Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Computer Communications 26 (2003) 1457–1469

computer
communications

www.elsevier.com/locate/comcom

Congestion control, differentiated services, and efficient capacity management through a novel pricing strategy[☆]

Adam J. O'Donnell, Harish Sethu*

Department of Electrical and Computer Engineering, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA

Abstract

Pricing is an effective tool to control congestion and achieve Quality of Service (QoS) provisioning for multiple differentiated levels of service. In this paper, we propose a practical, flexible and computationally simple pricing strategy that can achieve QoS provisioning in Differentiated Services networks with multiple priority classes operating in an efficient economic market, while also maintaining stable transmission rates from end-users. In contrast to previous work, in which dynamic pricing strategies are based on the state of congestion alone, our strategy adds a separate price component for the preferential service received by a packet. This permits an efficient market for network resources and services, with the price charged being dependent upon both the cost of the resources and the dynamically changing demand for it. In addition, this automatically enforces efficient capacity management in the allocation of resources among the various service classes, leading to a user-centric approach where a user is not charged a higher price unless preferential service is actually delivered. Our analytical and simulation results demonstrate that, with the combination of user adaptation and our pricing strategy, differentiated services can be achieved with stable transmission rates. This paper concludes with a discussion of various operational issues associated with actual deployment of such a pricing strategy.

Keywords: Internet economics; Differentiated services; Pricing; Congestion control; Quality of service; Capacity management

1. Introduction

The network resources in the Internet are dynamically shared among a large number of users, posing a significant challenge in the guaranteed provisioning of quality-of-service (QoS) to individual users. During the last several years, QoS issues in the Internet have attracted significant research interest as well as commercial investments. One of the ways to achieve QoS guarantees on a per-flow basis is to make a priori reservations of buffer and bandwidth resources in the network. This approach is used in the Integrated Services (IntServ) architecture [4], which relies upon a reservation setup protocol such as RSVP [26]. The per-flow management required at the routers, however, calls into question the scalability of this approach. The Differentiated Services architecture (DiffServ) [3] is an alternate method that achieves improved scalability by aggregating data packets into a small number of service classes and defining router behaviors expected by packets belonging to each of these classes.

DiffServ allows up to 64 different service classes that serve only to define the treatment a packet will receive in relation to other packets, but do not provide absolute guarantees on performance. In the absence of guarantees, as in IntServ, the role of capacity planning for traffic from various classes of service becomes critical to achieving satisfactory service. These bandwidth contracts, referred to as service-level agreements (SLAs), can provide reasonable guarantees only when established over long time scales [18]. The user demands for various levels of service can change rapidly due to a variety of reasons; participation in SLAs between providers, therefore, is not likely to lead to an efficient use of network resources. Mechanisms for capacity planning and congestion control through dynamic pricing, however, can be significantly more efficient and also more responsive to changes in, and the demand for, the network resources. This paper explores a practical, flexible and

---

[☆] A preliminary version of this paper appeared in Proceedings of the IEEE International Conference on Communications, April–May 2002, New York, NY, pp. 986–990.

[*] Corresponding author.

E-mail addresses: sethu@ece.drexel.edu (H. Sethu), adam@ece.drexel.edu (A. J. O'Donnell).

0140-3664/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0140-3664(03)00031-8

computationally simple user-centric pricing strategy that can achieve QoS provisioning in DiffServ networks with multiple priority classes operating in an efficient economic market, while also maintaining stable transmission rates from end-users.

1.1. Related work and motivation

Over the last several years, a number of research proposals have advocated the creation of an economic market for congestion control and differentiated services. One of the simplest means of achieving QoS differentiation is through Paris Metro Pricing [20], which assigns fixed prices for traffic from different service classes and logically partitions the network into separate channels, each with its own allocated bandwidth. All packets traveling through the same channel receive the same level of best-effort service; however, channels with higher prices provide better service because they attract less traffic. Such a scheme leads to an inefficient utilization of the network resources since bandwidth allocated to the higher-price channels may go unused.

An alternate method, which does not pre-allocate bandwidth to different service classes, is based on dividing traffic into multiple priority classes and using a different, but fixed, price associated with each service class [6,17]. Packets belonging to higher-priority classes are given greater forwarding priority at the intermediate routers, thus leading to QoS differentiation. Such a scheme is rendered more efficient by the addition of a congestion-dependent component to the price, allowing for a more competitive price to be offered. One such scheme is studied in Ref. [1], where the pricing changes at a time-scale suitable for typical human response times. In this system, a technical (non-economic) method of congestion control is required at shorter time-scales. The absence of an economic congestion control mechanism at short time-scales, however, may expose the system to abuse by software application vendors or sophisticated users.

Another scheme utilizing congestion-dependent pricing is studied in Ref. [22], where the charges are determined on a per-call basis and are assessed at the time the call is admitted. However, rapid variations in the demand within the duration of one call, typical of Internet traffic today, render such an approach less effective. This issue is addressed in the pricing strategy described in Ref. [25], where the price for each service class depends on the average demand for that service class and is negotiated for short intervals of time. This requires a resource negotiation protocol so that the network can commit resources for these short durations. However, a user may incorrectly anticipate his/her requirements and request a resource commitment but not actually use it. Therefore, such reservation-based approaches, even for short durations, can lead to higher charges for users and inefficient use of the network. In a somewhat related pricing strategy [7], resources may be provisioned per service class over the long term and then priced based on user demand over the short term. This can lead to improved utilization of the network resources over the short-term as users dynamically change transmission rates and/or switch service classes. Such provisioning may be engineered to maximize profits as shown in Ref. [7]; however, over the long term, static allocation of bandwidth among the service classes will lead to a less efficient market with sub-optimal long-term pricing and resource utilization.

An altogether different principle is used in another set of approaches based on the 'smart-market' proposal by MacKie-Mason and Varian [15], where packets are each marked with a bid price that reflects the need of the sending user for the packet to be transmitted. The router admits packets whose bid price is greater than a certain cutoff amount, which is in turn a function of the congestion state of the router. The originator of the packet is charged the lowest bid of all the packets admitted to the router during the time period. Gibbens and Kelly [9] also describe a price-based feedback system for congestion control where packets utilizing a congested router's services are marked as such and end-users are assessed a charge based upon the number of packets thus marked. This principle is extended in Ref. [2] and other related works, where a price variable maintained at each router represents the measure of congestion and is used to determine the probability with which a packet is marked at the router. The marking probability at each router is exponential in the current price, ensuring that the end-to-end probability that a packet is marked before it reaches its destination accurately captures the congestion along the path of the packet. As in Ref. [9], users may be charged based on the number of packets that are marked. A further extension of such a pricing scheme, which uses more than one bit in the packet for the price feedback, is discussed in Ref. [8]. In this pricing strategy, a value representing a price is assigned to each packet as it enters the router. This price is an expression of the social cost incurred by other users due to congestion. Users adapt their transmission rates based on the recent history of prices marked on the packets.

All of the above mentioned pricing schemes provide for economic regulation of individual bandwidth consumption but do not facilitate QoS provisioning for multiple levels of priority service. They all rely upon an FCFS discipline for packet scheduling at the routers and therefore, do not provide a means for one customer to obtain priority service over another. For example, a customer requiring very low delays but not much bandwidth cannot obtain the desired service when pricing schemes are based on the bandwidth consumed but not on the scheduling services rendered.

A pricing scheme that seeks to provide multiple levels of service, with price computations based on both the flow rates and the waiting times, is discussed in Ref. [11]. The computation of prices in this strategy is done periodically