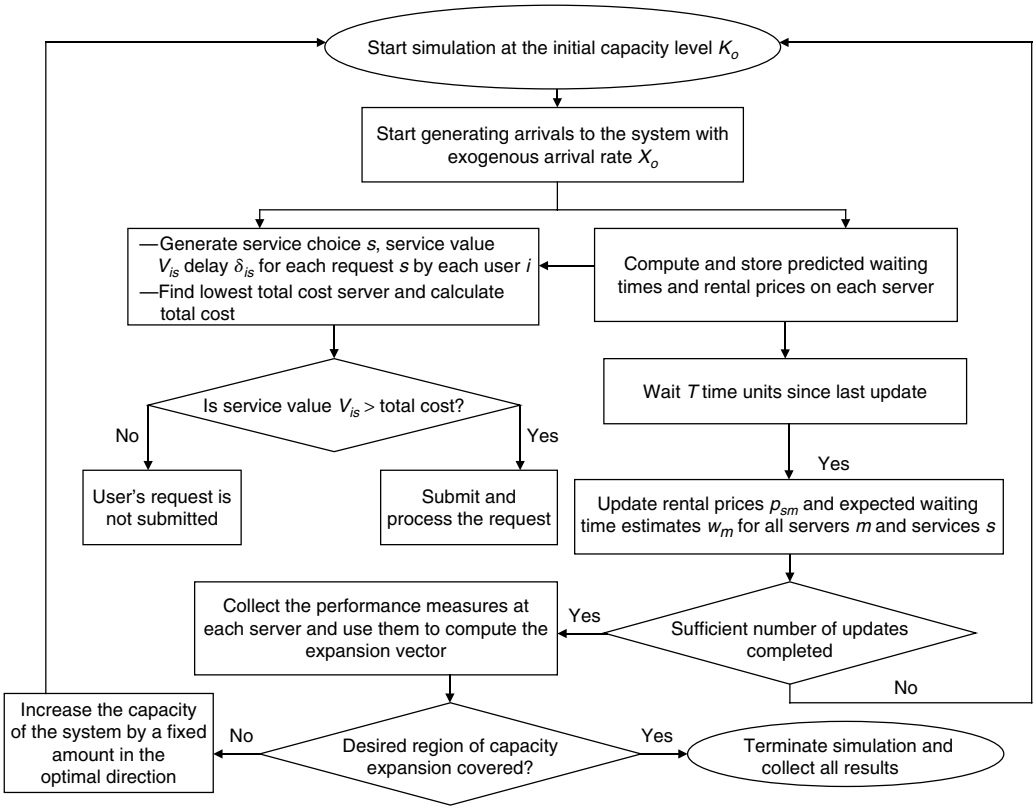
4.1.1. Congestion Pricing. In the absence of any usage-based pricing mechanism, as more users demand services, the quality of the service (in terms of data transfer rates) suffers.[11] Our benchmark model

[9] We do not compute the optimal step length each time the direction of capacity expansion is calculated for the following reasons: First, even if computed, expansion using the optimal step length (varying from one step to another) may not be feasible because the process may be constrained by other external factors such as scheduled budget expenditures, contracts with suppliers, movement of component prices in the market, engineering, and architectural concerns. Second, if we extended our gradient-based approach to include determination of the optimal step length with social welfare pricing and used it in our simulation model, the comparisons of net benefits delivered under different pricing schemes at identical levels of capacity at each step, as described in the next section, would not be possible. Finally, as a matter of practical concern, computation of optimal step length in our simulation model would involve running additional iterations of simulation model, drastically increasing the already impressive computational burden. Nonetheless, we develop an iterative approach that reallocates the capacity, at a given overall network capacity, based on marginal benefits at each component.

[10] The delay within the backbone can be easily modeled as in Gupta et al. (1997a).

[11] Note that some users might decide not to get the service because of excessive delays. However, users with a negligible delay costs

Figure 2 Flow Diagram of the Simulation Model



assumes that the network service providers are able to monitor the loads at different servers and that 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 et al. (1997a), then the equilibrium flow of a submitted request to each server is optimized.

Figure 2 provides a flow diagram of the simulation model. 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 noncongested arrival rate or the demand for network services). The realized arrival rate into the system, being price and delay sensitive, is always less than X_0.

In the simulation model, a service s 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 for all servers m that offer the service s: p_{sm} and the predicted average waiting times w_m are obtained. Prices and expected waiting times are updated each passage of the fixed interval of time. Users' values and delay costs are generated from normal distributions such that the mean delay costs are less than 1% of the mean job value. Users evaluate the aggregate expected cost of a service (i.e., delay cost + the service cost) against their value for that service. If the total cost of a service is higher than the value for that service, the user exits the system; otherwise, the user submits the request to the system.[12]

A user's request is sent to the server that was chosen as the least costly server. If the server queue is

factor δ_i when compared to their service value will try to obtain the service regardless of the delays. Thus, with no pricing mechanism, in the periods of high demand for the system's service, the users with the lowest value of time relative to their service value may crowd out the access to the services for users whose service value diminishes more rapidly with time.

[12] Realistically, this work would be done by a smart agent executing on the user's machine.

empty, the request is immediately processed. However, if there are other job requests in the server queue, then the requests are handled in a first-in-first-out (FIFO) manner.[13]

4.1.2. Flat-Rate Pricing. 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 experiments, users do not face any monetary charges for the individual submission of their service requests, but they still decide whether or not to submit a request based on the expected waiting time, and associated cost of waiting. This scheme is representative of all flat-rate pricing schemes.[14] The flow diagram for the flat-rate pricing case is identical to the one shown in Figure 2. The only difference in the execution of the simulation experiment is that no prices for individual submissions are being computed and stored at servers. Consequently, the total cost for a service includes only the cost of waiting.

4.2. Amount and Direction of Capacity Increments

In the first set of experiments, the capacity of the whole system is increased by a small, fixed amount in each successive simulation run. A small, fixed amount does not correspond to the vector of constant length in a Euclidean n-dimensional space as outlined in the description of the theoretical model, but it does provide for an accurate performance comparison of systems operating under two different pricing schemes at identical levels of systemwide capacities. In the congestion-based pricing case, the portion of the increase given to each individual server was determined by using a modification of the gradient method, which is explained in the previous section. Essentially, we move in the direction of the gradient of B and we adjust the stepsize so the aggregate change in capacity is fixed to $\theta = \nabla B/\|\nabla B\|$.

This method could be described as a modified gradient approach with variable stepsize. After increasing the capacity, the simulation is run again until the job request flows reach the level that maximizes the sum of net benefits generated on the network for the new level of network capacity. We construct the entire capacity expansion path by repeating the procedure until no additional benefits are derived from increasing the capacity. Note that we are purposely setting the unit capacity cost to zero in the expansion vector calculation during this experiment to observe the expansion process across the entire meaningful range of systemwide capacity levels. The issue of capacity cost is not ignored though. It is addressed by including the capacity cost curve in the expansion path graphs as shown in §5, and by performing the additional set of computations using the budget constraints.

As mentioned in §3, the direction of increase at the individual component level is not usually optimal unless a budget constraint is used.[15] To compare how well our simple gradient approach performs with respect to an "optimal" allocation, we develop an iterative approach that reallocates capacities until the marginal benefits at each network component are similar. The comparison of the results using our modified gradient-based approach, with and without reallocation, reveals that at the level of capacity, which has the maximum amount of net benefits realized, the difference in performance between these two methods is minor.

[13] In reality, servers as defined in this paper 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, these assumptions will not result in significant bias.

[14] Note that, in our welfare maximization analysis, the size of the upfront flat fee does not matter, because as pointed out in footnote 3, all payment transfers between system constituents have zero net effect. Assumption of the zero flat fee simply guarantees that exogenous arrival rate will be the same under both mechanisms. One could conceivably imagine a set of experiments where the exogenous arrival rate is lower (implying nonzero flat fee) and individual submission decisions are still based on the trade-off between request valuation and the delay cost. However, those experiments would result in the flow rates that cannot result in systemwide net benefits that are higher than under zero pricing. Therefore, zero pricing represents the best possible benchmark comparison.

[15] A budget constraint approach requires significant additional computational burden. In addition, in a large system with reasonable balanced design, the benefits gained by putting additional constraint may not be significant.

We have also conducted a second set of capacity expansion experiments, this time keeping the increment fixed in a Euclidean sense. The simple gradient method with a fixed Euclidean stepsize is quasi-optimal, provided the adjustment costs are quadratic and dominate budget concerns.[16] While this approach resulted in varying lengths of systemwide capacity increments, the overall expansion curve closely corresponded to the one obtained in the experiments when increments were of a fixed size.[17]

Experiments were also conducted under a flat pricing regime. In this case, we use another simple heuristic for expanding capacity. Essentially, we apportion 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 in our gradient expression, as indicated in Equation 12.

4.3. Simulated Network Elements and Services

The results presented here are based on a model with 50 servers[18] and 100 services. A server can provide several of the 100 services, and a service can be provided on up to 25 servers. For the purpose of the capacity expansion experiment, 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. Even though the servers are homogenous in terms of size, they are not homogenous with respect to the number and size of services offered. A sample of the service directory is provided in Table 1.

To address the impact of more severe differences of initial capacity distribution, another set of experiments was conducted, with the majority of the initial level of capacity apportioned to only one server. These results showed no qualitative difference from the experiment runs with a balanced starting scenario.[19] Therefore the simulation model seems to be robust against the starting bias in reaching the optimal level of capacity. The size of each service is randomly generated to be in the range of 10 Kb–15 Mb. The mean size of a service is 2.4 Mb. The service directory and the network configuration were kept constant for all experiments.

Table 1 A Sample of the Service Directory Used in the Experiments

Server number (total number of services offered)	0 (28)	1 (40)	2 (29)	3 (21)	4 (22)	5 (34)	6 (29)	48 (24)	49 (27)
Service 0 available	0	0	0	0	1	0	1	0	1
Service 1 available	0	0	0	0	0	0	0	0	0
Service 2 available	0	1	1	0	0	1	1	1	0
Service 3 available	0	0	1	0	1	0	0	0	0
Service 4 available	0	0	0	0	0	0	0	1	0
Service 98 available	0	0	1	0	0	0	0	0	0
Service 99 available	0	0	0	0	0	0	0	0	0

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 flat-rate pricing policy and (ii) the congestion-based pricing policy. At each capacity level, we executed multiple runs using differently seeded, random values for all exogenous arrival rates. We report the average value of benefits in the steady state, where benefits represent the aggregate amount of the values of a job currently in the system minus the total current delay costs, as expressed in (3). The benefit value reported at each capacity level is an average value during the last 2,000 update periods. Our results proved to be very robust with optimal levels of capacity experiencing no fluctuation under all four exogenous job arrival rates. The current analysis builds on and extends the findings presented by Gupta et al. (1997a) by comparing the impact of congestion-based pricing and flat-rate pricing in the long-term scenario that involves not only the policies for management of the network under a given level

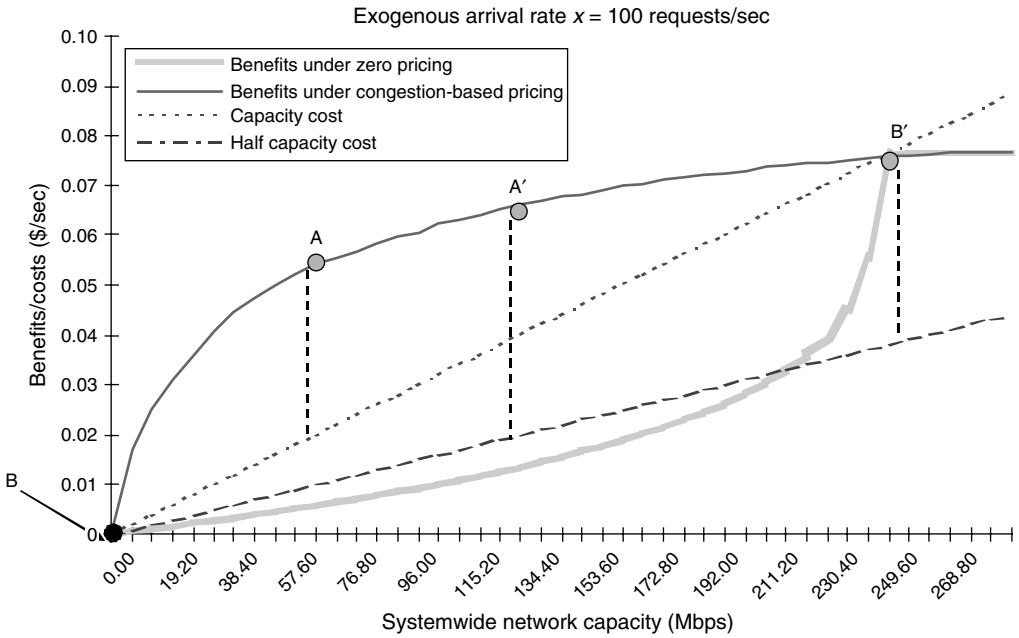
[16] A full optimization would entail finding the optimal stepsize at each level.

[17] For the purpose of review, we provide these results in Online Appendix A5.

[18] The word "server" as used in this section, refers to the data pipeline bottleneck, and its capacity is the bandwidth of that bottleneck.

[19] For the interested reader, we provide these results in online Appendix A6.

Figure 3 Benefits Generated ($/Second) Under Both Pricing Schemes for the Exogenous Arrival Rate of 100 Requests/Second



Note. Capacity is expressed in Megabits/second.

of systems capacity, but also the policies for expansion of the network and the impact of management decisions as it grows to match the demand.

5.1. Benefits Generated Through Congestion-Based and Flat-Rate Pricing as the Network Expands

As a measure of the network performance, we used the average benefits (per time unit) realized as a result of the new amount of capacity added to the system.[20] Figure 3 shows the results of the simulation runs for the arrival rate of 100 requests per second. We conducted this analysis at various exogenous arrival rates (50, 200, and 500 requests/second) to simulate different load conditions. The results are provided in Online Appendix A3. These results indicate that implications of our findings are robust under a wide array of exogenous load conditions. Also included in these graphs are two straight lines reflecting different levels of capacity cost. 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. The thin line represents the growth of realized systemwide benefits under congestion-based pricing as we increase the systemwide capacity. The benefits are growing in a concave fashion, reaching a point (a certain amount of capacity) after which additional expansion costs more than the value of the additional benefits. These points are marked as A and A′ for the two per unit cost lines, respectively. Notice the impact of the per unit cost of capacity on the maximal level[21] of systemwide capacity in the flat-rate pricing case. The optimal capacity is either zero (point B) when the per unit cost of capacity equals the average value of the users' request, or at a level slightly exceeding the entire average exogenous demand[22] when the per unit cost of capacity is twice as low (point B′).

This result shows that the impact of a pricing scheme on the maximal level of capacity expansion

[20] 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.

[21] We designate the maximal level where the new benefits or profits are maximized using either of the expansion approaches—with or without reallocation.

[22] Note that 100 arrivals/sec $*$ 2.4 Mb $=$ 240 Mbps.

depends on the infrastructure cost; this impact is summarized in the following observation

Observation 1. Flat-rate pricing results in a higher level of capacity than the benchmark congestion-based pricing strategy only when the per unit price of capacity is low compared to the user's valuation of services. Congestion-based pricing may make possible the deployment of systems in an earlier phase of the underlying technology development, with a higher price/performance ratio.

Interestingly, Observation 1 provides a rationale for a widely adopted usage-based (time-based) pricing prevalent in the early stages of Internet evolution when the cost of technology was relatively high and the service valuation relatively low. Figure 3 also compares the performances of both pricing schemes with respect to the amounts of net benefits delivered, which are the differences between systemwide benefits and systemwide capacity costs and are represented by vertical, dashed lines at maximal capacity points. The system under congestion-based pricing performs better, i.e., has a higher sum of net benefits than the system under the flat-rate pricing policy. Better performance under congestion-based pricing is also noticeable at every other level of systemwide network capacity, where simulation runs were conducted. This can be summarized in Observation 2.

Observation 2. Congestion-based pricing delivers an equal or higher systemwide amount of net benefits than the flat-rate pricing at any given system capacity level.

Note that a huge gap between benefits delivered under the two pricing schemes exists for a large range of systemwide capacities (between 40 Mbps and 220 Mbps in Figure 3, for example). This gap is a consequence of slow growth in benefits with respect to capacity under flat-rate pricing. In other words, a very small marginal increase in benefits occurs as long as the systems capacity is below the average exogenous demand. We noted a large difference in *submission rates, queue lengths*, and *average delays* in the flat-rate priced network for lower levels of capacity as compared to the congestion-priced network. For a given level of capacity (lower than the average exogenous demand), submission rates, corresponding queue lengths, and waiting times were much higher under flat-rate pricing. To investigate the importance of the average delay cost factor, we conducted a series of experiments with the average delay cost factor increased by a factor of 10. The results of this experiment indicate that, under flat-rate pricing, a larger average delay cost causes changes to the shape of the expansion curve. It approaches a linear shape, with the marginal benefits realized through the capacity increase at earlier stages larger than in the original case. Submission rates, corresponding queue lengths, and average delays drop significantly, resulting in higher benefits. Under congestion-based pricing, however, the shape of the curve does not ehibit large changes.[23] These findings can be summarized as follows.

Observation 3. At a given demand/capacity ratio, the size of the performance gap (the difference in the level of delivered system benefits) between the two pricing schemes depend on the users' average tolerance for delay. The gap tends to decrease as the average tolerance for delay decreases, i.e., when customers become more sensitive to delay.

The important managerial insight is that, as the network performance expectations increase over time, users are more likely to choose alternatives that can provide higher performance even at higher prices. Observation 3 can be explained by the fact that, even without the congestion-based pricing mechanism, a lower delay tolerance tends to eliminate low-valued jobs from the queues, freeing up remaining bandwidth for the higher-valued jobs. Still, we observed that a large portion of the performance gap remains between benefits realized under flat-rate pricing and congestion-based pricing, even if tolerance for delay is very low (i.e., delay cost is very high). This gap can be explained by the fact that congestion-based pricing considers the cost of the externality that a user's traffic imposes on other network users. Therefore, a user implicitly submits a request only if the value of her request exceeds the increased cost of delay imposed on the other users. Overall, the results show that 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 job requests with a higher value to their users.

[23] Benefits realized at every level of systemwide capacity did exhibit a slight decrease when compared to the original lower delay cost case.

Figure 4 Levels of Optimal Capacities (Mbps) Under Both Pricing Schemes as per Unit Cost of Capacity Drops

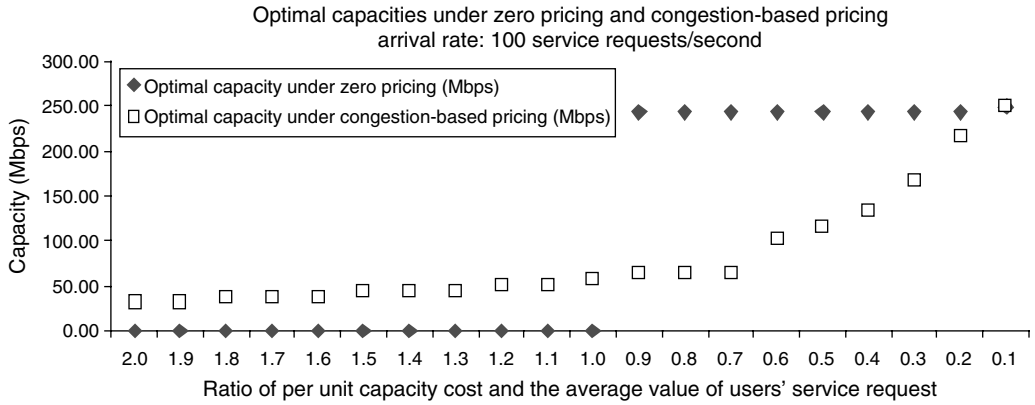


Figure 4 shows the dependence between the maximal level of capacity and the ratio of per unit cost of capacity and the average value of the user's request, for both pricing schemes. Under congestion-based pricing, the maximal capacity levels monotonously increase as the per unit cost of capacity decreases relative to the average value of the users' service requests. Under flat-rate pricing, the level of maximal capacity is zero as long as the per unit cost of capacity exceeds the average value of the users' service requests. When the per unit cost of capacity drops below the average value of the users' requests, the maximal level of systemwide capacity jumps to a level that can satisfy the entire average potential demand, exceeding the maximal capacity levels under congestion-based pricing.

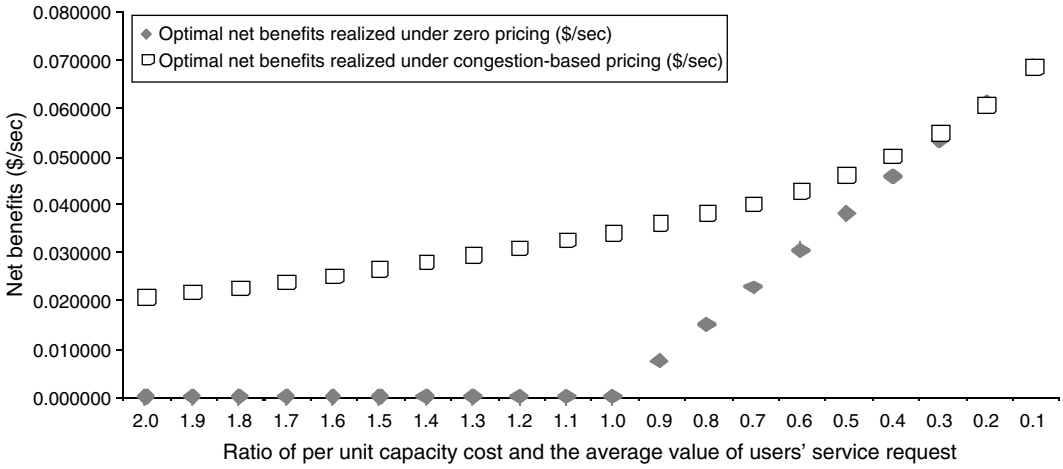
The results in Figure 4 clearly demonstrate one of the key points of this paper: *In a network of diverse servers offering a variety of services, the impact of congestion-based pricing on the incentives for network capacity investment will depend on the relationship between the per unit cost and the average value of the users' service requests*. If that ratio is high, congestion-based pricing will enable the existence of a network that simply would not be built with flat-rate pricing. However, if that ratio is sufficiently low, the implementation of flat-rate 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 low cost is so large that the congestion can be eliminated, resulting in zero prices for individual submissions. These findings can be summarized as follows.

Observation 4. The optimal level of systemwide capacity is higher under flat-rate pricing when the per unit cost of the capacity is lower than the average value of the users' requests. When that price/performance threshold is reached, i.e., the cost of capacity is equal to the average valuation; building a large system will be the optimal strategy under flat-rate 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 placing a high value on a service relative to the cost of its delivery, congestion-based pricing is necessary to provide investment incentives. However, once a high valuation is attached to the network usage as compared to capacity costs, fixed price access provides sufficient incentives to expand the network capacity.

Observation 4 provides additional economic rationale for the movement of Internet access pricing from usage-based pricing to flat-rate pricing. As enough network capacity was facilitated, and as the users' value of network services increased with respect to capacity costs, usage-based pricing became less desirable because the simpler flat-rate pricing could provide appropriate investment incentives.[24]

[24] Figure 3 shows that benefits generated under congestion-based pricing are never inferior to those guaranteed under zero pricing. However, as capacity levels increase, the difference becomes negligible, and easily offset by other positive factors tied to flat-rate charging (not accounted for in our model), such as simplicity and convenience.

Figure 5 Optimal Benefits ($/second) Under Both Pricing Schemes as the Unit Cost of Capacity Drops



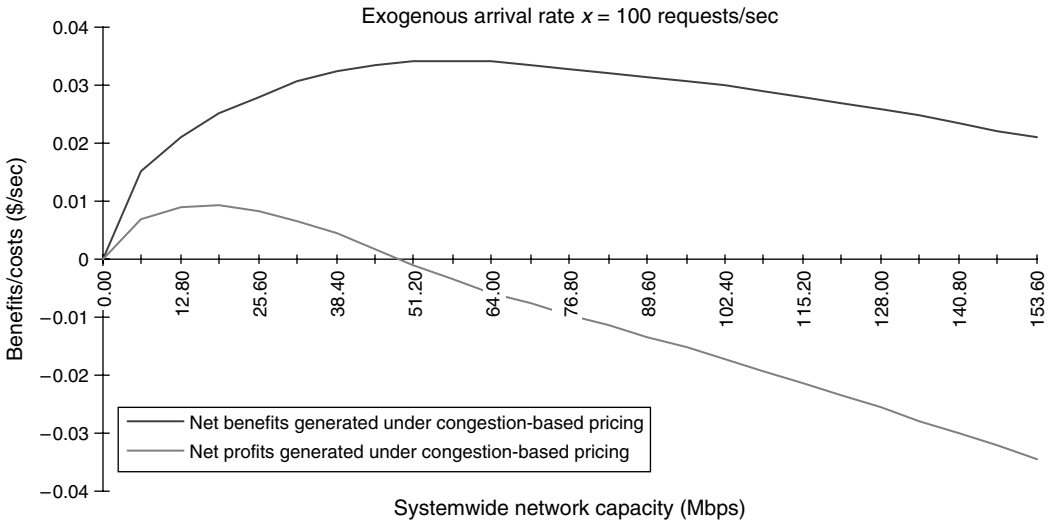
It is tempting to think that the price/performance ratio of telecommunication technology is such that it corresponds to the right-hand side of Figure 4, with flat-rate pricing providing more incentive to expand. This conclusion is valid if we assume that while the cost of the network capacity is dropping, the consumer's valuation of transmissions per unit of time remains the same. However, if the user's diminishing marginal value of quality and the frequency of completed transmissions is taken into account, the assumption may not hold. For example, the same service can be offered with increasing levels of quality (implying greater size in units of processing capacity), which results in greater valuation of that service to the user (such as better resolution of a video segment, or higher fidelity of an audio transmission). However, this growth in service's value to the user may be smaller compared to the growth in its size, and underlying required capacity needed to facilitate its increased size. Also, the rate of bandwidth usage by network users is increasing even for applications that are functionally the same; for example, more people use rich text e-mails now than earlier. Even if we assume that the current level of per unit cost of the capacity compared to the average value of the user's request is low (possibly very low), resulting in a higher level of capacity under flat-rate pricing, the performance question still remains. As we stated in Observation 2, congestion-based pricing always delivers higher systemwide net benefits at any given system capacity level, including the maximal capacity level.

Figure 5 shows the levels of net benefits delivered at the maximal capacity levels under both pricing schemes, as we gradually decrease the ratio between the cost of capacity and the average value of the users' request. This chart indicates that, while congestion-based pricing is not consistently more investment-friendly when compared with flat-rate pricing, it generally outperforms flat-rate pricing with respect to the amount of net benefits delivered. As the per unit cost of capacity decreases relative to the average value of users' requests, that difference becomes smaller. This observation indicates that a significant reduction in the cost of network resources renders the choice of access pricing schemes irrelevant. However, we emphasize again that while the absolute per unit of the capacity cost of network equipment and infrastructure is dropping at a fast rate, it is less clear whether such a cost relative to the users' perceived value of services, that such a network delivers is following the same trend.

5.2. Profits Generated Through Congestion-Based Pricing as the Network Expands

It appears that there is a discrepancy between our results and the present state of private and public networks, if one assumes that at present the per unit cost of network capacity is such that the ratio between this cost and the average value of the users' request is

Figure 6 Net Benefits and Net Profits ($/sec.) Under Congestion-Based Pricing and Original Unit Capacity Cost for the Exogenous Arrival Rate of 100 Requests/sec.



Note. Capacity is expressed in Mbits/sec.

low. According to our results, regardless of the pricing scheme, the optimal strategy is to expand networks to a level where there is very little congestion, and most requests are served with a small amount of delay. However, there is no proof that many networks, private or public, are at a capacity level where there is very little congestion. Nonetheless, we do see a major effort by telecommunication companies to expand the reach of broadband access. Furthermore, we would like to point out that it is not very realistic to assume that maximizing social net benefits is the objective for network service providers (excluding private corporate networks designed to serve the internal population of users to the maximum benefit of the company).[25] Therefore, in this subsection, we address the issue of capacity expansion with profit maximization as the objective. Figure 6 shows the profits and net benefits under congestion-based pricing as the network capacity expands. We conducted this analysis at various exogenous arrival rates (50, 200, and 500 requests per second) to simulate different load conditions. The results are provided in the online Appendix A4. These results indicate that implications of our findings are robust and none of the results and implications change under a wide array of exogenous load conditions.

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,[26] suggesting that if a benevolent monopolist was to expand the network to the optimal level, most of the expansion cost could be covered by the profits generated through congestion-based pricing. 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 underinvest. 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. Still, we cannot fully claim the generality of our simulation results, because 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 the profit-maximizing capacity distribution at each stage may be different

[25] Even in those cases, providers of network and other information technology (IT) services are often required to behave as profit centers.

[26] For our four arrival rates 97.7%, 80.5%, 88.3%, and 92% of capacity costs were covered by profits, respectively.

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. Clearly, profit-maximizing objective may not optimize load balancing for the total cost minimizers. A profit-maximizing firm, for example, may provide higher capacity reducing users' delay costs, and can therefore charge higher prices as compared to socially optimal prices. This, of course, assumes that return on investment (i.e., the aggregate savings in delay cost because of higher capacity, which can be extracted as higher prices) is higher than the cost of capital. Therefore we can no longer model this as a short-term problem independent of the investment costs. From another perspective, when a profit-maximizing entity is using the congestion pricing to maximize its profits, the equivalence trade-off between cost of delay and prices they can charge is inherently reached by under investing. Structurally, if we assume that users' decision problem is the same (i.e., minimization of total cost), the profit-maximization problem will still have to use the delay costs because of the additional constraints for individual rationality constraint (i.e., $V_{is} \geq \sum_m (p_{sm} + \delta_{is}\tau_{sm})x_{ism}$). This would result in structurally similar optimal prices as derived by Gupta et al. (1996). Intuitively, one of two cases could take place for higher profits to emerge: (i) the load on the machines is higher as compared to the socially optimal loads, which implies (comparative statically) that the capacity is lower as compared to the one resulting from the socially optimal prices; or (ii) if the capacity with the profit maximization is the same as the capacity with socially optimal prices, then the higher prices could be charged until the resulting prices reduce the load, and thereby providing higher net benefits to the users. We computationally solve the profit-maximizing problem in our simulations. In other words, we provide conditions so that equal or higher capacities can be supported with profit maximization. The result of these computational solutions seem to result in the first condition presented above, i.e., profit-maximization strategy results in higher loads on machines and lower capacity.

Essentially, our analysis models a situation where the network provider is primarily motivated by the maximization of value in the day-to-day operations and load balancing of its system. On other hand, more strategic capacity expansion decisions are made with profit maximization (or loss minimization) in mind. To model the profit-maximization problem in this context, we need to make assumptions about investment costs, which we want to express as a result in terms of value/cost ratio of technology allowing, in our opinion, for stronger managerial interpretations.

We summarize our findings in the following observation.

Observation 5. Profit-maximizing network expansion will result in a choice of systemwide capacity that is below the socially optimal systemwide capacity level.

There were no profits generated under our model of flat-rate pricing policy, because we do not model the choice of the optimal periodical fee for network users. However, Zhang et al. (1998) show that even the best flat-rate pricing policies will generate a lower level of profits compared to congestion-based pricing at a fixed level of capacity. This observation, may lead to a conclusion that the profit-maximizing capacity under flat-rate pricing may also be lower. However, the shape of profit function needs to be observed, and in our future work, we plan to provide an analysis of the impact of profit-maximizing flat-rate strategy on capacity expansion.

6. Conclusions, Limitations, and Recommendations for Future Research

Comparing two pricing policies and studying their impact on the level of capital investment in a simulated network economy provides a performance benchmark for the resource distribution on large public networks. We compared the investment policies under different pricing schemes and investment criteria such as profit maximization. The prevailing wisdom on network infrastructure investment 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*. The firms providing the network infrastructure are affected by a complex competitive telecommunications and network services landscape that are not addressed in the context our

model. Nonetheless, our simulation results provide a nuanced view of the picture. In particular, if the overall goal is to maximize social welfare, while our results support the prevailing wisdom when relative cost of capacity is low with respect to the value of services, we find that congestion-based pricing policy can result in a more robust capacity investment if the relative cost of capacity is high as compared to the value of services.

We derived an analytical expression for the optimal expansion of network resources using measurable performance parameters. We believe that the rational approach to the management and expansion of a company's internal computing resources would be far more beneficial than the common approaches toward investment in computing resources that rarely involve formal and precise cost-benefit analysis. Additionally, most of the current IT investment policies are based on prompting users to self-report their need for more resources, and, in turn, are inherently incentive incompatible. Our approach provides decision makers with a much better view of the companywide need for computational facilities.[27] However, a significant amount of additional research has to be performed before generally applicable guidelines for internal network expansion can be developed.

Our simulation results provide interesting insights into the evolution of Internet access pricing. For example, the usage-based pricing yielded to current flat-rate pricing. Our results suggest that this change in pricing practices occurs as people's value for network us in comparison to access costs decreased, and flat-rate pricing began to provide enough incentives for capacity expansion for basic Internet access. Interestingly, our results suggest that if users' value for network usage in comparison to capacity costs increases, because of the much higher amount of capacity needed to deliver more interactive services, the same flat-rate pricing might be a roadblock in providing a higher level of capacity. One practical insight that network and network service managers can derive from our findings is that it may be advisable to have pay-per-use or other usage-based pricing on top of basic access for high bandwidth access rather than higher flat-rate pricing. Such an approach can provide enough capital to cover an investment, and, in turn, may have the desired effect of higher service valuations by providing more services resulting in positive externalities.

Note that the socially optimal capacity provides the optimal service quality from an economic perspective. Deviation from the socially optimal capacities will result either in lowered welfare because of disproportional deterioration of service quality, resulting from under provision of capacity, or lowered welfare because of disproportionally lower gains in benefits because of overprovision of capacity. Our analysis of a profit-maximizing network provider only applies to settings where the users are atomistic and do not have the market power and technical sophistication to affect and monitor the network performance. Ganesh et al. (2007) have proposed an algorithm for the atomistic users that is analogous to the "fictitious play" in game theory, in which users respond to the congestion prices by choosing the arrival rate on the basis of the *predicted* price in each time slot. However, as opposed to our approach, their model does not account for the waiting cost. In addition, the rate of convergence of the expectation to predicted price is dependent on the number of time slots, which could be extremely large with a reasonable-sized network, and its users such as the ones simulated here.

Another interesting setting would be to consider network users as the part of a group (such as a large organization). In such a setting, a profit-maximizing network provider may not be able to under provision the capacity because of the penalties that may be imposed because of service level agreements (SLAs). This analysis is beyond the scope of this paper; however, Sen et al. (2008) provide an excellent analysis of design and monitoring of SLAs in stochastic environments.

Finally, we pointed out the influence of several factors at the most desirable levels of network capacity. Especially, we emphasized the importance of relative valuation of services compared to the cost of providing them, and a relative tolerance for the delay in receiving these services. We hope that these observations can provide motivation for careful monitoring of changes and trends in these factors when planning the

[27] To learn more about the incentive compatibility of our pricing mechanism, see Gupta et al. (2000).

future buildup and deployment of private and public networks, regardless of which pricing scheme is used.

Our approach and the model on which it is based does have limitations in its ability to provide a full range of managerial and policy implications. While our setting shows how incremental investments would be different in two specific pricing contexts, it still leaves a lot of important questions unanswered. For instance, content providers (as opposed to ISPs) have their own incentives to reduce latency, as well as charge customer premium pricing for improved service levels. In addition, as mentioned earlier, when the users are not atomistic and have market power, a different set of capacity investment strategies may have to be adopted by profit-maximizing service providers. We believe a variety of modeling and methodological approaches need to be employed to build a complete set of tool chest and managerial insights for this important class of problems.

We plan to investigate several research issues in the future. First, we would investigate a much richer realm of gradient estimation techniques to compute the optimality direction of capacity expansion. We will also test our analytical expression under different network specifications, e.g., by varying the number and availability of services, and monitoring how these changes affect the assumptions made to obtain the analytical solutions. Finally, this research will not be complete without an investigation of the impact of competitive pricing.

Acknowledgments

This research was funded, in part, by the National Science Foundation Grant IIS-0219825, but does not necessarily reflect the views of NSF. The first author's research is supported, in part, by NSF CAREER Grant IIS-0301239.

References

ABI Research Press Release. 2007. Cable bandwidth crisis approaching. ABI Research, New York. Accessed December 23, 2009, http://www.abiresearch.com/abiprdisplay.jsp?pressid=907.

Bailey, J., S. Gillett, D. Gingold, B. Leida, D. Melcher, J. Reagle, J. Roh, R. Rothstein. 1995. *MIT Internet Economics Workshop Notes*. MIT Press, Cambridge, MA. Accessed December 22, 2009, http://www.press.umich.edu/jep/works/BailWNotes.html.

Clarke, I., O. Sandberg, B. Wiley, T. W. Hong. 1999. Freenet: A distributed anonymous information storage and retrieval system. Freenet white paper, http://www.sandbergs.org/oskar/clarke00freenet.pdf.

Dewan, S. 1996. Pricing computer services under alternative control structures: Tradeoffs and trends. *Inform. Systems Res.* **7**(3) 301–307.

Dewan, S., H. Mendelson. 1990. User delay cost and internal pricing for a service facility. *Management Sci.* **36**(12) 1502–1517.

Ganesh, A., K. Laevens, R. Steinberg. 2007. Congestion pricing and noncooperative games in communication networks. *Oper. Res.* **55**(3) 430–438.

Gupta, A., D. Stahl, A. Whinston. 1996. An economic approach to networked computing with priority classes. *J. Organ. Comput. Electronic Commerce* **6**(1) 71–95.

Gupta, A., D. O. Stahl, A. B. Whinston. 1997a. A stochastic equilibrium model of Internet pricing. *J. Econom. Dynam. Control* **21** 697–722.

Gupta, A., D. O. Stahl, A. B. Whinston. 1997b. Priority pricing of integrated services networks. L. W. McKnight, J. P. Bailey, eds. *Internet Economics*. MIT Press, Cambridge, MA, 323–352.

Gupta, A., B. Jukic, M. Parameswaran, D. Stahl, A. Whinston. 1997c. Streamlining the digital economy: How to avert a tragedy of the commons. *IEEE Internet Comput.* **1**(6) 38–46.

Gupta, A., B. Jukic, D. Stahl, A. Whinston. 2000. Extracting consumers' private information for implementing incentive-compatible Internet traffic pricing. *J. Management Inform. Systems* **17**(1) 9–29.

Harris, T. E., P. Meier, J. W. Tukey. 1950. Timing of the distribution of events between observations. *Human Biol.* **22** 249–270.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Springer-Verlag, New York.

IEEE Internet Computing Editors. 1997. Bob Metcalfe on what's wrong with the Internet: It's the economy, stupid. *IEEE Internet Comput.* **1**(2) 6–17.

Israelsohn, J. 2003. Seven trends in fiber-optic communications. EDN Technical report, http://www.edn.com/index.asp?layout=article&articleid=CA313056.

Kaplan, E. L., P. Meier. 1958. Nonparametric estimation from incomplete observations. *Amer. Statist. Assoc. J.* **53**(282) 457–481.

Kleinrock, L. 1975. *Queueing Systems I and II*. John Wiley and Sons, New York.

MacKie-Mason, J., H. Varian. 1995. Pricing the Internet. B. Kahin, ed. *Public Access to the Internet*. MIT Press, Cambridge, MA, 269–314.

MacKie-Mason, J., L. Murphy, J. Murphy. 1995. The role of responsive pricing in the Internet. *MIT Workshop on Internet Econom.* Accesssed December 22, 2009, http://www.press.umich.edu/jep/works/MackieResp.html.

Mas-Collel, A., M. Whinston, J. Green. 1995. *Microeconomic Theory*. Oxford University Press, Oxford, UK.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Comm. ACM* **28**(3) 312–321.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the *M/M/*1 queue. *Oper. Res.* **38**(5) 870–883.

Naor, P. 1969. On the regulation of queue size by levying tolls. *Econometrica* **37**(1) 115–124.

Odlyzko, A. 2000. Should flat-rate Internet pricing continue? *IT Professional* **2**(5) 48–51.

Ratnasamy, S., P. Francis, P. Handley, R. Karp. 2001. A scalable content addressable network. *Proc. ACM Special Interest Group on Data Communications, San Diego*, 161–172.

Rowstron, A., P. Druschel. 2001. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. *International Federation for Information Processing ACM Internat. Conf. Distributed Systems Platforms (Middleware)*. Heidelberg, Germany, 329–350.

Savage, S., T. Anderson, A. Aggarwal, D. Becker, N. Cardwell, A. Collins, E. Hoffman, J. Snell, A. Vahdat, G. Voelker, J. Zahorjan. 1999. Detour: A case for informed Internet routing and transport. *IEEE Micro* **19**(1) 50–59.

Sen, S., T. S. Raghu, A. Vinze. 2008. Demand heterogeneity in IT infrastructure services: Modeling and evaluation of a dynamic approach to defining service-level agreements. *Inform. Syst. Res.* **20**(2) 258–276.

Stahl, D., Whinston A. 1994. An economic approach to client-server computing with priority classes. W. W. Cooper, A. B. Whinston, eds. *New Directions in Computational Economics*. Kluwer Academic Publishers, Boston, 71–95.

Stidham, S. 1985. Optimal control of admission to a queuing system. *IEEE Trans. Automatic Control* **38**(8) 705–713.

Stidham, S. 1992. Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Sci.* **38**(8) 1121–1139.

Stoica, I., R. Morris, D. Karger, M. F. Kaashoek, H. Balakrishnan. 2001. Chord: A scalable peer-to-peer lookup service for Internet applications. *Proc. ACM SIGCOMM, San Diego*, 149–160.

TelegeographyComms Press Release. 2005. Bandwidth demand outpaces price declines. Accessed December 22, 2009, http://www.telegeography.com/cu/article.php?article_id=6651.

TelegeographyComms Update. 2006. The bandwidth glut is over. Accessed December 22, 2009, http://www.telegeography.com/cu/article.php?article_id=12155.

Zhang, K., D. Stahl, A. Whinston. 1998. A simulation study of competitve Internet pricing: AOL flat rate versus GSW usage prices. *ACM Proc. First Internat. Conf. Inform. Comput. Econom., Charleston, SC*, 68–76.

Zhao, B., L. Huang, J. Stribling, S. Rhea, A. Joseph, J. Kubiatowicz. 2004. Tapestry: A resilient global-scale overlay for service deployment. *IEEE J. Selected Areas Comm.* **22**(1) 41–53.