Performance Evaluation 118 (2018) 33–47

Contents lists available at ScienceDirect

Performance Evaluation

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

Coupled queues with customer impatience

Ekaterina Evdokimova [a], Koen De Turck [b], Dieter Fiems [a,*]

[a] Department of Telecommunications and Information Processing, Ghent University, Belgium
[b] Laboratoire des Signaux & Systèmes, CentraleSupélec, France

ARTICLE INFO

Article history:
Available online 22 November 2017

Keywords:
Kitting process
Queues with customer impatience
Regular perturbation
Fluid limit

ABSTRACT

Motivated by assembly processes, we consider a Markovian queueing system with multiple coupled queues and customer impatience. Coupling means that departures from all constituent queues are synchronised and that service is interrupted whenever any of the queues is empty and only resumes when all queues are non-empty again. Even under Markovian assumptions, the state–space grows exponentially with the number of queues involved. To cope with this inherent state–space explosion problem, we investigate performance by means of two numerical approximation techniques based on series expansions, as well as by deriving the fluid limit. In addition, we provide closed-form expressions for the first terms in the series expansion of the mean queue content for the symmetric coupled queueing system. By an extensive set of numerical experiments, we show that the approximation methods complement each other, each one being accurate in a particular subset of the parameter space.

1. Introduction

We investigate the performance of a particular Markovian queueing system with K parallel queues, as depicted in Fig. 1. The queues have finite or capacity; let $C_k \in \mathbb{N}^+$ be the capacity of the kth queue. Customers arrive at the kth queue in accordance with a Poisson process with rate $\lambda_k > 0$, the arrival processes at the different queues being independent. We further assume that departures from the different queues are coupled. This means that there are simultaneous departures from all queues with rate μ as long as all queues are non-empty. If one of the queues is empty, no service takes place. Finally, customer impatience is assumed: each customer leaves the kth queue prior to service with abandonment rate α_k with the exception of customers whose service has started.

The queueing system described above is a natural abstraction for an assembly process with multiple inventories; see [1,2] and the references therein for advances in stochastic inventory models. The different queues represent part inventories for the different parts that are used during assembly. These inventories are continuously replenished by in-house production facilities (in accordance to a Poisson process), the inventories offering temporary storage to smooth out uncertainty in the various production processes. Parts are assumed to be perishable, meaning that they should be used before a (random) due-date or be discarded once this due-date is crossed. This perishability is captured by the abandonment processes from the different queues. Food-products are a prime example of perishable semi-finished products. However, perishable semi-finished products are also found in biochemical production, and in battery and semiconductor manufacturing [3]. Finally, assuming that assembly requires that all the necessary inputs are available, it can only proceed if the inventories (or queues) are not empty, which corresponds to the notion of the coupled departures introduced above.

* Corresponding author.
E-mail address: dieter.fiems@ugent.be (D. Fiems).

https://doi.org/10.1016/j.peva.2017.10.002
0166-5316/

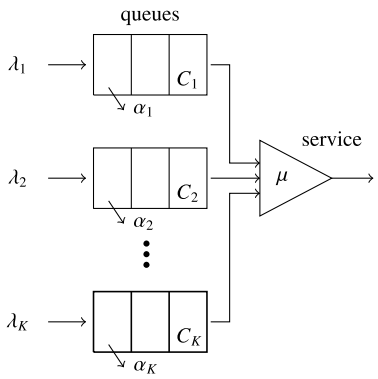

Fig. 1. Representation of the coupled queueing system with customer impatience.

The two-buffer coupled queueing system without customer impatience is well understood. If the buffer capacity is infinite, the uncontrolled queue process is null recurrent in the Markovian setting. The inherent instability of such queueing systems is demonstrated in [4] where the buffer content difference is studied in the two-queue case. Assuming finite capacity buffers, Hopp and Simon developed a model for a two-buffer kitting process with exponentially distributed processing times for kits and Poisson arrivals [5]. The exponential service times and Poisson arrival assumptions were later relaxed in [6] and [7], respectively.

Only a few authors have studied coupled (or paired) queueing systems with multiple (i.e. more than two) queues. In [8], Harrison studies stability of coupled queueing under very general assumptions: $K \geq 2$ infinite-capacity buffers, generally distributed interarrival times at the different buffers, and generally distributed service times. He proves that stability requires buffer control, or more precisely, that the distribution of the vector of waiting times (in the different queues) without control and infinite queue capacity is defective. When the queues are finite, such a control is not necessary. The queue content of the coupled queueing system with finite buffers is studied in [9], assuming exponential service and Poisson arrivals. As the size of the state–space of the associated Markov chain grows quickly with the number of queues involved, [9] presents an approximation for the queue content when the system is in the overloaded regime.

In contrast to the uncontrolled coupled queueing system, the controlled coupled queueing system has received considerable attention in the scientific literature. Ramakrishnan and Krishnamurthy adopt the term *synchronisation station* and present a recent account on approximations of such systems [10]. A particular type of control of coupled queues relates to fork-join type queueing system [11,12]. In fork-join systems, a job is forked into different sub-jobs, run on different servers. Upon completion of all sub-jobs, there is a final service joining the sub-jobs again. The server joining the sub-jobs operates as a coupled server, albeit with a controlled arrival process. Indeed, the sub-jobs that need to be merged, are already present in the fork-join system. These will be available for the coupled server after some delay.

Coupled queueing may also refer to different types of multi-queueing systems, most prominently to systems with discriminatory processor sharing. In discriminatory processor sharing the total service capacity is distributed amongst all queues that have waiting customers, some queues getting a larger share than others. Once one of the queues is empty, its share is moved to the queues with waiting customers. The authors in [13] investigate such a two-queue system where customers in both queues are served at unit rate when both queues are non-empty, while the non-empty queue is served at a higher rate when the other is empty. A similar system is studied in [14] in the heavy traffic regime while [15] allows for time varying arrival rates and the possibility that jobs abandon. In contrast to [13–15], jobs in the first queue do not leave the system but move to the second queue upon completion in [16]. Finally, [17] studies the stability of a more generic system with multiple queues where the service rate of each queue depends on the number of customers in all queues.

The present paper investigates approximations for multi-buffer coupled queuing systems with customer impatience, with service coupling as described above. We investigate two numerical approximation techniques as well as the fluid limit of the system at hand. The numerical approximation methods rely on a Maclaurin-series expansion of the steady-state probability vector, either around $\lambda = 0$ (light-traffic regime) or around $\alpha = \mu = 0$ (overloaded regime). Series expansion techniques for Markov chains are referred to as perturbation techniques, the power series algorithm or light-traffic approximations. While the naming is not absolute, perturbation methods are mainly motivated by sensitivity analysis of performance measures with respect to the system parameters. In particular singular perturbations where the perturbation does not preserve the class-structure of the non-perturbed chain, have received considerable attention in literature, see [18–20] and the references therein. The power series algorithm transforms a Markov chain of interest in a set of Markov chains parametrised by an auxiliary variable ρ. For $\rho = 0$, the chain can be solved efficiently, and one can also obtain the perturbation of the chain in ρ. For $\rho = 1$ the original Markov chain is retrieved such that the series expansion can be used to approximate the solution of the original Markov chain, provided the convergence region of the series expansion includes $\rho = 1$, see e.g. [21–24]. Finally, light-traffic approximations often correspond to a series expansion in the arrival rate at a queue. For an overview on the technique of series expansions in stochastic systems, we further refer the reader to the surveys in [25] and [26].