Network Economies for Internet Distributed Systems: Management Implications

Hans-Werner Gottinger*

STRATEC, Munich, Germany

Abstract: A macroscopic view of Internet -based distributed computer systems reveals the complexity of the organization and management of the resources and services they provide. The complexity arises from the system size (e.g. number of systems, number of users) and heterogeneity in applications (e.g. online transaction processing, e-commerce, multimedia, decision support, intelligent information search) and resources (CPU, memory, I/O bandwidth, network bandwidth and buffers, etc.) In a large distributed system, the set of systems, users and applications is continuously changing. In this paper we address some of the management issues of providing Quality of Service (QoS), pricing, and efficient allocation of resources (computational resources) in networks and systems facilitated through economic mechanism design.

Keywords: Internet economics, network economy, distributed systems, mechanism design, performance management.

'…*The Internet is in its essence an economy, a game, an arena where agents act selfishly and are affected by everybody's decisions...*'

Hall, Nikolova and Papadimitriou, in Deng and Graham [11].

1. INTRODUCTION

With advances in computer and networking technology, thousands of heterogeneous computers can be interconnected to provide a large collection of computing and communication resources. These systems are used by a growing and increasingly heterogeneous set of users which are identified with the present Internet. A macroscopic view of distributed computer systems reveals the complexity of the organization and management of the resources and services they provide. The complexity arises from the system size (e.g. number of systems, number of users) and heterogeneity in applications (e.g. online transaction processing, e-commerce, multimedia, intelligent information search, auctions) and resources (CPU, memory, I/O bandwidth, network bandwidth and buffers, etc.)

The complexity of resource allocation is further increased by several factors. First, in many distributed systems, the resources are in fact owned by multiple organizations. Second, the satisfaction of users and the performance of applications is determined by the simultaneous application of multiple resources. For example, a multimedia server application requires I/O bandwidth to retrieve content, CPU time to execute server logic and protocols, and networking bandwidth to deliver the content to clients. The performance of applications may also be altered by trading resources. For example, a multimedia server application may perform better by releasing memory and acquiring higher CPU priority, resulting in smaller buffers for I/O and networking but improving the performance of the communication protocol execution (Gupta and Stahl [1]).

Finally, in a large distributed system, the set of systems, users and applications are continuously changing. In this paper we address some of the issues in managing Quality of Service (QoS) and pricing, and efficient allocation of resources (computational resources) in networks and systems. Resource allocation in networks relate to computational models of networks, as developed in the works of Radner [2], Mount and Reiter [3] Mount and Reiter [4, Chap.4], van Zandt [5]. The economic features of Internet based network economies have beeen treated by Gottinger [6, Chap.9] which we specifically refer to. Here they emanate from certain types of queueing systems, Kleinrock [7] and Wolff [8] on generalized networks.

The structure of this paper is as follows. Sec. 2 exhibits some broader design criteria on large scale networks which underlie the heterogeneity of Internet based resource allocation and use.

Sec. 3 shows the major components of an interface architecture with which an 'economically enhanced resource manager' (Macias *et al.* [9]) is confronted.

Sec. 4 indicates more broadly the scope of mechanism design approaches that link economic modeling to computational resources, at the interface of economics, computer and management science.

Sec. 5 deals with a specific class of problems arising in mechanism design how in resource allocation processes pricing schemes have to be made 'incentive compatible'.

Sec. 6 relates to the basic structure of a data management economy which more recently features in major application areas as in grid computing, cloud computing, sponsored search auctions, broadcast protocols, and other areas like procurement auctions, spectrum auctions, communication networks, supply chain formation and social networks.

*Address correspondence to this author at STRATEC, Munich, Germany; Tel: ????????????????????????; Fax: ?????????????????????;
E-mail: hg528@bingo-ev.de

Strategic management issues emerging through resource provisioning and pricing are covered in Sec. 7, Conclusions follow in Sec. 8.

2. DESIGN GOALS

In designing resource allocation and control mechanisms in complex distributed systems and networks several goals need to be considered and could be traced in the literature in more detail, i.e. among others, Shenker *et al.* [10], Deng and Graham [11], and Neumann *et al.* [12]

- Partition large complex allocation problems into smaller, disjoint allocation problems
- Decentralize resource access, allocation and control mechanisms
- Design reliable, fault-tolerant and robust allocation mechanisms
- Design scalable architectures for resource access in a complex system
- Provide QoS guarantees to users of applications, on performance criteria and develop pricing mechanisms for QoS based charging. Some of the QoS criteria in distributed systems include the following: (i) average response time, (ii) throughput, (iii) application failure probability, (iv) information loss, packet loss in communication networks, (v) maximum response time (message delay)
- Define system performance criteria that reflect in aggregate the diverse individual criteria of users and applications
- Design a unified framework in which users have transparent access to the services of a distributed system, and services are provided in an efficient manner

3. THE RATIONALE OF ECONOMIC MODELS IN NETWORKING

There are intrinsic interfaces between human information processing and networking that show the usefulness of economic modeling.

Decentralization: in an economy, decentralization is provided by the fact that economic models consist of agents which selfishly attempt to achieve their goals. Suppose there are two types of economic agents: suppliers and consumers. A consumer attempts to optimize its individual performance criteria by obtaining the resources it requires, and is not concerned with system-wide performance. A supplier allocates its individual resources to consumers. A supplier's sole goal is to optimize its individual resources to consumers. A supplier's sole goal is to optimize its individual satisfaction (profit) derived from its choice of resource allocation to consumers.

Limiting Complexity: economic models provide several interesting contributions to resource sharing algorithms. The first is a set of tools for limiting the complexity by decentralizing the control of resources. The second is a set of mathematical models that can yield several new insights into resource sharing problems, such as the tools provided by Scarf [13].

Pricing and Performance: most economic models introduce money and pricing as the technique for coordinating the selfish behavior of agents. Each consumer is endowed with money that it uses to purchase required resources. Each supplier owns a set of resources, and charges consumers for the use of its resources. The supplier prices its resources based on the demand by the agents, and the available supply. Consumers buy resources or services such that the benefit they receive is maximized. Consumer-agents buy resources based on maximizing performance criteria. As a whole the system performance is determined by some combination of the individual performance criteria.

Usage Accounting, Billing and Dimensioning: by using economic models for service provisoning in distributed systems, accounting for QoS becomes an important task for suppliers, as they have to keep track of the resource usage in order to price the resources, and thereby charge or bill the users for QoS. In addition, pricing can be used to understand the user demands and thereby dimension systems appropriately.

Administrative Domains: often large distributed systems and computer networks spread over several domains, the control of resources is shared by multiple organizations that own distinct parts of the network. In such an environment, each organization will have a set of services that it supports. Economic principles of pricing and competition provide several valuable insights into decentralized control mechanisms between the multiple organizations and efficient service provisioning.

Scalability: a key issue in designing architectures for services in large computer networks and distributed systems is scalability. With the ever growing demand for new services, flexible service architectures that can scale to accommodate new services is needed. Economic models of competition provide, in a natural fashion, mechanisms for scaling services appropriately based on service demand and resource availability.

4. MECHANISM DESIGN APPROACHES

Network allocation and pricing could be looked at as part of mechanism design theory (Hurwicz and Reiter [14]) and in differential form by Williamson [15]. In a more economic historical context the justification for linking market mechanism to computational resource allocation may be attributed to the Austrian economist F.A. Hayek [16], so what we suggest

For an Internet based distributed system is a sort of Hayekian mechanism design (This may fly into the face of many interventionistic Internet economists). More specific mechanism design approaches for distributed networks and grid-type systems are covered by Narahari *et al.* [17] and Neumann *et al.* [12], see also Meinel and Tison [18]. In the context of computational resources, specifically, an algorithmic mechanism design uses a computational platform with an output specification and agents' preferences represented by utilities (Nisan and Ronen [19]).

In its general form for distributed systems, the user can indicate the 'type' of transmission and the workstation in turn reports this type to the network. To ensure truthful revelation of preferences, the reporting and billing mechanism must be incentive compatible.

Most studies of resource allocation mechanisms have used a performance model of the resource, where the very concept of the resource is defined in terms of measurable qualities of the service such as utilization, throughput, response time (delay) and so on. Optimization of resource allocation is defined in terms of these measurable qualities, as a basis of performance management. One novelty introduced by the economic approach is to design a system which takes into account the diverse QoS requirements of users, and therefore use multiobjective (utilities) optimization techniques to characterize and compute optimum allocations. Economic modeling of computer and communication resource sharing uses a uniform paradigm described by two level modeling: QoS requirements as inputs into a performance model that is subject to economic optimization.

In the first step, one transforms QoS requirements of users to a performance (example: queueing service model). This model establishes quantifiable parameterization of resource allocation. For example, average delay QoS requirement, when based on a FIFO queueing model, is a function of resources, bandwidth and buffer, and user traffic demands. These parameters are then used to establish an economic optimization model. The question of whether the resource is a piece of hardware, a network link, a software resource such as a database or a server, or a virtual network entity such as a TCP/IP connection is not of primary importance. The first modeling transformation eliminates the details and captures the relevant behaviors and the optimization parameters.

A reasonable approach to follow evolves in the following sequence. Many users present QoS demands, which are translated into demands on resources based on a performance model. The suppliers compute the optimal allocations based on principles of economic optimization and market mechanisms. Once the optimization is done, the results provide inputs to mechanisms for QoS provisioning, such as scheduling of resources and admission of users in networks and load balancing in distributed systems.

Optimal Allocation and QoS

We establish and solve a problem of allocating resources and providing services (QoS) to several classes of users at a single link (Gottinger [6], Chap. 9). The resources at the link are buffer space and bandwidth. The link (network provider) prices per unit buffer and bandwidth resources.

A simple example on the representation of QoS parameters is the bandwidth-buffer tradeoff. Bandwidth can be traded for buffer space and vice versa to provide the same QoS. If a bandwidth is scarce, then a resource pair that uses less bandwidth and more buffer space should be used. Resource pricing is targeted to exploit this tradeoff to achieve efficient utilization of the available resources. The pricing concept for a scarce resource is well-known in economics, but in the context of exploiting the bandwidth-buffer tradeoff, Low and Varaiya [20] used non-linear optimization theory to determine centralized optimal shadow prices in large networks. With respect to large scale application, however, the complex optimization process limits the frequency of pricing updates, which causes inaccurate information about available resources. In order to make pricing in the context of a buffer-bandwidth tradeoff more adjustable and flexible it should be based on decentralized pricing procedures according to competitive bidding in large markets where prices will be optimal prices if the markets are efficient. This would also allow flexible pricing which results in accurate representation of available resources in that prices are updated as the instance connect request arrives. The subsequent procedure is based on distributed pricing as a more feasible alternative to optimal pricing.

Here are the steps involved to invoke an incentive compatible pricing scheme based on QoS needs.

The consumers (user traffic classes), *via* economic agents, buy resources such that their QoS needs are satisfied. The network provider prices resources based on demand from the consumers. The ingredients are as follows:

- Economic models: use competitive economic models, of the type as outlined by Scarf [13], to determine the resource partitions between user traffic classes, which compete to obtain buffer and bandwidth resources from the switch suppliers.
- Optimal allocations using economic principles: look for Pareto optimal allocations that satisfy QoS needs of agents. Agents represent QoS *via* utility functions which capture the multiple performance objectives.
- Pricing based on QoS : compute equilibrium prices (or approximate prices) based on the QoS demands of consumers. Prices are set such that the market demand and supply are met. Prices help in determining the cost of providing a service. (In practical application this may be a hard task to do.)
- Priorities: using the economic framework, show a simple way to support priority service among the user-classes (or agents).
- Decentralization: show a natural separation between the interactions of the user-classes (represented by agents) and the network switch suppliers. The interaction is purely competitive and market based. This decentralization promotes scalable network system design.

Scheduling and Pricing Mechanisms

Consider a dynamic system where sessions arrive and leave a traffic class, and demand fluctuates over time. In such a setting, we investigate practical mechanisms, such as packet level scheduling to provide bandwidth and buffer guarantees, admission control mechanisms to provide class QoS guarantees, practical pricing to capture the changing demand, and charging mechanisms for user sessions within a class.

- Scheduling algorithms for class based QoS provisioning: provide novel scheduling mechanisms, which allocate bandwidth and buffer for meeting the demand from traffic classes. The scheduling mechanism allocates bandwidth, which is computed from the economic optimization.

- Admission Region and Control: compute the admission control region of the agents on the economic model. Due to the natural separation between those who control the admission of sessions into the traffic class, the admission region can be determined.
- We propose simple pricing models which capture the changing demand, and are easy to implement. We also propose novel QoS based charging mechanisms for sessions in a class with applications to charging in ATM Networks and Integrated Services Internet.

Network and Server Economies

Consider first a network economy, of many parallel routes or links, where several agents (representing user classes) compete for resources from several suppliers, where each supplier represents a route (or a path) between a source and destination. Agents buy resources from suppliers based on the QoS requirements of the class they represent. Suppliers price resources, independently, based on demand from the agents. The suppliers connect consumers to information providers, who are at the destination; the flow of information is from information providers to the consumers. We formulate and solve problems of resource allocation and pricing in such an environment.

Then consider a server economy in a distributed system. Again, we use a similar model of interaction between agents and suppliers (servers). The servers sell computational resources such as processing rate and memory to the agents for a price. The prices of resources are set independently by each server based on QoS demand from the agents. Agents represent user classes such as transactions in database servers or sessions for Web servers that have QoS requirements such as response time. Examples are given in Gottinger [21].

Server Economy: architecture for interaction: Consider a large scale distributed information system with many consumers and suppliers. Suppliers are content providers such as web servers, digital library servers, multimedia database and transaction servers. Consumers request for and access information objects from the various suppliers and pay a certain fee or no fee at all for the services rendered.

Consider that third party suppliers provide information about suppliers to consumers in order to let consumers find and choose the right set of suppliers.

Access and dissemination: consumers query third-party providers for information about the suppliers, such as services offered and the cost (price). Likewise, suppliers advertise their services and the costs *via* the third party providers in order to attract consumers. Consumers prefer an easy and simple way to query for supplier information, and suppliers prefer to advertise information securely and quickly across many regions or domains. For example, consider a user who wishes to view a multimedia object (such as a video movie). The user would like to know about the suppliers of this object, and the cost of retrieval of this object from each supplier.

Performance requirements: users wish to have good response time for their search results once the queries are submitted. However, there is a tradeoff. For more information about services offered, advanced searching mechanisms are needed, but at the cost of increased response time. In other words, users could have preferences over quality of search information and response time. For example, users might want to know the service costs in order to view a specific information object. In large networks, there could be many suppliers of this object, and users may not want to wait forever to know about all the suppliers and their prices. Instead, they would prefer to get as much information as possible within a certain period of time (response time).

From the above example, in order to let many consumers find suppliers, a scalable decentralized architecture is needed for information storage, access and updates.

Naming of services and service attributes of suppliers becomes a challenging issue when hundreds of suppliers spread across the globe. A simple naming scheme to connect consumers, across the Internet, with information about suppliers is essential. The naming scheme must be extensible for new suppliers who come into existence. A name registration mechanism for new suppliers and a de-registration mechanism (automatic) to remove non-existent suppliers is required. In addition, naming must be hierarchical, domain based (physical or spatial domains) for scalability and uniqueness. Inter-operability with respect to naming across domains is an additional challenging issue not covered in this paper.

The format of information storage must be simple enough to handle many consumer requests quickly within and across physical domains. For better functionality and more information, a complex format of information storage is necessary, but at the cost of reduced performance. For example, a consumer, in addition to current service cost, might want to know more information such as the cost of the same service during peak and off-peak hours, the history of a supplier, its services, and its reputation, in order to make a decision. This information has to be gathered when requested. In addition, the storage formats must be inter-operable across domains.

Performance: a good response time is important to make sure consumers get the information they demand about suppliers within a reasonable time period, so that decision-making by consumers is done in a timely fashion. In addition, the design of the right architectures for information storage and dissemination is necessary for a large scale market economy to function efficiently. Using the previous example, consumers and suppliers would prefer an efficient architecture to query for and post information. Consumers would prefer good response time in obtaining the information, and suppliers prefer a secure and fast update mechanism to provide up-to-date information about their services.

Security in transferring information and updating information at the bulletin boards (name servers) is crucial for efficient market operation and smooth interaction between consumers and suppliers. For this the third party suppliers (naming services) have to provide authentication and authorization services to make sure honest suppliers are the ones updating information about their services.

5. ALLOCATION AND PRICING MODELS

In economic models, there are two main ways to allocate resources among the competing agents. One of them is the exchange based economy and the other is the price based economy. In the exchange based economy, each agent is initially endowed with some amounts of the resources. They exchange resources until the marginal rate of substitution of the resources is the same for all the agents. The agents trade resources in the direction of increasing utility (for maximal preference). That is, two agents will agree on an exchange of resources (e.g. CPU for memory) which results in an improved utility for both agents. The Pareto optimal allocation is achieved when no further, mutually beneficial, resource exchanges can occur. Formally, an allocation of resources is Pareto Optimal when the utility derived by the competing economic-agents is at the maximum. Any deviation from this allocation could cause one or more economic agents to have a lower utility (which means the agents will be dissatisfied).

In a price based system, the resources are priced based on the demand, supply and the wealth in the economic system. The allocations are done based on the following mechanisms. Each agent is endowed with some wealth. Each agent computes the demand from the utility function and the budget constraint. The aggregate demand from all the agents is sent to the suppliers who then compute the new resource prices. If the demand for a resource is greater than its supply, the supplier raises the price of the resource. If there is surplus supply, the price is decreased. The agents again compute their demands given the current prices and present the demand to the suppliers. This process continues iteratively until the equilibrium price is achieved where demand equals the supply.

Bidding and auctioning resources is another form of resource allocation based on prices. There are several auctioning mechanisms such as the Sealed Bid Auction, Dutch Auction, and English Auction. The basic philosophy behind auctions and bidding is that the highest bidder (or in the Vickrey auction the second highest bidder) always gets the resources, and the current price for a resource is determined by the bid prices.

What are the Economically Hard Problems?

Some of the interesting problems encountered when designing an economic based computer system are discussed and stated.

- How do agents demand resources? This is a fundamental question regarding the agents preferences on the resources they consume. Are there smooth utility functions that can capture the agents preferences of the resources? Are there utility functions that can capture the diversity of the agents preferences?
- How are the prices adjusted to clear the economy or to clear the markets? In an economic model, efficient allocation of resources occurs when the demand equals the supply at a certain equilibrium price vector.
- What rational pricing mechanisms do the suppliers adopt? This question raises issues on pricing mechanisms that will attract agents (consumers).
- How do suppliers provide price guarantees to agents? This is a fundamental question in advertising and providing price guarantees to agents. Delays in information about prices, and surges in demand can cause prices to vary. Therefore agents can make bad decisions.
- What are the protocols by which the consumers and suppliers communicate to reserve resources?

What are the general allocation principles? Can economic models give insight into the allocation mechanisms that can cause the computer system to reach equilibrium? Can these principles be used practically to evolve the computer system in a way that price equilibrium can be achieved? Even devoting the entire WINE 2007 proceedings to those issues, with active participation of K. Arrow, H. Scarf and C. Papadimitriou (in Deng and Graham [11]), still many practical issues of implementation haven't been yet finally resolved.

6. THE DATA MANAGEMENT ECONOMY

Unlike the flow control and load balancing economies where users maximize an utility function to compute the required allocation, this economy considers data migration, replication and pricing strategies for a data management economy as evidenced by large scale e-commerce facilitated through new platforms in grid computing, cloud computing and related application areas (Kushida *et al.* [22]). The problem of data migration, storage and replication is formulated in an economic setting. Transactions that enter the system for service are charged by the processors for read and write access to data objects. Processors also lease resources to other processors to make profit using the revenue they earn.

The distributed system consists of N processing nodes connected *via* links. Each processor P_i ($i \in [1, N]$) has rate r_i at which it can process operations on local data. A link e_{ij} connects processor P_i to P_j. There are M data object denoted by $D_1, D_2, ..., D_M$. $S(D_i)$ defines the size of D_i in bytes. The economy treats these as abstract data objects. In a real system, they could correspond to relations, tuples, files, records or any other data structure. The data management problem is to minimize the mean transaction response time with the following as control variables.

- Number of copies of data object
- Assignment of copies to processing nodes
- Pricing strategies of suppliers

In the data management economy there are four types of agents. The consumers are transactions, and the suppliers are data object managers, local data agents and processors as through cloud computing. The economy functions in the following way. Each transaction T that arrives has an allocation of money M_T. Transactions pay to access data at a processor P_i. Data access is provided by the processor by leasing copies of data objects from data object managers. The local data agents act as an intermediary between a processor P_i and the object managers (remote). Two economic factors cause the data management economy to adapt the number of read copies of each object D_j to the read/write ratio. These are:

- The total revenue that all processors earn by selling Read(D_j) decreases as the initially set price of the agents given its wealth p_w increases
- The read lease price for D_j increases linearly with the number of copies c(j)
- The data management economy uses decentralized decision making to compute the number of read copies of each object. The business strategies of the processors are decoupled, and P_i uses only local information to estimate its revenue. The economy adapts itself to any read/write ratio without any external intervention. The economy is not completely self tuning, however, there is a subtle interaction between the following factors:
- Lease price function
- Transaction arrival rates
- The wealth of the transactions

7. STRATEGIC INTERNET MANAGEMENT ISSUES

Universal Access. A primary concern in regulating universal access to the Internet, next to security, had been the issue of pricing its services, the maintaining of competition among providers and strengthening incentives for private investment into the network infrastructure. Possible options emerged in identifying the issues toward a workable model:

(i) charging by access to telecommunications capacity, e.g., flat rate pricing and keeping distance independent pricing

(ii) consider network externalities in the economics and growth of networks

(iii) introduce usage-based linear prices

(iv) introduce usage-based nonlinear prices

The evolution of Internet pricing poses interesting problems. Flat-rate pricing has been one of the factors that promoted the Internet to expand at a dramatic rate. It has enabled low-cost dissemination, beta-testing and refinement of new tools and applications. The strength of many of these tools is their ability to operate in and rationalize a distributed and heterogeneous structure making it easier to identify and retrieve information sources. The increased demand that has arisen due to the power and new resources these tools have brought to the Internet (and in view of lagging a corresponding capacity expansion due to advanced fiber-optic technology) is likely to create more gridlock and a need for a new pricing model. This despite new regulatory proposals on "net neutrality" emerging, usage based pricing and service charges or more specific content pricing should make the Internet attractive to many new users and also incentivize innovation driven product development on the net. One paradox of usage based pricing is that its implementation may actually cost more on a transaction basis than the underlying cost of transport. Therefore, it very much depends on network accounting capabilities as a critical implementation tool.

Congestion Problems. A natural response by shifting resources to expand technology will be expensive and not necessarily a satisfactory solution in the long run. Some proposals rely on voluntary efforts to control congestion. Others have suggested that we essentially have to deal with the problem of overgrazing the commons, e.g. by overusing a generally accessible communication network. A few proposals would require users to indicate the priority they want each of the sessions to receive, and for routers to be programmed to maintain multiple queues for each priority class. If priority class is linked to the value the users attach to it, one could devise schemes of priority pricing. This is where application of mechanism design could help. At congested routers, packets are prioritized based on bids. In line with the design of a Vickrey auction, in order to make the scheme incentive compatible, users are not charged the price they bid, but rather are charged the bid of the lowest priority packet that is admitted to the network. It is well-known that this mechanism provides the right incentives for truthful revelation. Such a scheme has a number of desirable characteristics. In particular, not only do those users with the highest cost of delay get served first, but the prices also send the right signals for capacity expansion in a competitive market for network services. If all of the congestion revenues are reinvested in new capacity, then capacity will be expanded to the point where the marginal value is equal to its marginal cost.

Quality-of-Service Characteristics. With the Internet we observe a single QoS: "best effort packet service". Packets are transported first come, first serve with no guarantee of success.

Some packets may experience severe delays, while others may be dropped and never arrive.

Different kinds of data place different demands on network services. Email and file transfer requires 100 percent accuracy, but can easily tolerate delay. Real-time voice broadcasts require much higher bandwidth than file transfers and can tolerate minor delays but cannot tolerate significant distortions. Real-time video broadcasts or video telephony over VOIP have very low tolerance for delay and distortion. Because of these different requirements, network allocation algorithms should be designed to treat different types of traffic differently but the user must truthfully indicate which type of traffic (s)he is preferring, and this would only happen through incentive compatible pricing schemes.

Internet and Telecommunications Regulation. In contrast to traditional telecommunications services Internet transport itself is currently unregulated but services provided over telecommunication carriers are not. This principle has never been consistently applied to telephone companies since their services over fixed telephone lines also used to be regulated.

There have been increasing demands, sometimes supported by established telecommunication carriers that similar regulatory requirements should apply to the Internet. One particular claim is "universal access" to Internet services, that is, the provision of basic Internet access to all citizens at a very low price or even for free. What is a basic service, and should its provision be subsidized? For example, should there be an appropriate access subsidy for primary and secondary schools? A related question is whether the government should provide some data network services as public goods.

A particular interesting question concerns the interaction between pricing schemes and market structure for telecommunications services. If competing Internet service providers offer only connection pricing, inducing increasing congestion, would other service providers be able to attract high value "multimedia" users by charging usage prices but offering effective congestion control? On the other hand, would a flat rate connection price provider be able to undercut usage-price providers by capturing a large share of baseload customers who would prefer to pay for congestion with delay rather than with a fee. Could this develop into a fragmented market with different Internets? These developments may have profound impacts to shape a future telecommunications industry which may be taken over by different structured layers of the Internet.

8. CONCLUSIONS

In this paper we focus on applications of mechanism design to resource management problems in distributed systems and computer networks. These concepts are used to develop effective market based control mechanisms, and to show that the allocation of resources are Pareto optimal. The emphasis here is on management implications given the economics of the Internet.

We follow novel methodologies of decentralized control of resources, and pricing of resources based on varying, increasingly complex QoS demands of users. We bring together economic models and performance models of computer systems into one framework to solve problems of resource allocation and efficient QoS provisioning matching large-scale e-commerce applications. The methods can be applied to pricing services in ATM networks and (wireless) Integrated Services Internet of the future. We address some of the drawbacks to this form of modeling where several agents have to use market mechanisms to decide where to obtain service (which supplier?). If the demand for a resource varies substantially over short periods of time, then the actual prices of the resources will also vary causing several side effects such as indefinite migration of consumers between suppliers. This might potentially result in degradation of system performance where the resources are being underutilized due to the bad decisions (caused by poor market mechanisms) made by the users in choosing the suppliers. As in real economies, the resources in a computer system may not easily be substitutable. The future work is to design robust market mechanisms and rationalized pricing schemes which can handle surges in demand and variability, and can give price guarantees to consumers over longer periods of time some of which have been discussed by Spulber an Yoo ([23], Chap.12). Another drawback is that resources in a computer system are indivisible resulting in non-smooth utility functions which may yield sub-optimal allocations, and potential computational overhead.

In addition to models for QoS and pricing in computer networks, we are also working towards designing and building distributed systems using market based mechanisms to provide QoS and charge users either in a commercial environment or in a private controlled environment by allocating quotas *via* fictitious money (charging and accounting) by central administrators.

In summary, economic based management is useful for implementing and operating internet-type systems. The Internet currently connects hundreds of millions of users and thousands of sites. Several services exist on many of these sites, notably the World Wide Web (WWW) which provides access to various information sources distributed across the Internet. Many more services (multimedia applications, commercial transactions) are to be supported in the Internet. To access this large number of services, agents have to share limited network bandwidth and server capacities (processing speeds). Such large-scale networks require decentralized mechanisms to control access to services. Economic/managerial concepts such as pricing and competition can provide some solutions to reduce the complexity of service provisioning and decentralize the access mechanisms to the resources.

ACKNOWLEDGMENT

This paper is partially based on my book, Chaps 8 and 9 (Gottinger [6]) with some overlapping and updated material with permission granted by the publisher NovaScience, New York which is highly appreciated.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

REFERENCES

[1] Gupta A, Stahl DO. An Economic approach to networked computing with priority classes. Cambridge, Ma.:MIT Press 1995.

[2] Radner R. The Organization of decentralized information processing. Econometrica 1993; 62:1109-46.

[3] Mount KR, Reiter S. On modeling computing with human agents. Center for Math. Studies in Economics and Management Science. Northwestern Univ., Evanston, Ill. 1994, No.1080.

[4] Mount KR, Reiter S. Computation and complexity in economic behavior and organization. Cambridge: Cambridge Univ. Press 2002.

[5] Van Zandt T. The scheduling and organization of periodic associative computation: efficient networks. Rev Ec Design 1998; 3: 93-127.

[6] Gottinger HW. Strategic Economics in network industries. New York: NovaScience 2010.

[7] Kleinrock L. Queueing systems Vol. 2. New York: Wiley 1976.

[8] Wolff RW. Stochastic modeling and the theory of queues. Englewood Cliffs, N.J.: Prentice Hall 1989.

[9] Macias M, Smith G, Rana O, Guitart J, Torres J. Enforcing service level agreements using economically enhanced resource manager. In: [12]; 109-27.

[10] Shenker S, Clark D, Estrin D, Herzog S. Pricing in computer networks: reshaping the research agenda. J Telecom Policy 1995; 20: 30-45.

[11] Deng X, Graham FC. Internet and network economics. third international workshop. WINE 2007, San Diego, Berlin, New York: Springer 2007.

[12] Neumann D, Baker M, Altmann J, Rana OF, Eds. Economic models and algorithms for distributed systems. Basel: Birkhaeuser 2010.

[13] Scarf H. The Computation of economic equilibria. cowles commission monograph. New Haven and London: Yale Univ. Press 1973.

[14] Hurwicz L, Reiter S. Designing economic mechanisms. PB Ed., Cambridge: Cambridge Univ. Press 2006.

[15] Williamson SR. Communication in mechanism design, a differential approach. Cambridge: Cambridge Univ. Press 2008.

[16] Hayek FA. The Use of knowledge in society. Am Ec Rev 1945; 35:519-530.

[17] Narahari Y, Garg D, Narayanam R, Prakash H. Game theoretic problems in network economics and mechanism design solutions. London: Springer 2009.

[18] Meinel C, Tison S, Eds. STACS 99, 16th Ann Symp. Theoretical aspects of computer science. Trier, Germany, March, Berlin: Springer 1999.

[19] Nisan N, Ronen A. Algorithmic mechanism design. 18th intern conf distributed computing systems (ICDCS-98). Amsterdam 1998; 1-45.

[20] Low S, Varaiya P A. New approach to service provisioning in atm networks. IEEE Trans. Networking 1993; 1:7-14.

[21] Gottinger HW. Network economies for the internet - application models. iBusiness 2011; 3: 313-22.

[22] Kushida KE, Murray J, Zysman J. Diffusing the fog: cloud computing and implications for public policy, berkeley roundtable on the international economy (BRIE), BRIE Working Paper 197: 2011.

[23] Spulber DF, Yoo CS. Networks in telecommunications, economics and law. Cambridge: Cambridge Univ. Press 2009.

Received: May 29, 2012 Revised: August 16, 2012 Accepted: August 22, 2012