PERGAMON

Automatica 35 (1999) 2013–2030

www.elsevier.com/locate/automatica

Information delivery through broadcasting in satellite communication networks☆

Leandros Tassiulas[a,1], Chi-Jiun Su[b,*,2]

[a]Department of Electrical Engineering, University of Maryland, College Park MD 20742, USA

[b]Advanced Development Group, Hughes Network Systems, 11717 Exploration Lane, Germantown MD 20876, USA

Received 30 June 1998; revised 16 February 1999; received in final form 28 June 1999

The problem of server's broadcast scheduling and user's memory management for a broadcast delivery system is considered, optimal policies are identified and near-optimal strategies are proposed.

Abstract

Satellite broadcast is an important candidate for large-scale multimedia information distribution due to the inherent wide-range multicasting capability of satellites and the asymmetry of satellite communications (high bandwidth downlink, limited bandwidth uplink) that matches nicely the information flow asymmetry in multimedia applications. We consider a data broadcasting model that is encountered in most asymmetric satellite communication environments. The problem of scheduling the data broadcast such that average response time experienced by the users is low is considered. In a push-based system, where the users cannot place requests directly to the server and the broadcast schedule should be determined based solely on the access probabilities, we formulate a deterministic dynamic optimization problem, the solution of which provides the optimal broadcast schedule. Properties of the optimal solution are obtained and then we propose a suboptimal dynamic policy which achieves average response time close to the lower bound. In a pull-based system where the users may place requests about information items directly to the server, the scheduling can be based on the number of pending requests for each item. Suboptimal policies with good performance are obtained in this case as well. If a user has local memory, it can alleviate its access latency by selectively prefetching the items from the broadcast and storing them in the memory. A good memory management strategy can substantially reduce the user's access latency. An optimal memory management policy is identified, that minimizes the expected aggregate latency. Memory update strategies with limited look-ahead are presented as implementable approximations of the optimal policy as well. We also consider the problem of joint broadcast scheduling and user's cache management and propos a joint optimization scheme which can achieve the performance up to 40% better than the existing non-joint approach. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Broadcast data delivery; Information distribution methods; Data push/pull; Cache management; Broadcast scheduling; multicast

1. Introduction

Broadcast data delivery is rapidly becoming the method of choice for disseminating information to massive user populations in Direct Broadcast Satellite systems. Under the broadcasting approach as shown in Fig. 1, a server continuously and repeatedly broadcasts information to a user community with limited feedback about the user's needs due to the limited uplink communication capability from the user to the server. The information broadcast by the server is organized into units called *information items*. When a user needs a certain information item, it monitors the broadcast channel until the desired item is detected and captures it for use. The transmission of an information item satisfies simultaneously all the pending requests for that item, and hence it has a multicasting effect.

Broadcast delivery appears to be a promising alternative to the traditional client-to-server transaction due to the communication asymmetry inherent in several

☆This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Guest Editors Venkat Anantharam and Jean Walrand.

[1] Supported in part by an ONR YOUNG INVESTIGATIOR award N000149710735.

[2] The work was performed when the author was with University of Maryland, College Park.

* Corresponding author. Tel.: 001-301-601-7346; fax: 001-301-428-2750.

E-mail address: csu@hns.com (C.-J. Su)

PII: S0005-1098(99)00134-X

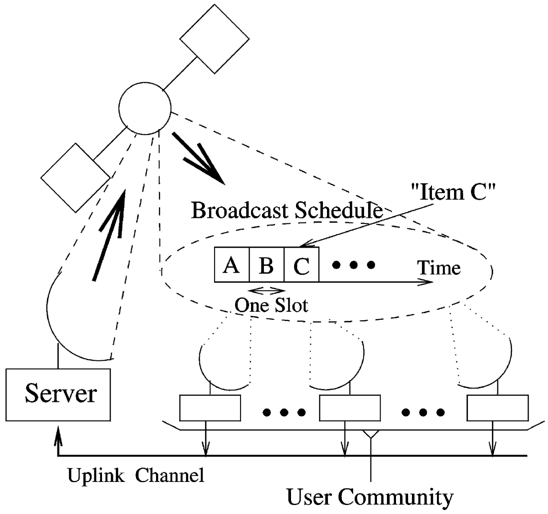

Fig. 1. A direct broadcast satellite data delivery system.

wireless communication environments. Communication asymmetry can generally arise in two ways: either by *physical asymmetry*, when an asymmetric bandwidth or power limitations exist between downlink and uplink communication, or by *asymmetric information flow*, as in an information retrieval system with a large population of users where the amount of information flowing from the server to the users is considerably larger than the flow from the users to the server. The main advantage of broadcast delivery is its scalability: it is independent of the number of users the system is serving.

There are several possible architectures for a broadcast delivery system depending on the capacity of the communication channel from the user population to the server. In *push-based* delivery, the users cannot inform the server about what they actually need due to the lack of, or, limited uplink communication channel from the users to the server. The server should design the schedule relying exclusively on the average user access statistics in this case. In *pull-based* delivery, there is a uplink channel available through which a user can submit access requests to the server and this information can be used for scheduling in a dynamic fashion. When there exists a limited uplink capacity from the users to the server, the users cannot send requests for all the information items in the server's database. In this case, information items are classified into two mutually exclusive sets, namely, DB_{push}, the set of frequently requested items, and DB_{pull}, the set of less popular items. Users are allowed to submit requests only for the items in DB_{pull} while requests for the items belonging to DB_{push} are serviced using push-based delivery.

There are several issues to be addressed in data broadcasting. One of them is the choice of delivery architectures — push, pull or hybrid — based on the capacity of the uplink channel. The main challenge from the server's side is the organization of the data in a broadcast schedule so as to minimize the average response time of user's requests. The problem on the users' side is related to user's local memory management in order to reduce the mismatch between the broadcast schedule and user's access pattern. Both server's broadcast scheduling and user's cache management can be jointly considered in an integrated approach which provides broadcast schedules and user's cache management schemes simultaneously. The issues in a multistage system are cache management in intermediate stages and joint buffer management and schedule design in local broadcast stations.

The problem of broadcast schedule design for push-based delivery has been considered in Gecsei (1983), Ammar and Wong (1987, 1985), Chiueh (1994), Acharya, Franklin and Zdonik (1995), Vaidya and Hameed (1996) and Hameed and Vaidya (1997). The motivation for the work in Gecsei (1983), Ammar and Wong (1987, 1985), was teletext systems. In Ammar and Wong (1985), it was shown that within a rather general class of scheduling policies for data broadcast for which the mean response time exists, a *periodic* policy is optimal in terms of minimizing the mean response time. The results about the parameters of the optimal schedule from Ammar and Wong (1987) were later extended to the case with information items of unequal length and to the case where the transmission of an item is successful with some probability in Vaidya and Hameed (1996). Scheduling for the pull-based system was considered in Dykeman, Ammar and Wong (1986), Dan, Sitaram and Shahabuddin (1994), Shachnai and Yu (1995), Aggarwal, Wolf and Yu (1996). The problem was first considered in Dykeman et al. (1986) in the context of videotex systems and later investigated for a quasi-video on-demand system with impatient users in Dan et al. (1994), Shachnai and Yu (1995) and Aggarwal et al. (1996). A push-based broadcast/pull-based unicast hybrid model was considered in Wong and Dykeman (1989) and Viswanathan (1994). In Stathatos, Roussopoulos and Baras (1997), adaptive partitioning of information items for the same model was provided in which items are grouped according to the estimated user's request generation statistics which is based on the number of requests received through the uplink channel. A push/pull-based broadcast hybrid model was considered in Acharya, Franklin and Zdonik (1997) where, at each slot, a coin weighted by the capacity of the uplink is tossed to determine whether a push-based or pull-based item is transmitted.

In Su and Tassiulas (1997a,b, 1999), the scheduling problem in a push-based system is formulated as a deterministic MDP. Dynamic scheduling policies are considered where the scheduling decision at a slot is based on the elapsed time since the last transmission of each item. Properties of the optimal policy are identified. Furthermore, a class of policies is identified which has near optimal performance, of the same level or slightly better