A Mechanism for Dynamic Re-routing of Real-time Channels*

Colin Parris *Hui Zhang* *Domenico Ferrari*

571 Evans Hall

Computer Science Division

University of California at Berkeley

Berkeley, CA 94720

(510)834-8234 (phone)

(510)643-7684 (fax)

Abstract

Various solutions have been proposed to provide real-time services (i.e., services with guaranteed performance requirements) in packet-switched networks. These solutions usually require fixed routing and resource reservation for each conversation. The routing and reservation decisions, combined with load fluctuations, introduce the problems of network unavailability and loss of network management flexibility. We believe that these problems can be alleviated by properly balancing the network load. In this paper, we present a mechanism that dynamically reroutes a real-time channel without disruption of service to the clients. This mechanism is one component in a framework to investigate load balancing in a real-time internetwork. We show that the mechanism can be incorporated into the Tenet real-time protocol suite with minimal changes and overhead.

1 Introduction

With the advent of high-speed networking, there is an increasing demand for new applications such as the transmission of digital video and audio over packet-switched networks. These applications have stringent real-time performance requirements in terms of throughput, delay, delay jitter and loss rate [2]. New solutions have been proposed to provide real-time services in packet-switched networks. These solutions usually require fixed routing and resource reservation on a per-conversation basis [4, 6, 9]. Resource allocation and routing decisions are made at the beginning

---

*This research was supported by the National Science Foundation and the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement NCR-8919038 with the Corporation for National Research Initiatives, by AT&T Bell Laboratories, Hitachi, Ltd., Hitachi America, Ltd., Pacific Bell, the University of California under a MICRO grant, and the International Computer Science Institute. The views and conclusions contained in this document are those of the authors, and should not be interpreted as representing official policies, either expressed or implied, of the U.S. Government or any of the sponsoring organizations.

of the conversation based on the resource availability and network load at that time, and are kept for the duration of the conversation. However, the resource availability and the network load may change during the lifetime of the connection. The resource allocation and routing decisions made at connection setup time, and the changes in resource availability and network load occurring during the lifetime of the connection may introduce two problems: (1) new requests may be denied due to unbalanced resource allocation (see the example in Section 2), and (2) some network management functionality may become less flexible. The severity of these problems increases in an internetworking environment.

One solution to these problems is, in certain cases, to balance the load in the network by dynamically re-routing existing connections. The challenge lies in the fact that this re-routing should not appreciably impact the quality of service of the existing connections.

This paper presents a mechanism for dynamic re-routing of real-time channels. Section 2 provides the motivation for load balancing, and discusses other uses of this re-routing mechanism. Section 3 gives an overview of our real-time environment and a framework for the investigation of load balancing. A dynamic re-routing mechanism is described in detail in Section 4. A summary and a discussion of future work are given in Section 5.

2 Motivation

This section discusses the motivation behind the use of load balancing and the subsequent design of a dynamic re-routing mechanism.

Load balancing in real-time internetworks is desirable to increase network availability, thereby increasing total revenue and maximizing network utilization; to allow network administrators to reclaim resources; and to reduce the impact of unscheduled, run-time maintenance on clients with real-time services.

In networks offering real-time services, special emphasis must be placed on making these services available to the client. The type of network unavailability that is being referred to here is a by-product of the admission control scheme, which is needed to guarantee service performance. To provide real-time services, resources have to be allocated for use by clients. Efficient routing and resource allocation decisions made on previous real-time clients reduce the probability that a new client's request is refused by the admission control scheme. The more efficient the routing and the resource allocation the greater the number of real-time channel requests that the admissions control scheme can accept. As it is very difficult to predict new client requests, efficiency can be increased by re-routing current connections such that new requests are accommodated.

A simple example is shown in Figure 1. In this example, the guaranteed resource is bandwidth. The maximum bandwidth of each link is 10 Mbps. Along the upper route , Upper_0_3, there is a single client who has reserved a bandwidth of 5 Mbps. There is also a single client who has reserved 5 Mbps of bandwidth along the lower route, Lower_0_3. A third client requests a single channel with with a bandwidth of 10 Mbps. While the network has an aggregate unallocated bandwidth of 10 Mbps, this bandwidth is distributed across two disjoint routes, hence the network is forced to deny the client's request, and the network's peak utilization remains at 50%[1]. With load balancing, the single connection along Upper_0_3 can be re-routed along Lower_0_3, thus making the Upper_0_3 route available to the new client. The network peak utilization is now 100%.