Performance Analysis of Polling Systems

O.J. Boxma & W.P. Groenendijk
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Editorial note. There are dozens of research projects carried out at CWI. It is the editors' policy to pay more attention to these projects. Therefore in the future short descriptions of such projects will be included in the CWI Quarterly. They are presented in a non-mathematical way. The following article belongs to this category.

Introduction

Queueing theory is concerned with the mathematical research of the performance of a system offering services for collective use. Such a system may be a hospital or bank, but also a 'flexible manufacturing system', telephone exchange or computer network. The object of study is formulated in abstract terms as a network of service units and customers requiring services at those units. The nature of the arrival processes and service requests is usually such that they have to be represented by stochastic processes. Hence the most important performance measures, like waiting times, workloads and queue lengths, are random variables. Accordingly, the main techniques of queueing theory stem from probability theory.

In the beginning of this century, queueing theory was developed as a tool for dimensioning telephone exchanges. In the sixties and seventies it turned out that queueing models could also lead to accurate predictions of the behaviour of complex computer systems. This gave a strong impulse to the research on queueing networks. Today, the distributed nature of modern computer-communication networks poses a new challenge to queueing theory. We are beginning to see systems with distributed communications, distributed storage, distributed processing and distributed control. Unfortunately, a thorough understanding of the basic principles of distributed systems is still lacking. Such principles are needed in order to predict performance, to explain behaviour and to establish design methodologies.

At CWI one project is mainly devoted to queueing theory and its application to the performance analysis of computer and communication networks. In the framework of the Government's Information Technology Promotion Plan INSP (1984-1989), computer performance research has been carried out that has resulted in two Ph.D. theses, that have been defended in the

beginning of 1990. The following exposition is concerned with the research area of the thesis of W.P. Groenendijk, viz., distributed control of multi-access communication channels. Below we discuss the queueing model under consideration, and the main results that have been obtained. But first we describe the kind of computer-communication network that has stimulated this research.

Many communication systems provide a broadcast channel which is shared by all connected stations. When two or more stations wish to transmit simultaneously, a conflict arises. The rules for resolving such conflicts are referred to as 'multi-access protocols'. The token ring protocol is one such protocol, that is being used in many local area networks.

In a token ring local area network, a number of stations (terminals, file servers, hosts, gateways, etc.) is connected to a common transmission medium in a ring topology (Figure 1). A special bit sequence called the *token* is passed from one station to the next; a station that 'possesses the token' is allowed to transmit messages. After completion of his transmission the station releases the token, giving the next station in turn an opportunity to transmit. This situation can be represented by the following queueing model, which is known as a *polling model.*

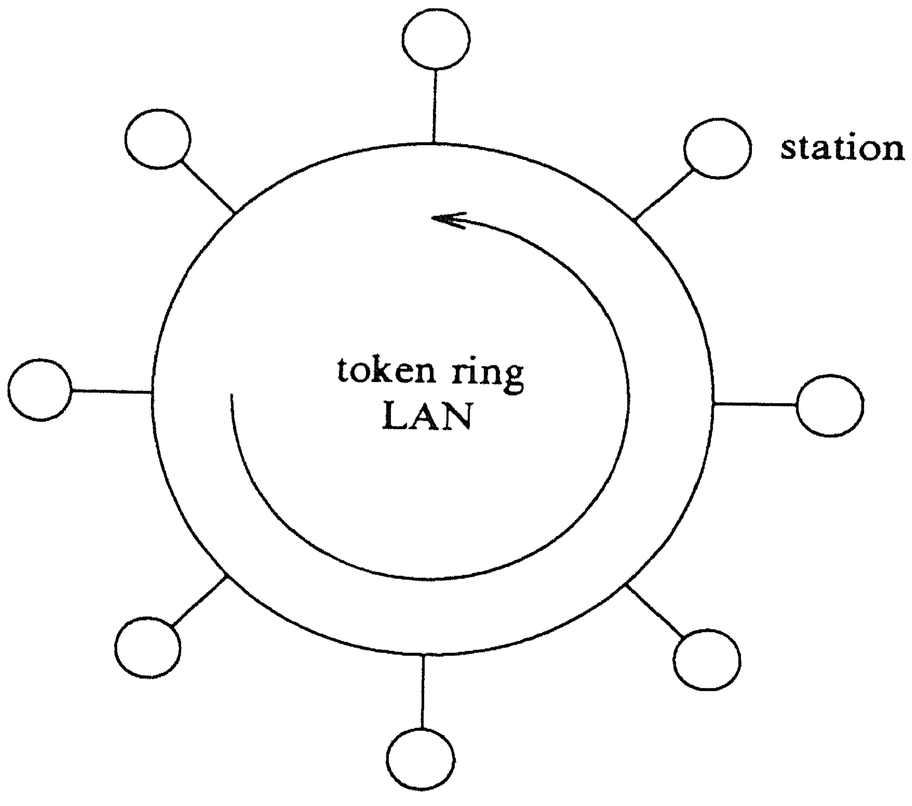

FIGURE 1. A token ring local area network

THE POLLING MODEL

A polling model is a single-server multi-queue model, in which the server attends to the queues in cyclic order (Figure 2). The N queues $Q_1,...,Q_N$ have infinitely large waiting rooms. Arrival times of customers at the queues are usually assumed to occur according to a Poisson process. Service requirements of customers at a queue are independent, identically distributed stochastic

variables; the same holds for the switch-over times of the server between queues. Arrival rates, service time and switch-over time distributions may differ from queue to queue.

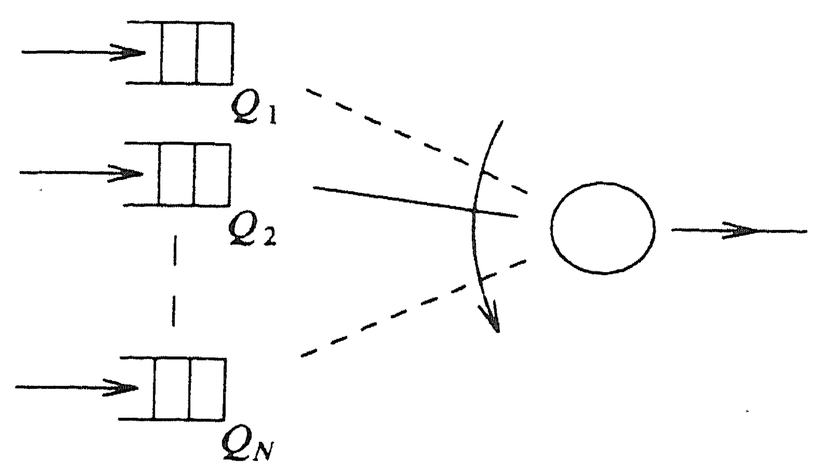

FIGURE 2. Queueing model of a polling system

A polling model describes the behaviour of a token ring local area network in a natural way. The server represents the token-passing mechanism, and the customers represent messages generated at the stations. But many other situations in which several users compete for access to a common resource can be described by this polling model. Examples are a repair man patrolling a number of machines which may be subject to breakdown, assembly work on a carousel in a production system and a signalized road traffic intersection (Figure 3). Depending on the application, various service disciplines at the queues may be considered. Common disciplines are *exhaustive service* (the server continues to work at a queue until it becomes empty), *gated service* (the server serves exactly those customers who were present when he arrived at the queue) and *1-limited service* (the server serves just one customer—if anyone is present—before moving on to the next queue).

ANALYSIS OF THE POLLING MODEL

For polling models with exhaustive and gated service, the steady-state mean waiting times at all N queues can be determined by solving a set of $O(N^2)$ linear equations. The case of 1-limited service is much less amenable to an exact analysis. For only two queues, the joint queue length distribution at both queues can be determined by formulating and solving a boundary value problem for analytic functions, a Riemann or Riemann-Hilbert problem; the mean waiting times are thus also found, being expressed as singular integrals. But the mean waiting times are unknown when there are at least three queues with 1-limited service. However, even in such a case there exists a simple expression for a certain weighted sum of all the steady-state mean waiting times. Twenty-five years ago Kleinrock showed that, under quite weak assumptions, in the case that all switch-over times between queues are zero,

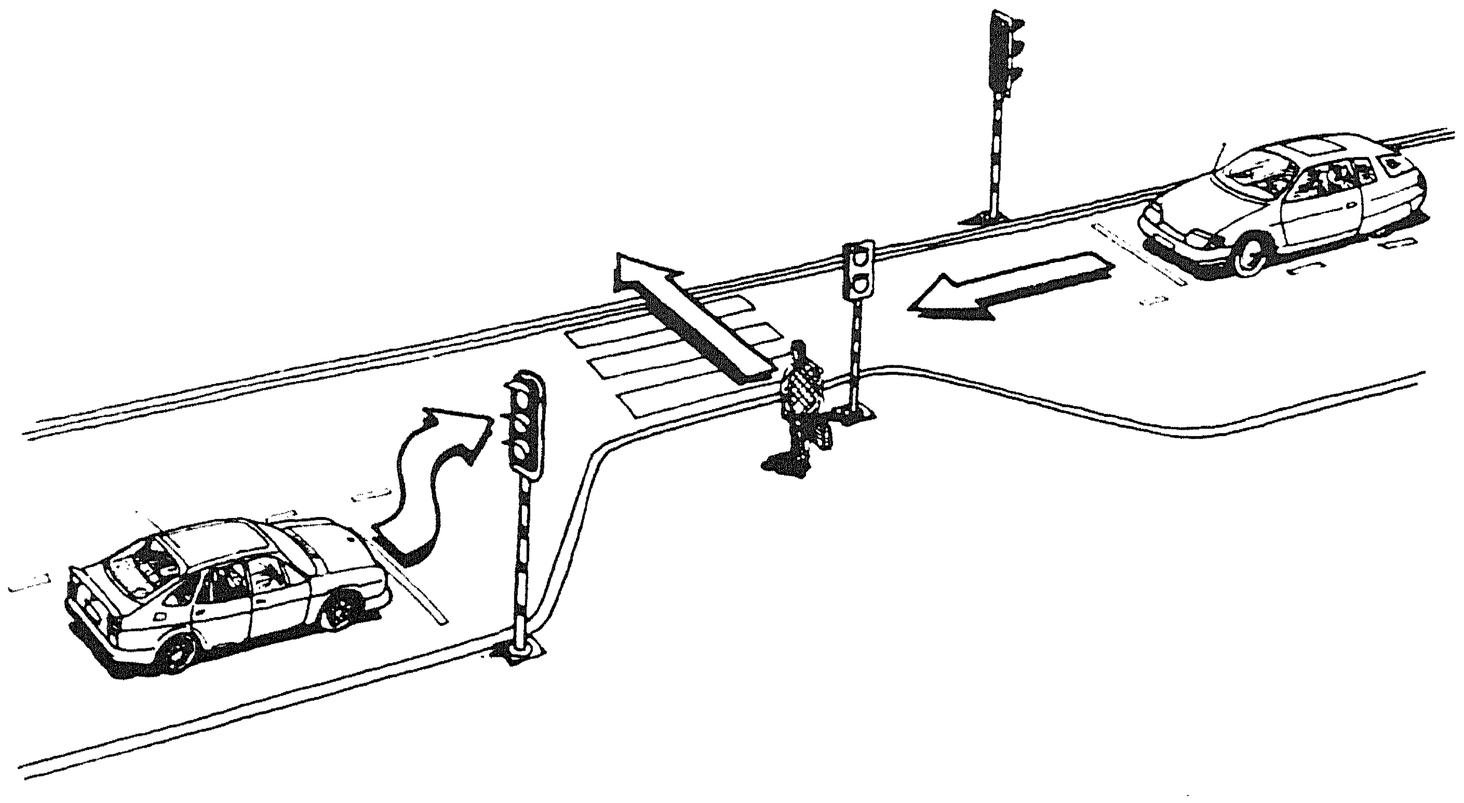

FIGURE 3. A signalized road traffic intersection

$$\sum_{i=1}^{N} \rho_i \, EW_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)}. \tag{1}$$

Here EW_i denotes the mean waiting time at Q_i, λ_i the arrival rate, ρ_i the offered traffic load and $\beta_i^{(2)}$ the second moment of the service requests; $\rho = \Sigma_i \rho_i$ denotes the total offered load. This is being called a *conservation law*: if the service discipline at a queue is changed, the weighted sum of mean waiting times (the left-hand side of (1)) remains the same, although the individual mean waiting times may change.

WORK CONSERVATION AND WORK DECOMPOSITION

The conservation law is a consequence of the 'principle of work conservation'. Suppose the scheduling policy, i.e., the procedure for deciding which customer(s) should be in service at any time, has the properties that it does not allow the server to be idle when at least one customer is present and does not affect the amount of service given to a customer or the arrival time of any customer. Comparing the sample paths of the 'workload process' for such a system under different scheduling disciplines leads to the observation that *the workload process is independent of the scheduling discipline.*

The principle of work conservation has in the past proven to be very useful. It enables one to analyze the workload process of queueing systems with a highly complicated scheduling discipline as if the scheduling discipline were a relatively simple one, such as the First Come First Served discipline.

For the token ring local area network mentioned above, the time for the token to be passed from station to station is in general not negligible. Correspondingly, in the polling model the time the server needs for switching

from station to station has to be taken into account. This fact considerably complicates the analysis: the principle of work conservation is no longer valid, since now the server may be idle (switching), although there is at least one customer in the system. However, under certain conditions there exists a natural modification of the principle of work conservation for polling systems with switch-over times, based on a *decomposition* of the amount of work in the system. This result (see [1], and generalizations in [2,3]) states that—under certain conditions—the amount of work in the polling system, $\mathbf{V}_{\text{with}}$, is in distribution equal to the sum of the amount of work in the simpler 'corresponding' system *without* switch-over times, $\mathbf{V}_{\text{without}}$, plus the amount of work, $\mathbf{Y}$, at an arbitrary moment during a period in which the server is switching from one queue to another:

$$\mathbf{V}_{\text{with}} = \mathbf{V}_{\text{without}} + \mathbf{Y}. \tag{2}$$

The work decomposition gives rise to similar expressions for a weighted sum of the mean waiting times as Formula (1). Taking means in (2) leads to (cf.(1))

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + E\mathbf{Y}. \tag{3}$$

Denote the mean total switching time in one round of the server by s, and the second moment by $s^{(2)}$. Evaluating $E\mathbf{Y}$ yields [1]:

$$\sum_{i=1}^{N} \rho_i E\mathbf{W}_i = \rho \frac{\sum_{i=1}^{N} \lambda_i \beta_i^{(2)}}{2(1-\rho)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}[\rho^2 - \sum_{i=1}^{N} \rho_i^2] + \sum_{i=1}^{N} E\mathbf{M}_i, \tag{4}$$

where $E\mathbf{M}_i$ denotes the mean amount of work in Q_i left behind by the server upon its departure from that queue (when $s \to 0$, the fraction of visits to Q_i in which the server finds Q_i empty tends to one, and the right-hand side of (4) reduces to the right-hand side of (1)). Formula (4) has been coined a *pseudo-conservation law*. The main difference with Kleinrock's conservation law is that now the left-hand side of the formula *does* depend on the service discipline at each queue, through *sum* $E\mathbf{M}_i$. For many service disciplines, amongst which are exhaustive, gated and 1-limited service, we are able to determine the right-hand side of (4) explicitly. Such pseudo-conservation laws often contain the only information available in polling systems with nonzero switching times. They are therefore of considerable practical importance. One of the main features of the pseudo-conservation laws is that they are very useful for testing *and* developing approximations for the individual mean waiting times, which seldom can be determined explicitly.

LAST YEAR'S RESULTS

Boxma and Groenendijk, in close cooperation with H. Levy (Tel-Aviv University) and J.A. Weststrate (Tilburg University), have extended the just described results in several directions.

- An accurate and generally applicable method for approximating mean waiting times has been devised, which may yield insight into the behaviour of a large class of polling systems.
- Formula (4) has been generalized to allow group arrivals. In combination with the developed mean waiting time approximation procedure, this has been used to analyze the performance of Transaction Driven Computer Systems.
- Pseudo-conservation laws have been derived for several non-cyclic server routing mechanisms. This was motivated by the consideration that cyclic routing is more and more becoming a naive strategy, dating from the days in which not enough computing power was available to implement something more sophisticated.
- The ultimate goal of performance modeling and analysis is performance improvement and system optimization. Pseudo-conservation laws appear to be a good starting-point for optimization in polling systems. In the case of non-cyclic service, they have already been used to minimize mean system workload by judiciously choosing the server route. This research direction will be pursued in a forthcoming Ph.D. project.

REFERENCES

1. O.J. BOXMA, W.P. GROENENDIJK (1987). Pseudo-conservation laws in cyclic-service systems. *J. Appl. Prob. 24*, 949-964.
2. O.J. BOXMA (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing systems 5*, 185-214.
3. W.P. GROENENDIJK (1990). *Conservation Laws in Polling Systems*, Ph.D. Thesis, University of Utrecht, January 1990.