Approximation and analysis of a call center with flexible and specialized servers

Robert A. Shumsky

William E. Simon Graduate School of Business Administration, University of Rochester, Rochester, NY 14627, USA (e-mail: shumsky@simon.rochester.edu)

Abstract. This paper describes a decomposition algorithm to estimate the performance of a call center with two types of customers and two server categories. In this system specialized servers can process only one customer type, while flexible servers handle both types. The algorithm divides the system’s state space into regions, and simple approximate models find the conditional system performance within each region. While the procedure described here is tailored for a system with a priority queue discipline and two customer classes, it can be adapted for systems with FCFS queue disciplines and for systems with more than two customer types. Performance measures generated by the procedure are sufficiently accurate for many service system design decisions, such as setting telephone call center staffing levels and long-term capacity planning. The procedure is also extremely fast, and its computational requirements do not grow with system congestion. Numerical tests demonstrate that its running time is significantly lower than traditional numerical methods for generating approximations. As an example of its use, we employ the procedure to demonstrate the benefits of server flexibility in a particular telephone call center.

Keywords: Queueing – Call centers – Capacity flexibility

1 Introduction

This paper describes an approximation procedure for the performance evaluation of queuing systems with servers who are partitioned into specialized and cross-trained groups. The approximation procedure is used to explore the benefits of

* The author thanks Gregory Dobson, who participated in the initial consulting project and problem formulation, as well as Paul Schweitzer, Harry Groenevelt, and two anonymous referees for their helpful suggestions.
capacity flexibility in a local utility’s telephone call center, which provided both the motivation and the test data for this study. The call center employs two types of servers: those who are trained to handle specific consumer billing questions and more experienced generalists who can handle all calls. Similar configurations are found in many environments, including emergency medical services (basic and advanced life support units), digital communications (specialized and general world-wide-web servers), and international services (single-language and multilingual servers).

The goal is to develop an approximation that is sufficiently accurate for use in call center capacity decisions, such as determining hourly staffing levels, planning training schedules, and assessing long-term staffing needs. Because we wish to use the performance measures to explore multiple scenarios and search for optimal staffing configurations, the procedure should also be fast, with running times of at most a few seconds when applied to systems with large numbers of servers.

While similar systems have been examined in the past (see, for example, Green, 1985; Stanford and Grassmann, 1993), previous studies employed numerical procedures for performance analysis that require minutes or hours to converge when applied to large or heavily congested systems. The computational burdens of these procedures grow rapidly with either the number of servers or the servers’ utilization. The approximation described here provides sufficient accuracy for our purposes while significantly reducing the necessary computation time. In particular, the size of the state space of the approximate system grows slowly as the number of servers increases, and the size of the state space does not grow with system utilization.

The procedure is a decomposition algorithm that is built upon a set of well-known approximations for systems with intermittent service (servers with vacations) and systems with intermittent (‘bursty’) arrival processes. These approximations are merged using techniques reminiscent of the aggregation/disaggregation methods that are often applied to systems with large state spaces (Schweitzer, 1991). However, we will see that the approximation presented here is fundamentally different from these general methods.

This procedure is tailored to approximate the performance of a system with a particular type of server flexibility as well as a particular queue discipline (in our application, customers who can only see flexible servers are given priority for those servers). However, we will see that the procedure is readily adaptable to an alternate queue discipline.

In the next section we review the related literature. Section 3 presents the queuing model and the approximation procedure, compares the procedure with traditional aggregation/disaggregation methods, and describes how it can be adapted for a system with a FCFS queue discipline. Section 4 describes numerical experiments to evaluate the accuracy and speed of the procedure. Given parameters based on the call center mentioned above, in Section 5 the approximation will be used to determine the relative value of specialist and flexible servers. Section 6 summarizes these results, and describes how the approximation procedure may be adapted for systems with more than two customer types.