Available online at www.sciencedirect.com
SCIENCE DIRECT®
Future Generation Computer Systems 22 (2006) 820–827
FGCS FUTURE GENERATION COMPUTER SYSTEMS
www.elsevier.com/locate/fgcs

Approximation in non-product form finite capacity queue systems

Nigel Thomas

School of Computing Science, University of Newcastle, UK

Available online 29 March 2006

Abstract

In this paper a class of finite length Markovian queueing models is studied that, in general, does not exhibit a product form solution. Good approximations can be derived for the marginal queue size distributions in this case, and hence measures such as the average response time can be calculated. However, because no product form exists, expressions for the joint queue size distribution are much more costly to derive, hence many performance measures of interest cannot be easily computed. An approximation for the joint queue size distributions is explored here, which improves on a naive product form assumption by considering various boundary cases. The approximations are explored numerically by example.

Keywords: Finite capacity queue; Variance approximation; Stochastic process algebra

1. Introduction

Systems of finite capacity queues are used to model the performance of a wide range of applications, for example software architectures [5], telephone networks [2] and manufacturing systems [11]. A finite Markovian system should always be numerically tractable; however, for large models the cost of solving such a model may be significant. Since the aim of a performance study invariably involves evaluation for a number of different parameter values, or worse, optimisation of certain parameters, a model will usually need to be solved a number of times. As such, techniques that seek to reduce the computational load involved in obtaining a solution are of great importance when building tools to support performance analysts. One such class of techniques involves finding product form solutions that allow the model to be decomposed into individual queues (or small numbers of queues) that can be solved in isolation to find the marginal queue size distributions. Because a product form exists, these marginal distributions can be combined to give exact expressions for the joint queue size distribution, i.e. the steady state probability distribution for the entire model. Unfortunately, product form solutions are not as common as we would like and the conditions by which they are derived are easily broken, particularly by blocking.

Approximate techniques for solving finite capacity queueing systems with blocking have received a very sizeable amount of effort over the past 20 to 30 years and many significant advances have been made. For example, Kouvatsos has used maximum entropy and minimum relative entropy to derive product form approximations for a wide variety of such models (see [10]). A number of surveys of finite capacity queueing systems have been published; Perros [14] surveyed approximation techniques for open queueing models, whereas Onvural [13] surveyed approaches in closed systems. Many key results are further reported by Perros [15] and more recently by Balsamo et al. [4]. A great many other non-product form systems have been studied in the past and expressions found for their marginal distributions. Clearly, marginal distributions have a significant role to play in the analysis of queueing systems, however it would also be desirable to be able to calculate other measures based on the joint queue size distributions.

More recently, Thomas et al. [18] have used the stochastic process algebra (SPA) PEPA [9] to explore an approximation technique for marginal distributions. This approach was inspired by the work of Gribaudo and Sereno [8] and is related to earlier work on stochastic marked graphs [7] and queueing systems [1] and to a method described by Mertziotakis [12].

The approximation relies on the notions of *behavioural independence* and *control* in PEPA, as well as the well formed notions of equivalence which are a feature of process

E-mail address: Nigel.Thomas@ncl.ac.uk.

0167-739X/$ - see front matter
doi:10.1016/j.future.2006.02.005

algebra. Put simply, *behavioural independence* requires that components in a model behave identically regardless of the current behaviour of other components in the model (this property is defined formally and discussed in detail in [17]). If a component is not behaviourally independent, then it must be dependent on some other component to perform one or more actions during its evolution. This property is referred to as *control*, and is more formally defined in [18]. The importance of the notion of control is two-fold. First, it allows an explicit characterisation of where a model fails to be behaviourally independent. If the number of instances of control (the occurrence of actions through which control is exerted) is small, then it may be possible to create an approximate model where these instances are in some way ignored, or an approximate solution where the intervals between controls are long enough for the model to approach steady state behaviour. The second important use for the notion of control is in building approximations that actually seek to exploit this property.

In this paper, this technique is applied to a class of finite capacity queueing models to derive approximations for the marginal queue size distributions. In SPA terms, these models represent components with a rigid structure and an in-built reward structure (the number of jobs in the queue). As such, the results presented here are applicable to other SPA models with similar properties, regardless of whether it represents a queue or not. SPA is not necessarily the best formalism for studying queueing models, as the resultant specification tends to be quite verbose. It is likely that more specific approaches will be able to derive reduced models more quickly than the general method described using SPA (see [3], for example). However, a number of queueing case studies have been published that show how complex and unusual queueing behaviour can be formally defined in SPA [6,16]. Furthermore, SPA is extremely well suited to specifying hybrid models containing both queues and more general components, as is the case for the reduced models derived in this paper.

The remainder of the paper is organised as follows. The class of models is defined in Section 2 and the approximation is discussed in Section 3. The approximation also provides certain boundary values for the joint distribution, and these are used to improve an approximation of the joint queue size distribution based on a simple product form assumption. This is discussed in Section 4, and the results are evaluated numerically using two examples in Section 5. Some brief conclusions and directions of future work are presented in Section 6.

2. The model

In general, the technique presented here can be applied to an arbitrary system of finite capacity queues with simple adaptation. However, for consistency and conciseness, the discussion here will be limited to single server models with a directional flow control. Nodes in the system, as illustrated in Fig. 1, consist of a single queue and server pair with arrivals from one or more other nodes and departures to one or more other nodes.

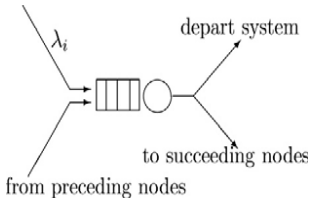

Fig. 1. A single node with external and internal arrivals and departures.

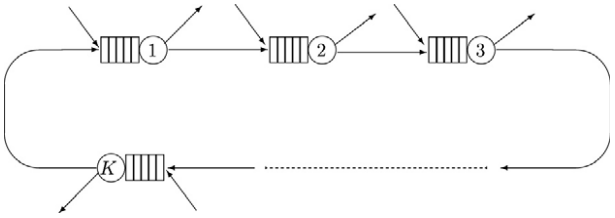

Fig. 2. A system of K queues with a flow control condition.

Service at node i is negative exponentially distributed with mean $1/\mu_i$, and there are N_i buffer spaces at node i. In addition, jobs may enter the system at node i in a Poisson stream at rate λ_i ($\lambda_i \geq 0$) and depart the system after service at node i with a given probability, q_i. In general, the choice of the next destination will be a priori and based on a fixed routing vector and the availability of non-empty queues. If all the queues at successor nodes are full, and the probability of departure at this node is zero, then the job is blocked in service. This means that service will not take place until a successor becomes available, and that service may be suspended at any time subject to the successor queues becoming full. This choice of *blocking before service* is, in essence, an arbitrary one in this instance. The models can be easily adapted to *blocking after service* or *blocking with repeated service* without loss of generality, and the solution method described is generally insensitive to the blocking mechanism.

A single direction flow control is now imposed such that each node i only receives jobs from node $i \ominus 1$, or from outside, and only sends jobs to node $i \oplus 1$, or jobs depart. If there are K nodes in the system, then $i \ominus 1 = i - 1$ if $i \geq 2$ and $i \ominus 1 = K$ if $i = 0$ and $i \oplus 1 = i + 1$ if $i \leq K - 1$ and $i \oplus 1 = 0$ if $i = K$. Furthermore, the restriction is imposed that, if node $i + 1$ is full, then the rate at which jobs continue to depart the system, μ_i', may be less than μ_i. In the first example considered in Section 5, jobs continue to leave the system at rate $\mu_i' = q_i \mu_i$. These restrictions make it clear that any node i is dependent on nodes $i \ominus 1$ and $i \oplus 1$, and so, in general, no product form will exist.

The complete model is illustrated in Fig. 2.

3. Derivation of marginal queue size distributions

An iterative approach is adopted (from [8]) to find approximate solutions to the marginal queue size distributions as follows.

(1) For each node i, generate a Markov modulated arrival process with two states, $A(i)$ and $B(i)$, such that in state $A(i)$ the arrival rate $\lambda_A(i) = \lambda_i + \mu_{i\ominus 1}$ and in state $B(i)$ the