Chapter 27
Scalable Algorithms for Server Allocation in Infostations

Alan A. Bertossi
University of Bologna, Italy

M. Cristina Pinotti
University of Perugia, Italy

Romeo Rizzi
University of Udine, Italy

Phalguni Gupta
Indian Institute of Technology Kanpur, India

ABSTRACT
The server allocation problem arises in isolated infostations, where mobile users going through the coverage area require immediate high-bit rate communications such as web surfing, file transferring, voice messaging, email and fax. Given a set of service requests, each characterized by a temporal interval and a category, an integer k, and an integer $h_c$ for each category c, the problem consists in assigning a server to each request in such a way that at most k mutually simultaneous requests are assigned to the same server at the same time, out of which at most $h_c$ are of category c, and the minimum number of servers is used. Since this problem is computationally intractable, a scalable 2-approximation online algorithm is exhibited. Generalizations of the problem are considered, which contain bin-packing, multiprocessor scheduling, and interval graph coloring as special cases, and admit scalable online algorithms providing constant approximations.

INTRODUCTION
An infostation is an isolated pocket area with small coverage (about a hundred of meters) of high bandwidth connectivity (at least a megabit per second) that collects information requests of mobile users.

DOI: 10.4018/978-1-60566-661-7.ch027
and delivers data while users are going through the coverage area. The available bandwidth usually depends on the distance between the mobile user and the center of the coverage area: increasing with decreasing distance. An infostation represents a way in the current generation of mobile communication technology for supporting at many-time many-where high-speed and high-quality services of various categories, like web surfing, file transferring, video messaging, emails and fax. It has been introduced to reduce the cost per bit on wireless communications, and hence to encourage the exchange of ever increasing volumes of information. Infostations are located along roadways, at airports, in campuses, and they provide access ports to Internet and/or access to services managed locally (Goodman, Borras, Mandayam, & Yates, 1997; Wu, Chu, Wine, Evans, & Frenkiel, 1999; Zander, 2000; Jayram, Kimbrel, Krauthgamer, Schieber, & Sviridenko, 2001).

It is desirable that the infostation be resource scalable, that is able to easily expand and contract its resource pool to accommodate a heavier or lighter load in terms of number and kind of users, and/or category of services. Indeed, the mobile user connection lasts for a temporal interval, which starts when the user first senses the infostation’s presence and finishes when it leaves the coverage area. Depending on the mobility options, three kinds of users are characterized: drive-through, walk-through, and sit-through. According to the mobility options, the response time must be immediate for drive-through, slightly delayed for walk-through, and delayed for sit-through. In general, several communication paradigms are possible: communications can be either broadcast or dedicated to a single user, data can be locally provided or retrieved from a remote gateway, and the bit-rate transmission can be fixed or variable, depending on the infostation model and on the mobility kind of the user.

Each mobile user going through the infostation may require a data service out of a finite set of possible service categories available. The admission control, i.e., the task of deciding whether or not a certain request will be admitted, is essential. In fact, a user going through an infostation to obtain a (toll) service is not disposed to have its request delayed or refused. Hence, the service dropping probability must be kept as low as possible. For this purpose, many admission control and bandwidth allocation schemes for infostations maintain a pool of servers so that when a request arrives it is immediately and irrevocably assigned to a server thus clearing the service dropping probability. Precisely, once a request is admitted, the infostation assigns a temporal interval and a proper bandwidth for serving the request, depending on the service category, on the size of the data required and on the mobility kind of the user, as shown in Table 1 for a sample of requests with their actual parameters. Moreover, the infostation decides whether the request may be served locally or through a remote gateway. In both cases, a server is allocated on demand to the request during the assigned temporal interval. The request is immediately assigned to its server without knowing the future, namely with no knowledge of the next request. Requests are thus served on-line, that is in an ongoing manner as they become available.

Each server, selected out of the predefined server pool, may serve more than one request simultane-