

Polling systems

Onno Boxma

EURANDOM and Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

Polling systems are queueing systems consisting of several queues cyclically visited by a single server. They find applications in computer-communications, manufacturing, road traffic, etc. The polling model that is most commonly studied is the following. Customers arrive at N queues according to independent Poisson processes, requiring generally distributed service times. When the server visits Q_i , $i = 1, \dots, N$, it serves a number of customers according to a certain visit discipline. Well-known disciplines are the exhaustive discipline (keep serving a queue until it has become empty) and the gated discipline (keep serving a queue until all those customers have been served who were already present when the server arrived at the queue). When the server moves from Q_i to Q_{i+1} , this usually requires some switchover time.

In the first hour, we shall give an overview of polling systems, focusing on workload (work conservation, work decomposition [2]), on waiting times (pseudo-conservation laws), and on queue lengths (relation to branching processes; [4]).

In the second hour, we discuss the issue of scheduling in polling systems. In almost the whole – vast – polling literature, it is assumed that the service order within each queue is First-Come-First-Served (FCFS). The special feature of our study is that, within each queue, we do not restrict ourselves to FCFS; we are interested in the effect of different service disciplines, like Last-Come-First-Served, Processor Sharing, Random Order of Service, Shortest Job First, and fixed priorities.

Our motivation is that scheduling through the introduction of priorities in a polling system can improve the performance of the system significantly without having to purchase additional resources. Priority polling systems can be used to study the Bluetooth and 802.11 protocols, or scheduling policies at routers and I/O subsystems in web servers. In such settings, it may be desirable to give different requests different priority levels due to the need for Quality-of-Service guarantees.

Using Mean Value Analysis, we first determine the mean waiting times at the various queues [5], for a queue with the gated or exhaustive visit discipline, and for various service orders within the queue. In the last part of the second hour, we indicate how one can also obtain the waiting-time *distributions* in several cases [1, 3].

References

- [1] M.A.A. Boon, I.J.B.F. Adan and O.J. Boxma (2008). A two-queue polling model with two priority levels in the first queue. EURANDOM report, April.
- [2] O.J. Boxma (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems* **5**, 185-214.

- [3] O.J. Boxma, J. Bruin and B. Fralix (2008). Waiting times in polling systems with various service disciplines. EURANDOM report, June.
- [4] J.A.C. Resing (1993). Polling systems and multitype branching processes. *Queueing Systems* **13**, 413-426.
- [5] A. Wierman, E.M.M. Winands and O.J. Boxma (2007). Scheduling in polling systems. *Performance Evaluation* **64**, 1009-1028.