Available online at www.sciencedirect.com

SCIENCE DIRECT®

European Journal of Operational Research 165 (2005) 794–809

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Stochastics and Statistics

Finite buffer polling models with routing

Scott E. Grasman [a,*], Tava Lennon Olsen [b], John R. Birge [c]

[a] *219 Engineering Management, University of Missouri—Rolla, Rolla, MO 65409-0370, USA*
[b] *John M. Olin School of Business, Campus Box 1133, Washington University in St. Louis, St. Louis, MO 63130-4899, USA*
[c] *McCormick School of Engineering and Applied Science, Office of the Dean, 2145 Sheridan Road, Northwestern University, Evanston, IL 60208, USA*

Received 6 June 2003; accepted 26 November 2003
Available online 6 May 2004

Abstract

This paper analyzes a finite buffer polling system with routing. Finite buffers are used to model the limited capacity of the system, and routing is used to represent the need for additional service. The most significant results of the analysis are the derivation of the generating function for queue length when buffer sizes are limited and a representation of the system workload. The queue lengths at polling instants are determined by solving a system of recursive equations; an embedded Markov chain analysis and numerical inversion are used to derive the queue length distributions. This system may be used to represent production models with setups and lost sales or expediting.

Keywords: Queueing; Polling; Routing; Setups

1. Introduction

A polling model is a system of multiple queues served by a single server, which requires a setup when switching queues. Polling models have been used extensively to model many computer and communications systems (see, e.g., Levy and Sidi [8] and Takagi [15]), and recently have been used to model other demand-systems such as production and inventory systems (see, e.g., Federgruen and Katalan [3] or Olsen [11]). For a comprehensive review of queueing analysis of polling models see Takagi [13,14,17]. Takagi [13] presents an overview of polling model analysis and applications, an extensive list of references, and various analysis and results. Takagi [14,17] provide updates on research involving polling models.

A number of papers address finite buffers or routing, but to our knowledge, none address both. Sidi et al. [10] utilize the buffer occupancy method to analyze a polling model with routed customers and infinite buffers. They provide results on the expected number of customers in the system at arbitrary instants and

* Corresponding author. Tel.: +1-573-341-7011.
E-mail addresses: grasmans@umr.edu (S.E. Grasman), olsen@olin.wustl.edu (T.L. Olsen), jrbirge@nwu.edu (J.R. Birge).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2003.11.030

the expected delay for a system with Poisson external arrivals, general cyclic service, and general switchover times. Takagi [16] analyzes a finite capacity polling model with Poisson arrivals, general service, and general switchover times. The Laplace–Stieltjes transform (LST) of the waiting time distribution is derived using the M/G/1/K queue with vacations results of Lee [7]. Jung and Un [6] provide an analysis of a finite buffer polling system with exhaustive service based on virtual buffering. This paper uses the buffer occupancy method and the M/G/1/K vacation results of Lee [7] to derive the mean waiting time and blocking probabilities.

In this paper, we model a finite buffer polling model with routing using the buffer occupancy method and the concept of virtual buffering. A finite buffer queue implies that the buffer at each queue has limited capacity and that when the queue is full, new arrivals are turned away. Since routing is allowed, upon completion of service a customer may leave the queueing network or be redirected to another queue in the network. Both routing and finite buffers are realistic modeling elements for the types of communications systems described in Levy and Sidi [8].

Under virtual buffering, which was introduced for polling models by Jung and Un [6], an infinite buffer is virtually present at a queue during vacation periods, i.e., service at other queues and switchover periods. At a polling instant, which is the instant when the server arrives at a new queue in order to serve that queue, the virtual buffer is removed and all excess customers are lost. Thus, the analysis calculates the number of arrivals (external and routed) to the virtual infinite buffer queue during vacation periods, and then calculates the buffer occupancy variables by considering the probability of k arrivals during the vacation period. When the server is serving the queue, the busy period and number served is that of an M/G/1/K queue.

The buffer occupancy approach is based on computing moments of the number of customers present at a polling instant. Specifically, the buffer occupancy approach computes the first and second moments of the number of customers present at a polling instant, which are required for deriving the expected queue length at arbitrary instants (as well as the mean delay of the system). The main principle of the buffer occupancy method is to follow the evolution of the system in the forward direction and compute the moments using a set of linear equations as introduced by Cooper and Murray [2] and Cooper [1]. The approach will be discussed in more detail in Section 2.

The remainder of the paper is organized as follows. The model considered is described in Section 2. Section 3 provides the queueing analysis, including the derivation of the generating function for the number of customers present at polling instants, calculation of the buffer occupancy variables, and an expression for system workload. Section 4 provides the derivation of the queue length distribution and expected queue length. Practical and numerical applicability of the model is discussed in Section 5, followed by conclusions and future work discussion in Section 6.

2. Model description

The polling model considered in this paper is depicted in Fig. 1 and consists of a single server and N finite buffer queues. When the buffer is not full, customers arrive according to independent Poisson processes with rate λ_i, $1 \leqslant i \leqslant N$. Customers that find the buffer full are lost. Customers arriving at queue i are called type i customers and have a service time with LST, $S_i^*(\theta)$, mean s_i and second moment $s_i^{(2)}$, $1 \leqslant i \leqslant N$.

The service at queue i follows the exhaustive discipline, in which the server continues to serve until the queue is empty and then proceeds to the next queue. Thus, all customers found in the queue at the beginning of service and all those that arrive and enter the queue during the service period are served in the given service period. After completion of service, customers may be routed to another queue, but not immediately back to the same queue, with probability, p_{ij}, $1 \leqslant i, j \leqslant N$, or may leave the system. Routed customers that find the queue full are lost. The server then moves to the next queue and incurs a switchover