A Pricing Framework For QoS Capable Internet

By

Safiullah Faizullah
A Pricing Framework For QoS Capable Internet

Abstract and Proposal Synopsis

With ever increasing popularity of Internet, since it became available to general public, researchers have been actively looking for means and ways to influence the behavior of her selfish users. Severe congestion can otherwise make everyone experience tremendous degradation in the service quality they could have received from the network. Pricing soon was realized as the tool to provide proper incentives (monetary, performance, and social are few to name) so that users’ self-interest will lead them to modify their usage according to their needs.

Currently, no adequate pricing scheme that answers these challenges exists. Therefore, necessitating a renewed search for an efficient solution that can result in better overall network performance.

On top of the early stage in pricing research, tremendous technological improvements resulted in better hardware, software and intelligent protocols necessitating active research in this regard. This created a networked environment that sparked a host of new and demanding services requiring stringent QoS guarantees. Therefore, reshaping the present (and elementary) pricing research so as to cope with these new challenges.

In this work, a scalable pricing framework (deployed at edge routers) for QoS capable internetwork (e.g., next generation Internet) supporting real time, adjustable real time, and non-real time traffic is proposed. The scheme, belonging to usage-based [1-3, 14, 4, 5] (and QoS dependent) methods, is independent of the underlining network and the mechanism for QoS provisioning and it can be deployed in any QoS capable (with least being best-effort) environment. The proposed pricing scheme is credit-based ensuring the fairness (from users’ point of view), comprehensibility, controllability, predictability, and stability. On the other hand, it provides means for the network providers to ensure- with high probability- cost (plus profit) recovery, competitiveness of prices, and encouragement (or discouragement) of client behaviors that will enhance (or degrade) the network’s efficiency (by appropriate charging mechanisms and suitable incentives). We believe that the implementation and usage costs of the proposed framework are low [1-3, 4]. The method covers all other aspects of a complete solution as well (refer to section 3 and 4).

After presenting the motivation in the next section, a comprehensive survey of related work is presented in section 2 (where detailed summaries of selected papers are given in Appendices A-E). The proposed scheme is presented in section 3 with implementation and deployment issues given in section 4. Goals and objectives are presented in section 5, which is followed by references.

1 Introduction and Motivation

In this section, we introduce the existing approaches for pricing services in today’s Internet as well as establish the shortcomings (and unsuitability) of these traditional pricing mechanisms. There has been very limited work in this regard. In addition, on top of this vacuum, the rapidly changing Internet characteristics (slowly transforming from best-effort to QoS capable technological giant) necessitate the need to devise new and
improved pricing framework suitable for multi-services environment (where each service can support and guarantee QoS). This change is made possible with tremendous technological improvements, which resulted in better hardware, software and intelligent protocols. Thus reshaping the whole pricing research. In section 1.1, traditional pricing schemes in Internet are introduced. The limited work on pricing schemes in QoS capable Internet (and the challenges they face) is introduced in section 1.2.

1.1 Internet and Traditional Pricing

The Internet has made a remarkable transition from research testbed (available only to a small society of aware and selfless users) to a commercial enterprise (available to a vast community of selfish users). However, this transition seems to be just the beginning, particularly with the promise and opportunities that comes with this need-of-the-day technology (more or less becoming like electricity, water and phone services of household). Some common examples of earlier Internet services include email, FTP, Telnet, IRC, Gopher, and WWW, with limited experience with real-time audio/video services such as MBONE. More recent services include audio- and video-conferencing, distributed interactive simulation activities (such as tank battle simulations), exchanging medical records and online graphic images, exchanging experimental weather maps and wireless personal digital assistants (PDAs). In the face of ever increasing popularity of the Internet (due partly to the popularity of its services), this process of transitions is poised to grow faster than the technological advancements, which have been remarkable so far. Particularly, with the advent of WWW the access to information on the Internet has become easier and more user friendly and highly data intensive. The result of availability of user-friendly interfaces and high speed of access (in gigabits) especially the one button click-to-download interfaces coupled with higher awareness about these services has already started creating sever congestion problems on the Internet. In the near future, personal publishing through WWW or its descendants and increased number of real-time audio/video applications such as teleconferencing (having stringent QoS requirements), e-commerce, e-education and many other services will be offered through the public data networks such as the Internet. This will further create significant congestion problems on the network, resulting in further degradation of service quality.

Traditionally, the Internet and its predecessors (DoD's ARPAnet and NSF's NSFNET) were funded by Federal government agencies, namely the Department of Defense and the National Science Foundation. Individual users have not been charged for their use of networks, and have not generally been aware of the impact of their use on network performance. As a result Internet users have increased substantially with unfriendly and selfish attitude. As pointed by Crawford D. [28], the Internet users are reported to have grown at a rate of 10 percent per month since 1990 when Commercial Internet Exchanges (CIX) were first connected to the Internet to allow commercial traffic. Very few technologies saw this kind of rapid growth in the number of users which resulted in increased demand for transmission capacity and hence increasing the cost of transmission. The phasing out of Federal government funding of Internet operation in the United States has necessitated some form of alternative funding, such as revenue from fee for service operation [1-3, 28]. The traditional pricing is either free (subsidized through
government funds) or flat-rate for unlimited usage. Some variations have been pricing bandwidth of the connection, or flat-rate to a certain hours and per hour charges thereafter [1-3, 4, 14].

However, the greatly increased usages of the Internet and the resultant performance degradation have focused attention on the inefficiencies of the traditional pricing structures and their shortcomings. Also, a renewed emphasis on the research to improve hardware, software and protocols is needed particularly in the absence of proper incentives to act as congestion control. While there have been dramatic and outstanding success in the infrastructure research (resulting in high bandwidth backbones having gigabits transfer capability, widespread availability of PCs, easy network connections from homes, faster routers and sophisticated protocols), there has been a severe vacuum in pricing research. Since the traffic demands increase as the bandwidth (and other resources) improves, it is a mistake to argue that over-provisioning (like the one in [15] with summary Appendix E) the capacity is the solution for achieving high network performance. Efficient pricing mechanisms coupled with traditional congestion control protocols are the ultimate solution to congestion control that will result in better overall network performance. These pricing mechanisms are based on user incentives (particularly performance versus monetary as well as administrative) seems to be the answer to the challenges posed by future Internet. An additional motivation for imposing a pricing scheme is to give users knowledge about the value of what they do to other people, and an interest to act so as to reduce harm done to others (social incentives). It is assumed that the system, which grants users the power to cause congestion, also provides users the power to reduce congestion and thereby avoid needless or inefficient harm. A generous user who is willing to use a system after hours needs to know when after hours actually occurs. A less socially benevolent user, if offered a discount for after hours usage, may reschedule his/her use, not out of charity or of concern for the public good, but because it is in his/her interest to save money. Finally, a user must have sufficient power over the system so that after having decided to save money by using resources when they are cheap, the actions taken have that result. A user who submits his/her contributions to a mailing list at night will not have any benevolent impact if his/her software accumulates mail until 8 am and then transmits his/her messages.

A potential pitfall of introducing a pricing scheme, however, is that it is not only the behavior of the consumers that may be affected, but also the behavior of the providers. Profit seeking providers will have as much knowledge, interest, and power in the system as any consumer. This requires us to search for pricing schemes that satisfy both network providers as well as network users as much as possible. From the point of view of network providers, the scheme should have the following properties: high probability of cost recovery, competitiveness of prices, encouragement (or discouragement) of client behaviors that will enhance (or degrade) the network’s efficiency, low implementation costs, and low usage costs. From the point of view of the users, the scheme should have the following properties: comprehensibility, controllability, predictability, stability, and fairness [14].
1.2 The Future Internet and Pricing

As a result of the rapid development of network technology, it is becoming increasingly efficient to provide different telecommunication services through one integrated-services network instead of multiple single-service networks. It is expected that in very near future integrated QoS capable networks (referred to as “FInternet” in this document) will emerge which provides a variety of transmission services, such as telephony, video, and file transfer and all the other traditional Internet services. The system will be capable to negotiate QoS parameters and upon accepting user connection, it will be responsible for guaranteeing the agreed quality. Best-effort will be one of such services. Pricing is important but non-critical in today’s Internet. In FInternet the issue of pricing is more relevant than it is today. For otherwise, every user can/will opt for highest QoS available thus creating a huge congestion problem. Thus the role of incentives.

Integrating multiple services into a single network generates economies of scope, however heterogeneous services complicate pricing decisions. For example, users watching HDTV through the network require up to hundreds of thousands of cells to be transmitted per second while users who make phone calls send/receive only a couple of hundred cells per second; telnet users require mean cell transmission delay to be kept below a few tens of milliseconds but e-mail senders will tolerate longer delay; web browsing generates a very bursty cell stream while constant-bit-rate file transfer results in a smooth cell stream; to carry a voice conversation with a decent quality, under certain encoding schemes, cell loss rate, i.e. the percentage of cells that are allowed to miss a maximum delay bound (usually 30-50 ms for voice conversation), should not exceed 5%, while to carry video service, cell loss rate should be kept much lower [5].

Note that in all these cases, there are great differences among the services offered by the network; therefore, one might ask whether the prices of these service should also differ, and if so, how? Also, when more than one parameters are involved, for example delay = 1ms and BDW= 4Mbs versus delay = 2ms and BDW= 8Mbs; how to price them- what ratio? which is (should be) costing more?

2 Related Work

In this section, we present a comprehensive survey of pricing mechanisms that are proposed in recent years. Appendices contain more detailed summaries of some of these papers [2, 14, 15, 16, 17]. The pricing structures in current data networks are ad hoc, complex and based on unrealistically simplified assumptions (about underlining network, charging/pricing policies, etc.).

Pricing in telephone network has been better studied. However, it only offers a single quality of service, and the costs of providing service are better known where setting prices to recover costs (and some profit) is fairly simple [21].

There is some literature on how to price a network that offers heterogeneous services. Starting with Cocchi et al. [1-3], there are a number of authors who have worked on the issue of pricing. Following is a brief summary of most important (in our view) of these studies.

2.1 A Survey of Pricing Schemes

As an example of current pricing strategies, in today’s Internet, regional networks are supported by governmental funding, and participating organizations are responsible for local area network costs. Hence, end users typically pay only for their local area network costs and the membership fees of the regional network. These costs are usage insensitive and relatively small. This strategy is simple and works reasonably well- as the bulk of the costs are borne by government and respective organizations with little contribution from the users.

Cocchi et al, [1-3] have studied pricing in reservation-less network (a single network, which provides multiple (four to be precise) services at different performance levels). Each user is characterized by a utility function and can request a ToS by setting bits in its packets. They gave a very impressive example showing that in comparison with flat-rate pricing for all services where it is reported that quality sensitive pricing more efficient than flat-rate pricing scheme. A price schedule based on performance objectives can enable every customer to derive a higher surplus from the service, and at the same time, generate greater profits for the service provider. Since resources cannot be reserved, users may suffer QoS degradation. The only way to measure “goodness” of a pricing scheme in this case is to measure the net from the network, which means that user utility functions have to be assumed. Because of the difficulty in determining a valid user utility function, this is undesirable. Therefore, this scheme (otherwise an important and elegant work) does not seem appropriate for reservation-oriented environments, where QoS guarantees can be made. Also, the network is based on simple non-congested network setups. Moreover, they set no condition about the distribution of the total utility (with some users having very high and others very low utilities can result in maximal value) and therefore, the system has the potential of being unfair. Moreover, in the optimal pricing models, the fact that different applications may have different performance objectives was usually not considered. Finally, this method is computationally expensive as it is based on fine-grain per byte accounting. A complete summary of [2] is given in Appendix B.

Since this work, a number of authors proposed pricing schemes with different emphasis and objectives that are analyzed in the following paragraphs.

Parris et al, [4] also present a pricing scheme, which study a host of important issues in connection with pricing (peak/off-peak traffic, elasticity of users’ demand, call blocking, etc). However, they assume that the network provides only two classes of services (both requiring some fixed amount bandwidth to be reserved for the connection). This scheme only considers one resource of real network, namely link bandwidth. In reality, any
successful scheme must consider a host of resources. Also, a single node network model is over-simplification of any pricing model.

Edell et al., [22] proposed pricing and billing at the TCP level where the scheme addresses the issue of willingness and ability to pay. There is an overhead associated with message passing in this schemes which is large. Also, the scheme misses the potentially considerable amount of non-TCP traffic where resource intensive real-time applications may not use TCP.

In [23], Stahl et al., has looked at network computing with priority classes. The prices charged for services vary dynamically through an auction-like process. The scheme is computationally expensive and difficult to scale.

Varian and Mackie-Mason [24], cite the economic theory which states that given a congested resource, the price one pays to send a message (one's utility) should reflect the loss of utility inflicted on other users whose messages are waiting. Their model contains two price elements: connection charges and usage charges. Users bid the maximum price they are willing to pay to send their messages/packet. The highest bid messages are sent first. In any given time interval, the lowest bid message that gets sent sets the price for all messages sent. The overhead in bidding each individual message/packet is considerable. Also, the scheme treats the Internet as one entity and does not talk about distributing the revenue for all the involved parties when a message/packet is sent. Moreover, a packet may be blocked indefinitely in some intermediate node after successfully winning the bids in all previous nodes in its route while failing to win the bid in the given node.

Most economic schemes such as the above ones, do not take into account the technical details of communications. These methods are difficult to scale.

Dewan, Whang and Mendelson's work [29, 30] assumes that the consumer's willingness to pay depends only on expected mean delay, and Mackie-Mason [31] assumes that consumers do not care about delay--only whether or not their packets are eventually transmitted. There is no way, for example, to accommodate a service that would impose a maximum delay limit. These formulations also do not consider the case of heterogeneous data rate and burstiness. Consequently, pricing policies developed in these studies cannot be applied in integrated-services networks in which services differ from each other in terms of performance objectives and traffic pattern (data rate and burstiness). Some of these factors are considered in the paper by Cocchi et al. [1-3], however, as mentioned above it does not discuss procedures for designing an optimal pricing scheme.

Parenteau et al, [13] considers the problems of Internet congestion and delays from economic point of view. They propose a solution aligning the economic incentives of users and providers. The paper is not covering a host of problems and properties of pricing.

and how they are related to Pareto and system optimality. Appendix A has a complete summary of a follow-up paper [17] where these and other terms are defined. The method assumes fixed-size packets and proceeds with packet loss rate as the sole QoS indicator. The authors use utility function (as a theoretical tool) to analyze and reason about network and application behavior. Their system suits best the best-effort traffic environments where the network is not required to provide stringent QoS guarantees which can be achieved, currently, by resource reservation. The contribution of the work is in the area of theoretical aspects of the network properties where they have shown many results including:

They have given a characterization for Nash equilibria and conditions for their existence and also studied their structure relating to Pareto and system optima. Nash equilibria—corresponding to a stable fixed points in non-cooperative games—need not be Pareto nor system optimal; in fact Nash equilibria need not even exist.

However, their study is based on simple networks and it is not clear how well the results will hold in large networks such as the Internet. Also, they do not touch on the most crucial issue of practical implementation, as their study is simulation based. The issue of pricing is not incorporated into their game-theoretic analysis. The work is more of QoS provisioning— in a limited form— than pricing.

Similar work is reported by Korilis [25] (Appendix D is the summary of a paper by this group [14]), Shenker [9-12, 27] and Orda et al. [26]. The flow or congestion control model of Korilis et al., the QoS provisioning work by Shenker et al. and the non-cooperative routing game of Orda et al. are all important research efforts in their own place (and of back), their relevance to pricing is of limited interests.

Fishburn et al. [15], presented a model with two types of demands for data transport and three network configurations. The main conclusion they drew is that that the provision of uniformly high QoS to all traffic is the best in the long run. This is an expensive solution and since the traffic demand increases as the bandwidth (and other resources) improves, it is an error to argue that over-provisioning the capacity is the solution for achieving high network performance as this cycle would be impossible to break. A complete summary of this paper is included in Appendix E.

Ferrari D. in his paper [14] presented a pricing model that is also based on simple networks. The contribution of the work is its analysis of the properties of a pricing scheme both from users and providers’ point of view. Charging formulae were given for real and non-real traffics. The paper is limited as it does not present a global treatment of the pricing problem nor it presents any practical implementation of the scheme in a real network. Appendix C presents a complete summary of this paper.
3 Outline of the Proposed Pricing Scheme

In this section, our proposed scheme is explained in details. It is a simple method yet covers all the important aspects of a practical pricing framework. It is scalable framework for QoS capable internetwork (e.g., future Internet “FInternet” consisting of a collection of domains) supporting real time, adjustable real time, and non-real time traffic. We consider Bandwidth, Delay, Delay Jitter, and Packet Loss Probability as QoS parameters in this work. The scheme, belonging to usage-based (and QoS dependent) methods, is independent of the underlining network and the mechanism for QoS provisioning, which can be deployed in any QoS capable (with least being best-effort) environment. The proposed pricing scheme is credit-based ensuring the fairness (from users’ point of view), comprehensibility, controllability, predictability, and stability. On the other hand, it provides means for the network providers to ensure- with high probability- cost (plus profit) recovery, competitiveness of prices, and encouragement (or discouragement) of client behaviors that will enhance (or degrade) the network’s efficiency (by appropriate charging mechanisms and suitable incentives- see section 3.4).

We believe that the implementation and usage costs of the proposed framework are low. The method covers all other important aspects of a complete solution as well.

3.1 Introduction

The main problem in pricing of Internet services is to find a metric that fairly represents the relative merit of each service. In this document we present a technique that not only includes the definition of such a metric but also covers the practical issue of implementation in the existing Internet.

In conventional QoS model the required bandwidth or the BitRate has no rigid relation with the PacketRate. However all the forwarding takes place based on packets. Therefore we argue that PacketRate should be used instead of BitRate as the QoS Parameter. All other QoS parameters also consider a per packet behavior (e.g.: Delay, Jitter and LossProbability). Hence Considering PacketRate instead of BitRate will help relate all the QoS Parameters.

The Packet Size is fixed for the constant bit rate service while it is variable for VBR service. The applications need to declare the mean packet size (and possibly the S.D. which would be helpful in QoS-Routing mechanism) in advance.

3.1.1 Assumptions and Non-assumptions

In this section the assumptions made in this work are stated. Also, for comparisons with others work in this area, we state the relaxations of some assumptions made by researchers whose work are presented in our survey section. Other assumptions will be stated in different sections as appropriate.
The first assumption is that the underlying network is QoS capable, and that the QoS is negotiated at connection setup. Based on the admission control scheme, such request is accepted or rejected. Users can make requests with degraded QoS parameters so that their requests be accepted in case their current request is rejected (in case of network overload). Alternatively, the network can suggest to users about QoS parameters that can be accepted. Users can also renegotiate, during the connections, their QoS parameters [32, 33]. We assume that users desiring best-effort level services can present their unrestricted QoS requirements and the network will accept the connection and such users be treated according to the network policies laid down for this category of service. We do not state (and for that matter do not care about) how this capability is provided or how the negotiated QoS parameters are guaranteed. This makes our pricing framework independent of network infrastructure and hence can be deployed in any such networks that will emerge as “Future Internet”.

Unlike previous work [1-3], no assumption about network- congested or otherwise- is made in this work. It is the problem of call admission/QoS provisioning. Congested network only affects the rate of call blocking. Besides it results in degradation of the quality of established call. In our scheme as it is credit-based where users will get credits.

We assume that a call will be rejected if the requested delay (one of QoS parameter) is less than the propagation delay- where we assume that it (i.e., propagation delay) is standardized delay metric between two routers independent of path taken. This is not part of our pricing scheme, rather should be handled by call setup mechanism.

In order to maximize profit (not revenue) the routing mechanism should be smart enough to select the route that incurs minimum cost yet promising the demanded QoS.

It is assumed that at the source a new packet is ready to send every 1/PacketRate second. The reason being that in case the source is not ready, the current approach will not charge. We are exploring several approached to charge for this slow behavior of the sender.

In this work, coarse pricing (per packet rather than per byte) is used which substantially reduces the overhead.

A service (like AT&T customer service) will be available to the users to find out the price of communication between two IPs at different time of day given the QoS parameters.

3.2 Type Of Services

We categorize the services as the following broad Type of Services, each having a different set of QoS parameters. Within each type of service different Qualities of Services are available by adjusting the QoS parameters.
1) **Real Time (RT) Service: (CBR, VBR)**
Services, which have a critical upper bound on the time at which packets are arrived at the destination, are termed as Real Time Services. Any packet that is delayed beyond this pre-specified time is of no use to the users.
Such services specify AcceptablePacketDelay as one of the QoS parameters. Other parameters include MeanPacketSize, PacketRate, AcceptableDelayJitter and AcceptablePacketLossProbability.
Examples of such service include Telephony, teleconferencing and many others.

2) **Adjustable Real Time (A-RT) Service: (A-CBR, A-VBR)**
We introduce this new type of service after observing the fact that half-duplex multimedia unicasting/multicasting applications like real audio/video do not really need a real-time connection in the true sense. These applications actually need a little less-demanding type of service which we call AdjustableRealTime service. We elaborate this mode of service by using the following example:
Suppose you are watching a pre-recorded movie or a live event (like a cricket-match or President’s speech) and suddenly the video pauses because of the lack of packets perhaps due to congestion in the network. Suppose after 30 seconds you resume receiving the packets. Now you have two options: One option is that like a real-time service you discard all packets which are delayed more than their ETA(ExpectedTimeOfArrival), and resume playing with the ones which are then received in-time. This will produce discontinuity by skipping a portion of the video/speech. The other and more desirable option is that you play all the packets without skipping at the normal video rate. This will result in a continuous video/speech although with an increased delay of 30 seconds but you won’t be missing any part of the video/speech.

If the later option is the desired mode of service than clearly it is not the RealTime mode of service as the ETA of all the packets after one packet is delayed is shifted. We call such type of service as Adjustable-RT. As a QoS requirement, one should be able to define some kind of upper bound on the frequency and duration of video pauses. We took care of this by including the PerPacketDelayIncrement Parameter in the set of QoS parameters.

3) **Non Real Time or Best Effort Service. (ABR)**
It is the best effort service and can be charged free/flat like today’s Internet.

### 3.3 Pricing of Services

This section talks about pricing different traffic categories. We also derive the canonical prices for packets (with different ToS having QoS parameters).

#### 3.3.1 Pricing for RT Service

PacketLossProbability: For pricing purposes PacketLossProbability directly translates to the bandwidth (i.e., packet rate). A connection having PacketRate of 20 packets/second
with 10% PacketLossProbability can easily be considered as a connection of 18 packets/second PacketRate (only to determine billing price of the required QoS). i.e., this connection will be paying the same per packet amount as the one having 18 packets per seconds with zero loss probability but this connection may end-up paying more if all of its 20 packets arrive at the destination without loss.

**DelayJitter:** We argue that jitter is generally an un-desirable phenomenon. However there are applications that can survive with some limited amount of jitter. Such applications use playout-buffers at the destination to absorb the effect of jitter by increasing the delay. Therefore we translate the jitter parameter into an increase in the delay parameter for pricing purposes. (it is research issue to determine how much delay be increased. By 1 SD, 2 SD or 3 SD).

**PaketRate:** At the call setup phase the user specifies the PacketRate parameter. We assume that the policer doesn’t let the sender sending packets at higher rate (by using leaky-bucket). If the sender cannot keep up with the PacketRate and fails to send next packet within (1/Packet Rate) time interval then we call it a miss and charge for a packet of MeanPacketSize (because resources are assumed to be reserved for this connection).

**AcceptablePacketDelay:** Since end-to-end packet delay consists of two factors – Viz. the Queueing delay and the Propagation delay. If the AcceptablePacketDelay is less than the propagation delay then the network simply cannot provide the required quality of connection and hence should deny service at the call setup phase. Otherwise the difference AcceptableDelay-PropagationDelay gives the AcceptableQueueingDelay. It is fair to charge higher amount for connections that need lower QueueingDelays.

**Distance between Hosts:** It is logical to assumed that under similar network conditions, connections between distantly located hosts are subject to more queuing delays compared to closely located hosts in terms of hop count. Also large distance connections consume more resources. Therefore we propose to charge a higher price for large distance connections compared to the low distance connections.

Here we have two options (a research issue) one is to use the actual distance covered by each packet by looking at its TTL field. The other is to use a standardized distances (can be averaged over time) between hosts independent of the path taken. In the first case the cost of each packet may vary depending upon the path it takes and the user have no idea that how much he would be charged for that packet. In the other case we can give the user a fixed price at the beginning of the connection, however, the drawback is to maintain a table containing standard distances between all possibly communicating hosts (by using heuristics or other alternatives- a research issue).

**PerPacket Accounting:**
Now, given the PacketRate and the PacketDelay we can easily tell the ETA of each packet at the destination. We claim that any Packet that arrives within its ETA, satisfies
the required QoS and therefore should be charged for. While any packet that arrives after ETA, does not satisfy the required QoS and should be given some kind of credit.

Each packet that arrives in time will be valued at:

\[\text{Canonical Price} = \text{PacketSize} \times \text{Distance} / \text{AcceptableQueueingDelay}.\]

Any packet that is missed by the sender will be also valued at:

\[\text{Canonical Price} = \text{MeanPacketSize} \times \text{Distance} / \text{AcceptableQueueingDelay}.\]

The packets that are delayed beyond their ETA will be credited at:

\[\text{Canonical Price} = \text{PacketSize} \times \text{Distance} / \text{AcceptableQueueingDelay}.\]

Finally, the packets that are dropped—due to catastrophic failure or congestion—will be credited at:

\[\text{Canonical Price} = \text{MeanPacketSize} \times \text{Distance} / \text{AcceptableQueueingDelay}.\]

The net price that is charged (credited) for each packet is determined by multiplying the net value by a charging coefficient (constant) that reflects any policy issues (for example, recovering costs with some profit). Possibly we can add a small fixed cost factor (equal to the cost of non-realtime packets) to it. This issue is discussed later in this document.

Note that zero AcceptableQueueingDelay is deemed unaffordable. Reason being that since everyone cannot be given zero delay—for some other user has to be delayed (resulting in unfairness) or that over-provisioned network is available (economically unfeasible). The economic law requires us to increase the costs of scarce resource so that people are discouraged to buy.

### 3.3.2 Pricing for A-RT Service

The only difference in RT and A-RT is that ETA of subsequent packets is shifted by PerPacketDelayIncrement factor. This results in an increased AcceptableQueueingDelay for each packet and hence using the same Value-formulas as of RT service the value of subsequent packets goes down with the life of connection. Hence A-RT connection will end up paying less compared to the similar RT connection even if the same charging coefficient is used.

### 3.4 Charging Methods

In this section a number of charging related issues are discussed. Firstly, we assume that in addition to usage charges, the charging formulae given below have terms for fixed access charges and connection charges. Fixed access charges are needed for capacity expansion while connection charges are due to reservation of resources along the path to the destination of a connection. Connection charges are applied when a connection is
established. Access charges may be eliminated from these formulae in case such charges are not applied per connection (rather being applied per month for example).

As mentioned earlier, there are a number of desirable properties (both from users’ perspective and network providers point of view) which such formulae should posses. It is “work in progress” to establish and show that our charging formulae have such properties.

\[
C_{\text{traffic type}}(\sigma_{\text{QoS}}) = \alpha_{\text{traffic type}}^* P(\sigma_{\text{QoS}}) + \beta^* R(\sigma_{\text{QoS}}) + \gamma
\]

- **\( C_{\text{traffic type}}(\sigma_{\text{QoS}}) \)**: Connection cost for traffic type (such as real-time,…,etc.) traffic.
- **\( P(\sigma_{\text{QoS}}) \)**: Canonical Price calculated by the given pricing mechanism- see section 3.3.
- **\( R(\sigma_{\text{QoS}}) \)**: Resource reservation charge (may include connection establishment charge)
- **\( \sigma_{\text{QoS}} \)**: QoS parameters (Bandwidth, Delay, Delay Jitter, …)
- **\( \alpha_{\text{traffic type}} \)**: Coefficient for usage charges – for a traffic type (real-time, adjustable-real-time, best-effort)
- **\( \beta \)**: Coefficient for reservation
- **\( \gamma \)**: Fixed access charge

In this work, since our pricing framework is independent of underlying network infrastructure, we can only apply a fixed charge for resource reservation (it is rather connection establishment charge). This can be justified as resource consumption is captured in usage charges indirectly. It is “work in progress” to determine whether this component should be charged for those connection requests which are accepted or it should be charged for any connection request (accepted or rejected so that users will not even more congest the congested network).

Also, we assume that the access charge component (in the above charging formula) is not per connection (the access providing entity can charge this per month and during the revenue distribution this amount is paid to it by the billing mechanism).

Therefore the following are the formulae for different type of service that we consider in this work.

\[
C_{rt}(\sigma_{\text{QoS}}) = \alpha_{rt}^* P(\sigma_{\text{QoS}}) + R
\]

\[
C_{art}(\sigma_{\text{QoS}}) = \alpha_{art}^* P(\sigma_{\text{QoS}}) + R
\]

\[
C_{be}(\sigma_{\text{QoS}}) = \alpha_{be}^* P(\sigma_{\text{QoS}})
\]

\[
C_{rt/art/be}(\sigma_{\text{QoS}}) \quad \text{Connection cost for real-time, adjustable real-time and best-effort respectively.}
\]

- **\( P(\sigma_{\text{QoS}}) \)**: Canonical Price as described in section 2.2
**R** : Fixed connection establishment charge

**σ_{QoS}** : QoS parameters

**α_{rt}** : Coefficient for usage charges – real-time traffic

**α_{art}** : Coefficient for usage charges – non-real-time traffic

**α_{be}** : Coefficient for usage charges – best-effort traffic

Note that it makes sense to assume that $\alpha_{rt}$ is higher in value than $\alpha_{art}$ (with $\alpha_{be}$ being the lowest of all three).

We assume these parameters are defined by pricing agents- discussed below- and not individual routers that are involved in the connection.

Network providers can make these coefficients sensitive to different time periods (peak times or non-peak times, weekends/special occasions; borrowing from telephone). They can also promote (in order to increase users) by manipulating these coefficients in favor of users and/or eliminate connection establishment charges. Since network performance degradation may become apparent in extreme cases, the swing should be carefully designed. Also, in order to increase revenue, these coefficients can be made high. However, studies show that in an effort to generate higher revenues, users may be deterred to use the networks and hence causing a sharp decline in revenue generation [13, 4, 14]. It is “work in progress” to estimate these coefficients in order to successfully keep this balance.

Note that it is possible to ensure a connection (borrowing from postal service which makes sense only with real-time or adjustable-real-time traffic) where the users expect great losses in case any packets are lost. In case of unsatisfactory service, the users can be credited higher while it will be cost higher when the negotiated QoS is satisfied.

Another issue is whether to charge sender (similar to call collect) or receiver. It is “work in progress” to identify means to support this phenomenon. A bit in header or specific ports can be used to achieve this.

Multicasting is the same as far as charging is concerned. Revenue distribution is complex however which is discussed below.

### 3.5 Abstract Formulation

The formulation is based on Nash equilibria (see Appendix A, [17-19]), utility functions and some general observations about the Internet and its selfish users (same as [3, 16-19]) who respond to incentives. It will become apparent that our scheme is not exactly based on these assumptions, as the selfishness of the users are somehow controlled by call setup and policing mechanism. However, in order to clarify different aspects of our scheme, we will base our formulation on the assumptions that a user is selfish and would like to use the network as much as s/he can and no user will unilaterally reduce their usage.
One point worth noticing is that (in a network) Nash equilibrium is exception rather than being the norm. This is why, in this work, the emphasis is not on the theoretical aspects of pricing framework rather the focus was on practical issues.

Unlike the work by Cocchi et al. [1-3], the utility functions should for networks with QoS (rather than just ToS) provisioning which implying that they should be continuous functions. These utility functions are used to capture users’ overall satisfactions. Kihong et al. [17-19] also studied the utility functions in their work.

The economic theory that basically states that given a congested resource, the price one pays to send a message (i.e., his/her utility) should reflect the loss of utility inflicted on other users whose messages did not get the same treatment. Our abstract formulation is revolving around this phenomenon. The utility function needs to show the performance one’s application gets and the price they pay for.

In general the utility function takes the following form:

$$U(P(\sigma\text{QoS})) = V_{\text{traffictype}}(P(\sigma\text{QoS})) - C_{\text{traffictype}}(\sigma\text{QoS})$$

Where $V_{\text{traffictype}}(P(\sigma\text{QoS}))$ is the apparent degradation for (in favor of other users) to a user while $C_{\text{traffictype}}(\sigma\text{QoS})$ is the price one is charged. The canonical price $P(\sigma\text{QoS}) \to +\mathbb{R}$ (is a mapping from $\sigma\text{QoS}$ (the set of QoS parameter) to positive real number $+\mathbb{R}$). Also, $U(P(\sigma\text{QoS})) \to +\mathbb{R}$ is a mapping from a positive real number to another positive real number showing the worth utility perceived by a user (for a usage request with quality given $\sigma\text{QoS}$ by and for which s/he paid $C_{\text{traffictype}}(\sigma\text{QoS})$).

The challenge we face is to formulate suitable $V_{\text{traffictype}}$ for real-time, adjustable real-time, and non-real-time traffic. It is “work in progress” to define such utility functions (which are multi-variable continuous functions) for the abstract formulation of our Pricing Framework. Conditions for network operating at peak efficiency will be derived afterwards. Here, network utilization is defined as ratio between the total amount of bandwidth reserved (for all connections) and total bandwidth of network.

### 3.6 Metering, Accounting, Billing, Advertisements and Revenue Distribution

Metering is another important and problematic issue that is needed to be part of Pricing Framework. What is needed here is to log traffic by some pricing agents. In order for our Pricing Framework to be scalable, we choose endpoints (edge routers) to be the place for metering where dedicated agents receive duplicate headers from corresponding edge routers (to which receiving hosts are attached; see Figure 1). They are only dummy hosts with no overhead to the network. They discard the messages after logging the header information. In addition, these agents can be used to perform other activities such as coefficients estimations, providing charging information, acting as call admission agent (not part of our Pricing Framework), deciding queue size (not part of our Pricing
Framework) for network providers and a host of other activities. It can also bill electronically and receive payments via network too.

In this work, the billing is done per connection and accounting is done on specific time periods (of all completed connections).
Distribution of revenue –like accounting- is done per connections where actual billing will happen in predetermined time periods for all completed connections. As we mentioned earlier, every router can set its charging coefficients which will be reflected at charging the user (in this work they are assumed to be uniform) but revenue must distribute according (proportional) to delay in each router. The routers causing the packets to be delayed must have favored packets from other connection and hence it is fair to receive less revenue. It is “work in progress” to hammer the details of this proportionality. Note that the routers only set the coefficients in specific time periods independent of any particular connection and that the pricing agents at the endpoints only do the pricing related work.

Adjustable real-time can be useful for sponsored multi-media service where a network provider, in return for some reduce (possibly free) cost, shows (for example by using a tool like realplayer) advertisements to the receiver during pauses.

In multicasting the revenue distribution is more complex and which, in addition to being proportional to the delay experienced at routers, is also proportional to the number of receivers depending on them. It is “work in progress” to devise the details of this issue.

4 Implementation and Deployment Issues

Our Pricing Framework can be implemented in a complete decentralized manner. The implementation of a simplified version of our Pricing Framework is in progress. Since we assume that the underlying network is QoS capable, we need to adopt/simulate such a network.

It is research issue to study the feasibility and implement the QoS capable network, which incorporates the new and practical traffic classes (such as adjustable-real-time) with a changed set of QoS parameters that were introduced in this work.

It is our goal to test the scheme on a number of networks with different QoS provisioning methodologies and study the results.

A handshake is needed between the QoS provisioning mechanism of the network and the pricing scheme so that the later know about the QoS parameters that the network guaranteed to the user. Also, users can enquire about their accounts. In case, the QoS provisioning mechanism supports renegotiations [32, 33], this handshake needs to take place every time such renegotiation is accepted by the network- so that the canonical price (and finally charging) is calculated accordingly. Pricing agent can be used to implement a renegotiation mechanism such as the one reported in [32].

As we mentioned earlier, we assume that all the routers set the same charging coefficients and therefore they receive revenues proportional to the delays caused by them. Note that the routers only set the coefficients in specific time periods independent of any particular connection and that the pricing agents at the endpoints only do the pricing related work. However, it is also our goal to simulate networks where routers set different charging
coefficients (perhaps by using MPLS-like methods [34]). However, this will complicate the distribution of revenue too.

5 Goals, Objectives and Results

Following is a partial (and growing) list of goals and objectives, which we desire to achieve, and results which we want to observe.

All the “work in progress” items (indicated throughout this document) are subject to modification.

As mentioned earlier all the pricing related work is done either at edge routers or dummy hosts attached to edges routers (called pricing agents) and therefore it is easy to scale. Hence it is desired to simulate large networks. The approach will be implemented in completely decentralized manner where pricing related activities to be performed at endpoints.

Abstract formulation of our Pricing Framework is to be pursued. For this purpose, suitable utility functions to be devised.

We would like to see the effects of pricing on network utilization and performance.

We would like to see the effects of congestion control schemes coupled with pricing versus only congestion control schemes.

We would like to see the effects of different charging coefficients \((\alpha_{rt}, \alpha_{art}, \alpha_{re}, \text{ and } R)\), on network revenue generation.

Comparison of different charging schemes is to be studied.

We would like to see the effects of connections establishment charging in pricing.

For fixed revenue, \(D\), we would like to study which charging coefficients (with peak/non-peak, ….etc.) will generate \(D\).

We would like study network throughput with/without pricing versus revenue (and versus users’ satisfaction).
6 References


7 Appendices

A number of papers have been summarized with greater detail and the results are given below.

Appendix A

An architecture for Noncooperative QoS Provision in Many-Switch Systems
by Shaogang Chen & Kihonh Park

1 Introduction

With changing users' QoS requirements, it is important to organize today's "best-effort" bandwidth, for example that of Internet, into stratified services with graded QoS properties such that these requirements are effectively met. This is because, in general, transporting application traffic over reserved channels incurs a high cost. This is particularly useful for applications that possess diverse but flexible QoS requirement. It will be overkill to transport such traffic over reserved channels while sending it over "best-effort" service would be equally unsatisfactory.

Based on this the authors suggest that a dual architecture capable of supporting reserved and stratified best-effort service is needed, which, helps amortize the cost of inefficiencies due to overprovisioned resources of guaranteed traffic.

The authors' earlier included proposing a game-theoretic analysis of the multi-service QoS provision system for noncooperative single-switch network systems. They were able to show when a Nash equilibrium exist and under what conditions they are system optimal. They generalized-by reduction based on notion of selfishness- to many-switch systems. Following are some definitions which are used in the paper.

1.1 Definitions:

A configuration is a Nash equilibrium if each user can not improve his individual lot through unilateral actions affecting his traffic allocations. If every user finds himself in this situation then the system is in stable rest point.

A configuration is a Pareto optimum if in order to improve the lot of some user/player, the lot of others must be sacrificed.

A configuration is system optimal if the sum of the individual lots is maximized.

2. The Core of the Work:

This paper focuses on the distributed QoS control problem associated with noncooperative many-switch systems. It formulated the many-switch QoS assignment problem as a constrained optimization problem- known to be NP-hard even for single-
switch- and transforming this constrained optimization problem to an unconstrained form using the framework of Lagrangian multipliers (providing an approximation procedure) which can be implemented in decentralized manner. Distributed control is provided by:

1. local optimization at every switch by choosing a service class that meets the local QoS requirement and which minimizes resource usage
2. global optimization via end-to-end feedback loop that adjusts the Lagrangian multipliers- shared across all switches- to satisfy the target end-to-end QoS at minimum cost.

The paper shows the efficacy of the method by simulating a WAN environment based on a vBNS-like topology with multiple traffic flows processing diverse QoS requirements. They also compare the performance of their method against a reservation scheme that a priori allocates fixed service classes based on assumed knowledge of all traffic flows as well as FIFO packet scheduling and random service class assignment. The technique is tested using various problem instances (classified as "easy", "intermediate" & "difficult"). It has been shown that their technique performs well.

2.1 Network Model:
A network setup where a set of routers and stations are connected via some topology. The routers implement GPS packet scheduling where packets labeled by their service class number receive service commensurate with the resources allocated for that service class and the traffic impinging on that service class. QoS control is exercised on an aggregate flow basis. Other things equal, the larger the service weights the better the QoS.

Assuming fixed routes, the end-to-end QoS experienced by an application flow is determined by service levels received at each of the routers along a path. There is a calculus for computing end-to-end QoS in terms of the QoS rendered locally at each of the switches. Assuming there are m service classes at every switch k (1 <= k <= r), then to flow i there correspond r choice variable Tki (1 <= Tik <= m) which determine which service class application flow i is assigned to at hop k. In this work, it is assumed routing subsystem is separate and available.

2.2 User Model:
Assume n applications/users where each user I (1 <= i <= n) has a traffic demand given by its mean rate Li. A user's QoS requirement can be represented by a utility function Ui which captures the "satisfaction" experienced by user i when receiving a certain QoS.

Fixing a switch k(1 <= k <= r), user i's flow Lik (note Li1 = Li and Li1 => Li2 => ... => Lir) can be chosen by user to be assigned to one or more of the m different service classes at k. This assignment is represented by assignment vector Vik = [ Li1k, Li2k, ..., Lijk]t where Lijk => 0 and sum (Lijk) = Lik for any j. Thus, the aggregate flow entering into a service class j is given by Aik = sum (Lijk) for all flow i.
Note, that selfishness will mean that each application I will try to take actions- setting Ti (user i's service class assignment vector)- so as to maximize the individual utility Ui.

2.3 Noncooperative QoS provision Game:

Many-switch Network Game:
The paper states that the end-to-end QoS received by a user I, Xi, is a function of T (where T = [T1, T2, ..., Tn]t- Xi = Xi(T)-, and given i'th traffic demand Li, they arrive at the individual utility Ui(Li, Xi), and which they say is consistent with their earlier work (which I did not read yet). And therefore Nash equilibria, Pareto optima, and system optima can be defined with respect to Ui(Li, Xi).

2.4 Network Game with pricing:

Pricing is introduced as a mechanism for monitoring relative resource usage by imposing the relation

if Cak => Cbk then Pak => Pbk, where (1 <= a, b <= m),

where Cjk is the QoS rendered in service class j at switch k and Pjk is its price - the superior the service the higher the price. Other things equal, for GPS switches, Cak => Cbk iff the relative resources consumption per unit flow is higher in service class a than b.

The pricing scheme of this paper is fine-granular in the sense that pricing occurs at every switch on per-flow basis.

The cost or relative resource usage accrued to user i's traffic flow can be defined in a number of ways ranging from "user-friendly ones" to "provider-friendly ones".

One version of the noncooperative many-switch QoS provision game with pricing is given by:
Min Li*sum(P-Lik-k) (minimize for Ti) subject to Xi(T) <= Gi (where Gi is the QoS requirement vector representing user i's bounds on desired end-to-end QoS.

That is the user seeks a minimum cost assignment that satisfies the user's QoS requirement. This problem is NP-hard- can be reduced to a version of multiple choice knapsack- even for a single user and hence approximation solutions must be sought.

The noncooperative QoS provision game is defined by solving the above optimization problem for each user.

The author formulate a method- Lagrangian formulation- which enable transforming this problem into an equivalent unconstrained form which leads to a set of independent optimization problems (one for each switch) coupled only by a set of Lagrangian multipliers. It has a simple interpretation: the larger
the multiplier the more stringent the QoS rendered by the system. This provides a nice distributed environment.

The transformed form is also NP-complete and therefore approximation solutions are suggested. This decentralized approximation procedure consists of a local and a global portion. The local optimization needs a nonlocal information which is provided at the header of the packets (the header has been modified to suite this method).

3 Simulations:

ns network simulator was modified. The results for vBNS-based network topology were shown. The traffic configuration included several application flow and background flow. Three service classes (with one reserved for background flows) were studied. Results were compared with single-switch reduction technique and a fixed reservation-based allocation method.

Appendix B


1 Introduction:

This paper studies the role of pricing policies in multiple service class networks. It is argued that technological progress- in software, hardware, protocol standards, …etc- needed for better network performance is not the only important issue in this regard, it is also a function of incentives users encounter when using the network. Hence, this issue of incentives (which can take many forms such as performance, monetary, social, … etc) must be considered. This paper focuses on monetary and performance incentives.

The main point that the paper argues is that some form of graduated prices are required in order for any multiclass service discipline to have the desired effect. They try to establish- through simulation- that it is possible to set the prices so that every user is more satisfied with the combined cost and performance of a network with graduated prices.

The authors argue that by pricing the service classes appropriately, one can offer monetary incentives for reducing the QoS requests. Hence, it is expected that pricing of the various service classes to be a vehicle commonly used to encourage users to make reasonable choices. For some users the performance penalty received for requesting a less-than-optimal service class is offset by the reduced price of the service. For the other users the monetary penalty incurred by using the more expensive, higher quality service classes is offset by the improved performance they receive.
This way, it is being stressed, the prices allow us to spread the benefits of multiple service classes around to all users, rather than just having these benefits remain exclusively with users who are performance sensitive.

The authors consider a very simple multiclass service discipline and compare two pricing policies: flat per-byte fees vs. graduated fees for different priority classes.

By measuring user satisfaction as a function of both the cost and QoS received, and studying a simple network configuration with several different applications using standard transport-layer protocols, the study shows that every simulated user is more satisfied with the graduated pricing scheme.

2 Core of the Work:

After observing the characteristics of today’s Internet (1-limited bandwidth, 2-restricted access, 3-single type of service (TOS) & 4-no usage fees), the paper prophecies future of Internet to be having different characteristics (1-backbone would be upgraded, 2-this combined with the increase and wide spread of PCs much better public accessibility, 3-future traffic mechanism is likely to be more sophisticated than single TOS, 4-implementation of usage fees).

As mentioned above the paper investigating-in simple network model- the interplay between the monetary incentives provided by the prices and the performance incentives provided by the service classes. The authors try to answer affirmatively the following question: can we set prices in such a way that the performance penalty received for requesting a less-than-optimal service class is offset by the reduced price of the class, while at the same time not making optimal service classes so expensive that even performance-sensitive users do not use them?

2.1 Network model:

In the simple model investigated, users are sensitive to both the QoS they receive and the price they have to pay for that service. Authors model this sensitivity-as is common practice in economics- through a combined “utility function” which describe user’s level of satisfaction with the combined network performance and cost.

Simple utility functions representing users of several different applications (Voice, Telnet, FTP, Email, … etc) are constructed.

The emphasis is not on complex but simple network to provide a simple example that illustrate the point.
2.2 Multiclass Service Discipline

Each packets have two flags: priority service flag & no-drop flag. Hence, there are four service classes. Preference is given to those packets with priority flags on. Within a priority class, the service is provided in a FIFO order. The packets- in case of congestion- is discarded in the following order: both priorities off, service priority flag on & no-drop priority flag off, service priority flag off & no-drop priority flag on, and both priorities on.

2.3 Pricing Scheme

Two pricing policies are considered: flat per-byte fees and graduated fees for different priority classes. The flat rate is denoted by \( P_{\text{flat}} \) and \( P_{0,0} P_{1,0}, P_{0,1} & P_{1,1} \) are the prices corresponding to four service classes (first bit is the priority flag & second bit is no-drop flag). The following relation holds trivially \( P_{1,1} < (P_{1,0} ; P_{0,1}) < P_{0,0} \). A base price of \( P_{\text{priority}} \) is set for each packet and an additional for each flag set. Hence, \( P_{0,0} = P_{\text{priority}}, P_{1,0} = P_{0,1} = 2P_{\text{priority}} & P_{1,1} = 3P_{\text{priority}}. \)

In order to compare these two policies, the authors require that both schemes recover the same net revenue, which is referred as \( D \). This means to choose \( P_{\text{flat}} \) and \( P_{\text{priority}} \) so that total revenue is \( D \).

2.4 Network Configuration, User Model & Applications

A simple network where two host are connected to two gateways by means of 10Mbps Ethernets, respectively and where the two gateways are connected via a bottleneck link is studied.

Users are represented by network application which send data from host-1 to host-2. They care about the cost of running an application, denoted by \( C \). They also care about the performance of their application-which is a function of network performance. Let \( V \) denote the performance degradation apparent to users due network performance- the higher \( V \) is, the worst the application is performing. For each application \( V_{\text{application}} \) is defined. \( V \) measures the performance sensitivity as an application.

Four applications namely Email, FTP, Telnet & Voice are considered. Also, \( V_{\text{email}}, V_{\text{ftp}}, V_{\text{telnet}}, V_{\text{voice}} \) are defined as follows:
\( V_{\text{email}} = 0.1(\text{avg message delay, in sec}) + (\% \text{ of messages not deliv in loose delay bounds}). \)
\( V_{\text{ftp}} = (\text{avg file transfer time, in sec}). \)
\( V_{\text{telnet}} = (\text{avg packet round trip time, in sec}). \)
\( V_{\text{voice}} = (\% \text{ packet not obeying tight delay bound}). \)
2.5 Utility Functions

The overall level of satisfaction of an user is a function of both the application performance and the cost of running that application. The authors model as a utility function $U$ as $U = -V-C$ for this study. Higher $U$ mean higher level of satisfaction.

2.6 User Behavior

How user react to different incentives are discussed. The authors assume that total traffic generated by users are independent of the price. Users try to maximize their overall level of satisfaction $U$. If only performance incentives are present, then all users would set both flags on. If only monetary incentives are present the users would set both flags off. If both incentives are present, the situation is less obvious.

In flat pricing scheme, the monetary incentives are irrelevant, all users set both flags on. In priority pricing scheme, it is assumed that users only set a priority flag if it can improves the network performance in a way which is relevant to their application. So, in case of Email both flags are off, in case of Telnet both are on, in case of FTP service priority off and no-drop is set on, and in case of Voice service priority on and no-drop is set off.

3 Simulation & Result

Email, Telnet and FTP use TCP and Voice use UDP-given the strict delay constraints, retransmission is not useful.

Users repeatedly requests service from their applications and the performance of applications are averaged over all such instances. Each request is characterized by size $s$ and interval $t$ (from last invocation of the application) which are modeled using random processes.

For Email and FTP the max packet size of request is 500 bytes, for Telnet the packet size is 50 bytes, and 180 bytes packets are used for Voice. The strict delay bound in Voice is 200ms and the loose delay bound Email is 5 minutes.

Two configurations were used to test the priority pricing and flat pricing schemes. In first configuration, there were 2 Voice applications, 4 FTP, 5 Email and 2 Telnet. In the second configuration, there were 2 Voice applications, 3 FTP, 5 Email and 4 Telnet. The bottleneck link in first configuration was 772 Kbps and it was 600 Kbps in the second case. $D$ was $100$ in both cases. $P_{flat} = 2.43 \times 10^{-7}$ and $P_{priority} = 1.28 \times 10^{-7}$ in first configuration and $P_{flat} = 3.05 \times 10^{-7}$ and $P_{priority} = 1.75 \times 10^{-7}$ in the second case. In both cases $P_{0,0}$ was half of that of $P_{flat}$, and both $P_{1,0} & P_{0,1}$ were comparable to $P_{flat}$. Only $P_{1,1}$ is 50% higher than $P_{flat}$.
The main result drawn is that every simulated user has higher level of satisfaction with the priority pricing scheme with straightforward case (i.e., \( P_{0,0} = P_{\text{priority}}, P_{1,0} = P_{0,1} = 2P_{\text{priority}} \) & \( P_{1,1} = 3P_{\text{priority}} \)). There are however, many other pricing schemes which can be used and proven to be better than the flat price schemes. The authors claim that there are many such schemes.

With flat price, all simulated applications use the highest quality service class available and the network performance is poor. 82% Voice packets are delivered in first configuration, 65% in second configuration. The Telnet delay bound ranged from 250ms to 1 sec in both cases. In contrast in priority pricing, when users are motivated to choose the appropriate class, 99% Voice packets are delivered with their delay bound and Telnet delays have decreased by an order of magnitude. Improved QoS with additional would further resulted in even better performance gains. Email was the only application which received worse service under priority pricing- performance degraded modestly but the monetary benefits of the reduced price (half of the flat price) outweighed any prospective performance gains.

The contribution of this paper has been in realizing that only technological improvements in hardware and protocol standards can not bring about an overall enhanced network performance in future Internet. However these improvements should be complemented by incentives that users encounter when using the network.

The proposed schemes need to be tested in complex network configurations and with different traffic loads. Also, different pricing schemes be evaluated. Also, other related user preference issues such as demand elasticity (where users reduce offered load in response to increased price) and substitutions (where users switch to other applications such as from Email to Voice) need to be incorporated into realistic models.

### Appendix C


#### 1 Introduction:

In Internet of the future, which is expected to offer real-time as well as non-real-time services, the charging policy must be service-dependent.

The authors start from a list of the properties the policy should have, and derive from it a formula that satisfies most of them. They also briefly discuss the evaluation of the coefficients in the formula and the experiments that could be run for its validation.

#### 2 Core of the Work:

Formulating pricing policy is hard for single-service and even harder for multi-service one. In multi-service case, the problem is deciding different values of the various type-of-service (ToS) and within each type those of the various quality-of-service (QoS). Hence,
the problem is complicated not only by multiple service but also by multiple QoS parameters (the relative impacts on price of a communication of the delay bounds, jitter bounds, the packet loss probability, and so on) and the many values each of them can take.

This paper is working in the context of integrated-services packet-switching internetworks. This network is capable carrying different types of traffic, all of which is either (1) non-real-time (nrt) or (2) real-time (rt) traffic. Service to traffic in (1) is best-effort and traffic in (2) can choose a real-time ToS (and can request performance and reliability guarantees).

The paper assumes that there is only one ToS in (2) and only unicast communication is allowed.

Charges are service-sensitive and also usage-sensitive (being usage-sensitive is important but not compulsory and the terms corresponding to usage-sensitivity can be replaced by ones which are not).

2.1 Properties of Charging policies:
The authors outline the desirable properties of charging policies from the point-of-view of both network managers and users.

From the point of view of network managers, the properties are:
(a). High probability of cost recovery,
(b). Competitiveness of prices,
(c). Encouragement (or discouragement) of client behaviors that will enhance (or degrade) the network’s efficiency,
(d). Low implementation costs,
(e). Low usage costs,

From the point of view of the clients, the properties are:
(f). Comprehensibility,
(g). Controllability,
(h). Predictability,
(i). Stability,
(j). Fairness.

Property (a) suggests reasonable charges so that enough revenue be generated to offset cost- and possibly produce profit- but charges should be competitive with other networks and hence property (b). The charging should reward efficient and penalize inefficient use and therefore property (c). The costs of implementing and using the policy should be low (properties (d) and (e)). Clients want the charging formula to be easy to understand (property (f)), to control the charges (property (g)-by modifying QoS requirements- and predict charges (property (h))- either using past experience based on similar communication or obtaining estimates for new communications. Property (i) demands that prices do not change frequently and property (j) requires that all users be dealt the same way.
3 An Approach to Charging for QoS

An integrated-services packet-switching internetworks must offer connection-oriented services at the internetwork layer for real-time components of its traffic, while connectionless services are desirable for non-real-time ones at the same layer.

During connection establishment, resources are reserved along the path of the connection; these resources will be usable by non-real-time traffic whenever they are not occupied by the real-time connections for which they have been assigned. However, these resources will not be assigned to other real-time connections until this connection is terminated.

It is, therefore, fair to charge for resource reserved for connections. This is called reservation charges and will be higher for larger amounts of resource and proportional to the duration of the reservation, i.e., its lifetime. Obviously, the non-real-time component of traffic pays no reservation charges.

The charging formula has a usage-sensitive for both of these traffic types (real-time and non-real-time) and which is proportional to the number of bytes transported (charging for those received are fairer) to account for bandwidth utilized, and to the number of packets to account for header-processing work by the routers along the path (which should be proportional to number of hops).

4 The Charging Formula

The path of a real-time connection is assumed to be satisfactorily modeled by a cascade of queueing stations-each containing a single server where a router is modeled by a series-two or more- them. Each station has its own queue and represents either a header-processing or packet-transmission resource. This model has three types of resources: buffer space B, computing capacity C and schedulability D (with P(B), P(C) & P(D) being their price per unit respectively), and a connection requests b, c and d of total amount for its lifetime L. Hence the reservation charge for this station is $L^* [bP(B) + cP(C) + dP(D)]$.

The charging formulae therefore are:

$1)$ Charges(rt) = ReservationCharges + UsageCharges(rt) where

$1$’ ReservationCharges = $S^* L^* [bP(B) + cP(C) + dP(D)]$ and

$1$’’ UsageCharges(rt) = $K(rt)^* h^* V$.

$2$) Charges(nrt) = UsageCharges(nrt) = $K(nrt)^* h^* V$.

$L$ = lifetime of the connection,
$S$ = sum of all station on the path,
$K(*)$ = usage charges coefficients,
$h$ = number of hops in path (in rt case) or in “normal” path (in nrt case),
$V$ = total volume in bytes
and the rest as above.
The paper suggests ways to estimate both usage coefficients $K(*)$ and reservation coefficients $P(*)$ for each station. It is not a trivial problem as there are less formulae/equations than the variables. This problem becomes even harder if the stations are significantly different from each other.

5 Conclusions

Charging formulae for two types of traffic were suggested which adhere to the desired properties laid down by the authors to a reasonable extent. They do not satisfy all the properties. The authors indicate that a perfect formula which satisfy all of the properties may not exist.

The main contribution of this work is the excellent discussion about the main properties of any desirable charging formula. The proposed formulae have to be enhanced before it can take a practical role.

Appendix D


1 Introduction:
Quality of service mechanisms allow users identify paths suitable for their desired performance and reserve-efficiently- necessary resources.

The authors propose the use of pricing mechanisms as means to regulate users' decisions in networkwide manner using congestion-based pricing scheme.

They propose a path selection algorithm which satisfy users requirement at minimum cost.

The papers shows that the underlying non-cooperative game among users- as users are selfish- has unique equilibrium, for any choice of cost function.

Then it shows the existence of incentive compatible price function that drive the network into an equilibrium point. Then, the case of sub-optimal paths—the case of multi-objective path optimization— is considered.

2 Core of the Work:

The paper notes that a key issue in the design of broadband networks(which can support multiple application with diverse QoS requirements) is how to provide resources such that to meet the requirement of each connection meeting the goal of network efficiency. QoS routing is a major building block in this regard. But a problem with establishing connection with QoS guarantees is how to reserve resources along the route.
An important problem which is not been sufficiently addressed is that of efficient allocation of resources, namely "rates" or "bandwidth" -not only from single-user point of view, but from network point of view. It is assumed that pricing schemes can be used to enforce this.

What constitute "efficient" resource utilization and how it can be achieved through pricing mechanisms- open problems-are the subject of this paper.

The authors consider a general network with nodal scheduler that belong to the rate-based class- as in specification of Guaranteed Service for IP. Each connection is characterized by its source-destination nodes, maximal packet size, maximal burst, and upper bound on end-to-end delay. The properties of rate-based scheduler allow derivation of an upper bound on end-to-end delay of a connection when it is routed over a given path at a given reserved rate. Reserving a unit of rate over a link incurs a cost- the price of the link. Focusing on congestion pricing, the authors assume that the price of the link is a function of the aggregate rate reserved at the link. Each connection is to minimize the total usage cost while satisfying its end-to-end delay constraint.

The interaction among various connection/users that decide independently on their individual routing strategies can be modeled as a game. Any operating point of the network is a Nash equilibrium of that game, that is, a collection of routing strategies from which no user has the incentive to deviate unilaterally.

Network managers- who determine the link cost functions- desire to drive users to a Nash equilibrium that is efficient from network point of view).

The network efficiency is defined as minimizing a global cost function that quantifies the overall network performance and is the sum of link cost functions. Managers seek pricing scheme which enforce unique Nash equilibrium that minimizes this global function. Any such pricing scheme is called incentive compatible.

This paper show- for a given set of link price functions which obey certain assumptions- do produce a unique Nash equilibrium and address the problem of incentive compatible pricing strategies.

The paper suggests methods for path selection (user problem) and derive global functions (manager problem)- need to minimized- for homogeneous as well as general networks. This formulation is based on the assumptions that there are infinitely many infinitely small users. In addition, the authors assume that the network can accommodate the total offered load (i.e., flows on links are less than the link capacities).

It also notes the exceptions such as the problem of big user (as the existence of such a user violates their general formulations) and derive methods in this case separately.
The paper also talk about incentive compatible pricing strategies for multi-objective path optimization- besides end-to-end delay, the consider jitters also. After showing that such a path selection problem is NP-complete, they point out that there are pseudo-polynomial solutions available and also suggest pricing schemes for some of these approximations.

The paper has simplifying assumptions- such as that flows on links are less than the link capacities. The time complexities are high. Pricing strategies (i.e, congestion based) considered here are "manager/service provider friendly". In congestion based pricing, higher demands- congestion- leads to higher price irrespective of actual QoS rendered. The merit of the schemes is not investigated in action- it is not yet incorporated in any protocol.
Appendix E  
Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet  
by Peter C. Fishburn & Andrew M. Odlyzko

The paper that there is a wide dissatisfaction with the delays and losses in current Internet and hence a need for differential qualities- which will raise the question of differential pricing.

It considers a model with 2 types of demands for data transport, differing in sensitivity to congestion. It studies three network configurations:

1- with separate networks for the 2 types of traffic
2- with single network that provides uniformly high QoS and
3- with single physical network that provides differential QoS.

The author try to show that the provision of uniformly high QoS to all traffic-even though not the least expensive is the best in the long run. They argue that the additional cost is not very large. They base their argument on the fact that in this dynamic environment the traffic is growing but the cost are decreasing. And this additional cost may well be worth paying to attain simplicity of single network treating all traffic the same and have a simple charging mechanism.

The paper is arguing in a way in favor of "Net-head" and not "Bell-head" noting that the objections to "Net-head" of being expensive is no longer a strong case as the environments is changing drastically.

They argue if available capacity doubles each year or so while total cost increases much more slowly , so that price per unit volume decreases rapidly, it make sense to use high uniform QoS for every body and avoid the complexities associated with involved charging policy.

They site the price reduction in fiber optics as an example. They compare the situation here with the microprocessors price/power history and predict that similar would be the case.

They also argue that existing work on QoS does not contain any projections of the degree to which the different proposals for providing QoS will lower network utilization.

Prices for the demand types are based on transfer volume-usage based- and determined by equality between network costs and revenues (Revenue = cost). Uncertainties were accounted for by considering alternate features for demands and costs, including
economies of scale for costs and possible effects of competition and technological advances.

The paper considers two environments for simulation: the first in which costs change only because of potential volume doubling from period to period. The second is reflects not only doubling assumption but also cost reductions due to other factors such as competitions, ..etc. They compare prices, demand satisfactions, and revenues for the three networks.

The paper draws some conclusions. Stating that network comparisons can be sensitive to demand and cost scenarios. In terms of price, premium-service one-price network benefits delay-sensitive users and penalizes delay insensitive users. And the two-tiered network usually gives modest advantage over separated network in this case.

The main conclusion they draw is that that the provision of uniformly high QoS to all traffic is the best in the long run.

The main problem I have with the reasoning of this paper. On one hand they present some historical data (about microprocessor and fiber optics) and suggest that the same would be trend here. On the other hand, they ignore more relevant data and flag it as anomaly. For example the study written in 1992 and published in 1993 [Irvin] developed two models for leased prices. Both predicted a drop of about 50% by 1998 while the prices increased by 50% since 1992. Also, they ignore the difference in patterns of growth in voice versus Internet traffics. While voice traffic increased 10% per year, Internet traffic doubled each year in 1990s except 1995-1996 where it grew by a factor of 10 in each of these two years. Both of these cases- price and utilization- suggest that we can use patterns in other technologies- as they are- as evidence to suggest that just provide uniform high QoS to all traffic. Secondly, as we saw with other technologies, a simple increase can not prevent the users to find ways to use the additional resources to their fullest and demand for more. It was not long when 64K main memory was adequate where 64M is soon termed as too low. Same goes with modems,..etc.