Efficient visit frequencies for polling tables: minimization of waiting cost

O.J. Boxma

Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands; Faculty of Economics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

H. Levy

Department of Computer Science, The Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel

J.A. Weststrate

Faculty of Economics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Received 4 February 1991; revised 14 May 1991

Polling systems have been used as a central model for the modeling and analysis of many communication systems. Examples include the Token Ring network and a communications switch. The common property of these systems is the need to efficiently share a single resource (server) among $N$ entities (stations). In spite of the massive research effort in this area, very little work has been devoted to the issue of how to efficiently operate these systems.

In the present paper we deal with this problem, namely with how to efficiently allocate the server's attention among the $N$ stations. We consider a framework in which a predetermined fixed visit order (polling table) is used to establish the order by which the server visits the stations, and we address the problem of how to construct an efficient (optimal) polling table. In selecting a polling table the objective is to minimize the mean waiting cost of the system, a weighted sum of the mean delays with arbitrary cost parameters. Since the optimization problem involved is very hard, we use an approximate approach. Using two independent analyses, based on a lower bound and on mean delay approximations, we derive very simple rules for the determination of efficient polling tables. The two rules are very similar and even coincide in most cases. Extensive numerical examination shows that the rules perform well and that in most cases the system operates very close to its optimal operation point.

Keywords: Polling table; mean waiting times; minimization of waiting cost.

© J.C. Baltzer A.G. Scientific Publishing Company
1. Introduction

Polling systems have been used to model a large variety of applications in which a single resource is shared among customers accumulating in N distinct queues. These applications include: (1) the Token Ring network in which a single communication channel is shared among N stations each equipped with a buffer to store its messages, (2) a switching node in a communication network which processes messages coming from N different sources, (3) many non-generic applications in which a single program processes requests (messages) coming from N different entities. Many other applications in computer communications and in other fields are commonly represented by this model; for details the reader may refer to Levy and Sidi [19] and Grillo [14].

The use of this model in many scientific and engineering applications has caused the emergence of several variations of the model and of various strategies for operating these systems. This emergence, in turn, has triggered a massive research effort spanned over the last two decades. Nonetheless, this research has focused almost solely on analysis issues, namely of devising mathematical procedures for deriving the performance measures of these models. Almost no work has been done in the optimization of these models, namely on the problem of how to operate polling systems efficiently in order to improve their performance. As a result, designers of these systems have not been equipped with effective tools and guidelines for their efficient operation.

This paper is motivated by the need to derive rules for the efficient operation of polling systems. In Boxma, Levy and Weststrate [7,8] we addressed the problem of finding efficient polling tables in order to minimize the mean amount of work or, equivalently, the weighted sum $\sum_{i=1}^{N} \rho_i W_i$ (where $\rho_i$ and $W_i$ are the utilization of and the waiting time at queue $i$) in systems with exhaustive and gated service.

In this work, we are interested in studying the efficient operation of polling systems in a significantly wider framework. First, the objective function considered here is the sum $\sum_{i=1}^{N} \lambda_i c_i W_i$ in which $\lambda_i$ is the arrival rate at queue $i$ and $c_i$ is an arbitrary parameter reflecting the cost of waiting one unit of time at queue $i$. This sum, therefore, reflects the expected waiting cost per time unit under arbitrarily selected (linear) cost parameters. This is perhaps the single most important performance measure of the system. Second, in terms of modeling aspects, we are concerned here with a framework wider than the one used in Boxma, Levy and Weststrate [7,8]: we consider systems in which the service policy is either exhaustive, gated or limited-1 or a mixture of those (while previously only exhaustive and gated service were considered).

The model under consideration consists of a single server and N queues, $Q_1, \ldots, Q_N$. The server visits the queues in a fixed order specified by a polling table (periodic polling) in which each queue occurs at least once (cf. Eisenberg [11], Baker and Rubin [1]). Common examples are the cyclic order (the table...