European Journal of Operational Research 196 (2009) 996–1007

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

An analytic finite capacity queueing network model capturing the propagation of congestion and blocking

Carolina Osorio *, Michel Bierlaire

Ecole Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory, Station 18, CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 8 November 2007
Accepted 23 April 2008
Available online 7 May 2008

Keywords:
Queueing
Queueing networks
Finite capacity
Blocking

ABSTRACT

Analytic queueing network models often assume infinite capacity queues due to the difficulty of grasping the between-queue correlation. This correlation can help to explain the propagation of congestion. We present an analytic queueing network model which preserves the finite capacity of the queues and uses structural parameters to grasp the between-queue correlation. Unlike pre-existing models it maintains the network topology and the queue capacities exogenous. Additionally, congestion is directly modeled via a novel formulation of the state space of the queues which explicitly captures the blocking phase. The model can therefore describe the sources and effects of congestion.

The model is formulated for networks with an arbitrary topology, multiple server queues and blocking-after-service. It is validated by comparison with both pre-existing methods and simulation results. It is then applied to study patient flow in a network of units of the Geneva University Hospital. The model has allowed us to identify three main sources of bed blocking and to quantify their impact upon the different hospital units.

1. Introduction

Detecting the sources and effects of congestion within a network allows us to better understand its behavior and to improve its performance. The study of congestion is relevant in a variety of sectors ranging from the analysis of spillbacks (i.e. the backwards propagation of congestion) in urban traffic or pedestrian traffic (Cheah and Smith, 1994) to that of hospital bed blocking (Koizumi et al., 2005) or prison cell blocking (Korporaal et al., 2000).

The most common approach to analyze network congestion is the development of simulation models that capture the details of the underlying system. They are cumbersome to use within an optimization framework. On the other hand, analytic models naturally fit within such a framework but are rarely developed due to the complexity of modeling the propagation of congestion while preserving a flexible model. We focus on analytic models and more specifically on analytic queueing network models.

When modeling a network using a queueing theory framework it is crucial to capture the interactions between the queues. Consider a network of hospital units (e.g. operative and post-operative units) where each unit is modeled as a specific queue and where it is the patient flow that is of main interest. For such a network understanding the correlation between the occupation of the different units can help to avoid bed blocking and to improve a patients recovery procedure. More generally, the between-queue correlation helps to explain the propagation of congestion as well as its effects (such as spillbacks). Moreover, in networks containing loops spillbacks are of special interest because they may lead to deadlocks (also known as gridlocks) (Daganzo, 1996).

The most researched queueing network model is the Jackson network model (Jackson, 1963, 1957) which assumes infinite capacity for all queues. Infinite capacity is a strong assumption that is often maintained due to the difficulty of grasping the between-queue correlation of finite capacity networks. In order to capture this correlation we resort to models with finite capacity queues. The main challenge of such an approach lies in adequately grasping this correlation while also maintaining a tractable model.

Exact finite capacity queueing network (FCQN) models exist only for networks with two or three queues with specific topologies. For more general networks FCQN models are based on approximation methods. Existing analytic FCQN models based on approximation methods either revise queue capacities or vary the network topologies. If queue capacities are revised then they become endogenous parameters. Moreover, approximations need to be used to ensure their integrality and their positivity is only checked a posteriori. We propose an FCQN model which preserves these parameters as exogenous.

Moreover, in this model congestion is not regarded as an underlying phenomenon but is directly modeled. More specifically, we

* Corresponding author. Tel.: +41 21 693 9327; fax: +41 21 693 8060.
E-mail addresses: carolina.osoriopizano@epfl.ch (C. Osorio), michel.bierlaire@epfl.ch (M. Bierlaire).

0377-2217/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2008.04.035

propose a novel formulation of the state space of the queues that explicitly models the blocking phase. Few analytic models incorporating blocking have been developed and there is a recently recognized need for them: "The next generation of the methodology would include an approximation of the blocking of patients in the queueing model" (Cochran and Bharti, 2006). Our formulation yields performance measures that describe both the sources and the effects of congestion.

This paper is structured as follows. We describe the FCQN framework and then review the existing models. The proposed model is then described, followed by its validation versus both pre-existing methods and simulation results. The model is then applied to the study of patient flow within a network of units of the Geneva University Hospital.

2. General framework

We are interested in evaluating the performance of a network of queues. A job is the generic name for the units of interest that flow through the network, e.g. a pedestrian, a prisoner, a patient. We consider open queueing networks where jobs are allowed to leave the network and where the external arrivals arise from an infinite population of jobs. We now describe the general process that a job goes through upon arrival to a queue. Jobs arriving to a queue are either served immediately or wait until a server becomes available. Once a job is served it is routed to its next queue according to a probabilistic routing model. We call this queue the target queue. If this target queue has finite capacity then it may be full. If it is full then the job is **blocked** at its current location. Once there is a place at the target queue the job is unblocked and proceeds to the target queue. The jobs are unblocked with a first in first out (FIFO) mechanism.

Various blocking mechanisms have been defined in the literature (Balsamo et al., 2001). They differ either in the moment the job is considered to be blocked (e.g. before or after-service) or in the routing mechanism of blocked jobs. The blocking mechanism that we have just described is known as blocking-after-service.

The average arrival rate to queue i is denoted λ_i. Queue i has c_i parallel servers, each one serving with an average rate μ_i. The total number of jobs allowed in the queue is called the capacity of the queue, k_i, the buffer size is $k_i - c_i$. The possible routings among queues are given by the transition probability matrix (p_{ij}), where p_{ij} denotes the probability that a job at queue i is routed to queue j.

3. Literature review

A first survey of FCQN models was made by Perros (1984), who later on also wrote a historical overview of the research motivations and advances in networks with blocking (Perros, 2003). A detailed introductory book was written by Balsamo et al. (2001). Surveys focusing on specific application fields exist for the software architecture sector (Balsamo et al., 2003), the production and manufacturing sector (Papadopoulos and Heavey, 1996) and on retrial queues for the telecommunications sector (Artalejo, 1999).

The joint stationary distribution of the network, which contains the probability of each possible state of the network, allows us to derive the main network performance measures. We distinguish between models that allow the exact evaluation of this joint stationary distribution and those based on approximation methods.

3.1. Exact methods

Exact methods consist of either closed form expressions or numerical evaluation of the joint stationary distribution. For an FCQN the between-queue correlation suggests a non-product form joint stationary distribution. Thus closed form expressions are difficult to obtain and are only available for single server networks with two or three queues in tandem topologies (Grassman and Derkic, 2000; Langaris and Conolly, 1984; Latouche and Neuts, 1980; Konheim and Reiser, 1978; Konheim and Reiser, 1976) or two queues in closed networks (Akyildiz and von Brand, 1994; Balsamo and Donatiello, 1989).

On the other hand, exact numerical evaluation of the joint stationary distribution can be obtained by solving the global balance equations (these are detailed in Section 4.1). A detailed description of these numerical methods can be found in Stewart (2000). These equations require the construction of the transition rate matrix, i.e. the description of the transition rates between all feasible states of the network. This time consuming task is therefore only conceivable for small networks (i.e. small in the number of queues and their capacity). This approach also lacks flexibility because changes in the network topology require redefining the transition rate matrix. If the networks of interest have a more general topology or an arbitrary size then their analysis is done by models based on approximation methods.

3.2. Approximation methods

Models based on approximation methods can be classified into either simulation-based or analytic models. The use of simulation models is the most popular approach to evaluate the performance of finite capacity queueing networks. Surveys of simulation models exist for sectors such as transportation (Nagel, 2002; Ben-Akiva et al., 2001), healthcare (Fone et al., 2003; Jun et al., 1999), computer science (Sadoun, 2000; Obaidat, 1990) and the analysis of call centers (Koole and Mandelbaum, 2002; Mandelbaum, 2001). This approach although more realistic and detailed, is cumbersome to optimize, and its accuracy is strongly dependent on the quality of the calibration data (Korporaal et al., 2000). Analytic models are simpler, less data expensive and more flexible.

The main motivation of analytic models based on approximation methods is to reduce the dimensionality of the system under study. Decomposition methods achieve this by decomposing the network into subnetworks and modeling each subnetwork independently. The structural parameters of each subnetwork (e.g. average arrival and service rates) depend on the state of other subnetworks and thus capture the correlation with other subnetworks. The main difficulty lies in obtaining good approximations for these parameters so that the stationary distribution of the subnetwork is a good estimate of its marginal stationary distribution. Given a subnetwork its stationary distribution can be obtained by either establishing a behavioral analogy with a network whose distribution has a closed (and often product) form, or by exact numerical evaluation of the global balance equations which now have a smaller dimension but are often nonlinear.

Existing models based on decomposition methods have defined subnetworks consisting of single queues, pairs of queues or triplets. We call these methods single, two queue and three queue decomposition methods, respectively. If not stated otherwise the models concern open finite capacity networks with exponentially distributed service times.

The most commonly used decomposition method is single queue decomposition. The first model based on this method dates back to the work of Hillier and Boling (1967) who considered tandem single server networks. One of the most used models based on single queue decomposition concerns single server feed-forward networks where each finite capacity queue is transformed into an M/M/1 queue, and the blocking is taken into account by revising the arrival and service rates of the queues (Takahashi et al., 1980). An extension of this model to queues with multiple servers is given