

A Call-Routing Problem with Service-Level Constraints*

Noah Gans

OPIM Department
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6366
gans@wharton.upenn.edu

Yong-Pin Zhou

Dept. of Management Science
School of Business Administration
University of Washington, Seattle
Seattle, WA 98195-3200
yongpin@u.washington.edu

January, 2001

Abstract

We consider a queueing system, commonly found in inbound telephone call centers, that processes two types of work. Type-H jobs arrive at rate λ_H , are processed at rate μ_H and are served first come, first served within class. A service-level constraint of the form $\mathbf{E}\{\text{delay}\} \leq \alpha$ or $\mathbf{P}\{\text{delay} \leq \beta\} \geq \alpha$ limits the delay in queue that these jobs may face. An infinite backlog of type-L jobs awaits processing at rate μ_L , and there is no service-level constraint on this type of work. A pool of c identical servers processes all jobs, and a system controller must maximize the rate at which type-L jobs are processed, subject to the service-level constraint placed on the type-H work.

We formulate the problem as a constrained, average-cost Markov Decision Process and determine the structure of effective routing policies. When the expected service times of the two classes are the same, these policies are globally optimal, and the computation time required to find the optimal policy is about that required to calculate the normalizing constant for a simple M/M/c system. When the expected service times of the two classes differ, the policies are optimal within the class of priority policies, and the determination of optimal policy parameters can be determined through the solution of a linear program with $O(c^3)$ variables and $O(c^2)$ constraints.

*Research supported by the Wharton Financial Institutions Center and by NSF Grant SBR-9733739.