Abstract

We focus on a setting where users of a real-time networked application need to be assigned to servers, e.g., assignment of hosts to Internet gateways in a wireless mesh network (WMN). The service delay experienced by a user is a sum of the network-incurred delay, which depends on its network distance from the server, and a server-incurred delay, stemming from the load on the server. We introduce the problem of load-distance balancing, which seeks to minimize the maximum service delay among all users. We address the challenge of finding a near-optimal assignment in a distributed manner, without global communication, in a large network. We present a scalable algorithm for doing so, and evaluate our solution with a case study of its application in an urban WMN.

1 Introduction

The increasing demand for real-time access to networked services is driving service providers to deploy multiple geographically dispersed service points, or servers. This trend can be observed in various systems, ranging from wireless mesh networks (WMNs) [4] to content delivery networks (CDNs) and massively multiplayer online gaming (MMOG) grids [10]. In such settings, every application session is typically mapped to a single server. For example, WMNs provide Internet