Resource Sharing for Book-Ahead and Instantaneous-Request Calls Using a CLT Approximation

R. SRIKANT
rsrikant@uiuc.edu
Coordinated Science Laboratory and the Department of General Engineering, University of Illinois, 1308 W. Main Street, Urbana, IL 61801, USA

WARD WHITT
wow@research.att.com
AT&T Labs, Room A117, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA

Abstract. This paper extends the admission control algorithm for book-ahead and instantaneous-request calls proposed by Greenberg, Srikant and Whitt (1997) to cover multiple classes of instantaneous-request calls, each with their own traffic characteristics and their own performance requirements. As before, book-ahead calls specify their starting and finishing times, and are assumed to book far ahead relative to the holding times of the instantaneous-request calls. The book-ahead calls may be constrained by an upper-limit on the capacity that can be reserved for them. Instantaneous-request calls are admitted if the probability of interruption (or some other form of service degradation in response to the conflict) for that call is below a threshold, but now this threshold can be class-dependent, and now the interrupt probability is calculated by a normal approximation based on the central limit theorem. Simulation experiments show that the normal approximation performs as well as the previous detailed calculation in single-class examples, and that the normal approximation can be applied to multi-class examples.

Keywords: book-ahead calls, advance reservation, admission control, integrated-services networks, multimedia, video teleconferencing, loss networks, grades of service

AMS subject classification: 65B15, 65B99, 65C05

1. Introduction and summary

This paper is a sequel to Greenberg et al. [13] in which we studied resource sharing in a telecommunications system when some customers are allowed to book ahead their service requests, i.e., make advanced reservations. In particular, in [13] we developed an admission control algorithm for book-ahead (BA) calls that specify their starting and finishing times in advance and immediate-request (IR) calls that start to receive service immediately, if admitted, and have unspecified holding times governed by a probability distribution. We think of the IR calls as ordinary voice calls and the BA calls as large-bandwidth calls, such as video conference calls, possibly using multiple resources as with multicast, that might well need advance reservation in order to get access (avoid

* Research supported by an NSF CAREER Award NCR 9701525.
very high blocking probabilities). After allowing the BA calls to book ahead (usually subject to some constraints, such as a minimum book-ahead time and an upper-limit on the total capacity that can be reserved), we want to provide as good service as possible to the IR calls. There are also other potential applications, e.g., related to the Internet resource reservation protocol (RSVP); see [13] for additional references. See [6] for a recent theoretical study of booking ahead for a resource with capacity for a single customer.

Our main idea in [13] was to achieve more efficient use of limited resources than can be achieved by strict partitioning by allowing a small probability of conflict between an admitted IR call of uncertain duration and BA calls with scheduled starting times. We think of the BA calls tending to book relatively far ahead (in the time scale of IR call holding times) and the most recently arriving IR calls being interrupted if necessary, although the actual resolution of the conflict could be different. In many applications, it will not actually be necessary to interrupt calls. Instead, the bandwidth or quality of service may be temporarily reduced, e.g., by bit dropping or coarser encoding in video. The admission control algorithm based on interruptions can also be applied with other forms of service degradation. Then the interruption probability should be interpreted as the probability of conflict.

A main conclusion of [13] was that allowing occasional service interruptions or service degradation can yield significantly greater resource utilization and revenue than admission control schemes that do not allow them. It was also observed that a relatively simple nearly-decomposable Markov-chain (ND-MC) algorithm provides a useful approximation for long-run average performance when BA calls book for ahead and have relatively long holding times, and provides an upper bound on revenue more generally. However, most of [13] was devoted to the design and performance of the admission control algorithm. That will be our focus here as well.

Thinking of the BA calls as large-bandwidth calls, such as video conference calls, and the IR calls as ordinary voice calls, in [13] we let the BA calls have very general bandwidth requirements, book-ahead times (time until starting to receive service) and holding times (service durations), but we considered only a single class of IR calls with unit bandwidth requirements and a common holding-time distribution. We let BA calls be admitted subject to an upper limit on the reserved bandwidth for BA calls and possibly a minimum book-ahead time. We let each successive IR call be admitted if the probability that it will be interrupted is below a specified threshold. In [13] our admission control algorithm strongly exploited the fact that there was only a single IR class with common bandwidth requirements.

The main purpose of this paper is to extend our previous admission control algorithm to cover multiple classes of IR calls, with possibly different bandwidth requirements, holding-time distributions and interrupt-probability thresholds. To achieve this goal, we replace our previous interrupt probability calculation by a normal approximation based on the central limit theorem (CLT). We show that the normal approximation is effective, first, by showing that it does as well as our previous algorithm in the previous