ESTIMATING VARIANCE OF OUTPUT FROM CYCLIC EXPONENTIAL QUEUEING SYSTEMS

Izak DUENYAS and Wallace J. HOPP

Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA

Received 9 November 1989; revised 13 April 1990

Abstract

Production systems, particularly those making use of a “pull” production control mechanism, are well-modeled as closed queueing networks. Average throughput is clearly one important performance measure for these systems. However, many control decisions require information concerning the variability of the output process as well as throughput. Because of this, the standard deviation of the number of outputs during a specified interval is a practical performance measure for production systems.

In this paper, we consider the standard deviation of the number of outputs during a time interval from a closed queueing network consisting of $M$ single server exponential queues. Because computing this quantity exactly is extremely cumbersome, we introduce a simple approximation that makes use of (1) known results for the variance of the time a marked job takes to complete a round trip and (2) an approximate correction term for the covariance between successive round trips. We show through comparisons with simulation that our method is quite accurate under a variety of conditions.

Keywords: Closed queueing networks, cyclic queues, approximations.

1. Introduction

In an early extension of his ground-breaking work on queueing networks, Jackson pointed out that closed networks are appropriate models for manufacturing systems, since

... real production systems are usually subject to influences which make for increased stability by tending, as the amount of work-in-process grows, to reduce the rate at which new work is injected or to increase the rate at which processing takes place [11].

Whitt concurred with this view, pointing out that in production systems

... closed models are often applied because it seems natural to regard the number of jobs in the system as the independent variable and the throughput as the dependent variable [23].
In recent years, the popularity of pull production systems, such as Kanban [7], have made closed queueing networks even more appropriate models of manufacturing systems. Kanban is well-modeled as a closed queueing network with blocking, while CONWIP (constant work-in-process), which is a broadly applicable pull system (see [20]), is well-modeled as a closed queueing network without blocking. We are interested in closed queueing networks without blocking for two reasons. First, they represent a more analytically tractable starting point for the study of variance of output processes than systems with blocking. Second, they are directly applicable to CONWIP-type production control systems, which have been proposed by various authors (see [5,19,20,21]).

In addition to Jackson's original work, a number of approaches have been developed for analyzing closed queueing networks under a variety of modeling assumptions. These include Mean Value Analysis (MVA) for exponential systems [14], the Fixed Population Mean Method (FPMM) for approximating closed networks with open networks using the Queueing Network Analyzer (QNA) decomposition method [22,23], decomposition approaches other than QNA (see, e.g., [18]), Brownian models for the heavy traffic case [8], among others.

These methods are almost exclusively designed for computing mean quantities, such as throughput, mean queue length, and mean waiting times. For many decisions, such as capacity management and long-range aggregate planning, these mean quantities may be sufficient. However, for making short-range decisions concerning buffer stocks, safety lead times, and production quotas, we need information about the variability of the production process. For instance, setting an economic quota for regular time production in a system that makes use of "under-capacity" scheduling where quota shortfalls are made up on overtime (see e.g., [17]), requires balancing lost revenue versus the cost of running overtime. To perform this optimization, we must be able to compute the average cost for a given production quota, which requires knowledge of the distribution of regular time production. If we make a two moment approximation of this distribution, then we simply need the standard deviation of regular time production in addition to throughput. Several models for addressing the quota-setting problem that make use of the standard deviation of regular time production as an input are given in [10].

In some manufacturing situations it may be practical to empirically estimate the standard deviation of the output during a specified length of time. However, to be able to evaluate candidate changes in a production line, we would like to be able to calculate this standard deviation analytically. In particular, since WIP level is the primary decision variable in a CONWIP system, we would like to be able to estimate standard deviation of output as a function of the number of jobs in the network.

While the Brownian models can be used to generate the distribution of the queue length processes, and there are means for doing this for systems with exponential servers (e.g., [6,3]), neither approach yields information about the