Computers & Operations Research 73 (2016) 141–149

Contents lists available at ScienceDirect

Computers & Operations Research

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

Breaking the dimensionality curse in multi-server queues

Alexandre Brandwajn [a], Thomas Begin [b,*]

[a] PALLAS International Corporation, San Jose, CA, USA
[b] Univ Lyon, UCB Lyon 1, CNRS, ENS de Lyon, Inria, LIP UMR 5668, France

ARTICLE INFO

Article history:
Received 20 April 2015
Received in revised form
25 November 2015
Accepted 8 April 2016
Available online 9 April 2016

Keywords:
Multi-server systems
G/*G*/*c* queue
State-dependent *Ph*/*Ph*/*c*-like queue
Dimensionality curse
Reduced state description
Approximate solution

ABSTRACT

Ph/*Ph*/*c* and and *Ph*/*Ph*/*c*/*N* queues can be viewed as a common model of multi-server facilities. We propose a simple approximate solution for the equilibrium probabilities in such queues based on a reduced state description in order to circumvent the well-known and dreaded combinatorial growth of the number of states inherent in the classical state description. The number of equations to solve in our approach increases linearly with the number of servers and phases in the service time distribution. A simple fixed-point iteration is used to solve these equations. Our approach applies both to open models with unrestricted buffer size and to queues with finite-size buffers.

The results of a large number of empirical studies indicate that the overall accuracy of the proposed approximation appears very good. For instance, the median relative error for the mean number in the queue over thousands of examples is below 0.1% and the relative error exceeds 5% in less than 1.5% of cases explored. The accuracy of the proposed approximation becomes particularly good for systems with more than 8 servers, and tends to become excellent as the number of servers increases.

1. Introduction

A number of areas of computer applications and systems offer examples of multi-servers facilities. For instance, many-core CPUs with 32 or more cores are around the corner [7]. Parallel Access Volumes in mainframe storage [17] provide a potentially large number of "exposures" for simultaneous access to information. Call centers with hundreds of agents [9] are an element of everyday life. In the area of fiber optical cables, WDM multiplexing allows over a hundred simultaneous signals on a single fiber.

Such systems can be naturally represented as multi-server queues in which requests arrive, queue for service if all servers are found busy, and eventually leave the system after receiving service from one of the servers. Unfortunately, if one realistically assumes general distributions of times between request arrivals and general service times, the resulting *G*/*G*/*c* queueing model does not possess a known analytical solution except in some special cases [12,20,3,1]. Additionally, under higher loads, realistic models must account for finite buffer space (queueing room) which may prevent requests from joining the queue when the buffer is full.

A common approach is then to replace the "general" distributions by so-called "phase-type" distributions [13,5,18] as any distribution can be approximated arbitrarily closely by a phase-type distribution (e.g., [13]). This has the distinct advantage of leading to a system of linear equations if one is interested in the steady-state probability distributions in such a system. In queueing terms, the *G*/*G*/*c* queue is replaced by the *Ph*/*Ph*/*c* queue. The latter can be solved numerically using matrix geometric methods [19,16,4]. This approach works great as long as the number of servers and/or phases in the arrival and especially service process is not too large. However, as mentioned above, the number of servers in many realistic examples varies from several tens to many hundreds, and the traditional phase-type approach is known to suffer from the "dimensionality curse" in that the number of states (and, hence, equations to solve in the linear system) grows combinatorially as the number of servers and /or phases increases. This precludes the direct use of this approach in many interesting and important areas.

In the area of approximate solutions to such systems, several authors attempt to summarize the properties of general distributions in *G*/*G*/*c* queues by their first 2 (rarely, 3) moments [11]. Although the resulting approximations are usually simple to implement and their execution is fast, unfortunately, they fail to account for the intrinsic dependence of the *G*/*G*/*c* queue on higher-order properties of the distributions involved [10,2] (see also Appendix). Fluid queues represent another avenue for approximation based on the fact that, as the number of requests in a queueing system tends to infinity, one can consider the flow of discrete requests as a continuous flow and hence apply fluid mechanics equations to describe the system. These methods have been applied, for example, to represent call centers [15,9] and the *G*/*G*/*c*/*N* queue [22]. By their principle, these approximations appear best suited for the study of limiting processing capacities of such

* Corresponding author.

http://dx.doi.org/10.1016/j.cor.2016.04.011
0305-0548/

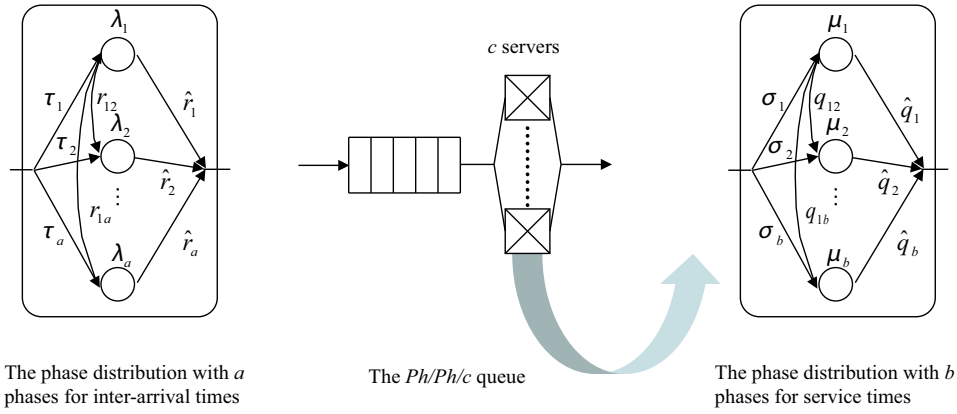

Fig. 1. *Ph*/*Ph*/*c* queue with unrestricted buffer.

queues.

In the particular case when the arrivals can be treated as generated by a Poisson or quasi-Poisson source, there has been recent progress in obtaining computationally manageable approximations applicable to systems with hundreds of servers. Khazaei et al. [14] propose to use an adaptation of the embedded Markov Chain method in the case of a pure Poisson arrival process. They show the good accuracy of their numerical results for service time coefficients of variation not exceeding 1.4. The finite buffer size in their numerical results is relatively small and kept at less than half the number of servers. Their approach does not seem easy to apply to systems with state dependencies or more general arrival processes. Brandwajn and Begin [8] introduce an approximation based on a reduced state description and demonstrate the accuracy of their approach for much larger range of coefficients of variation of the service time distribution (up to 7) and buffer sizes.

Clearly, in many situations the arrival process cannot be viewed as Poisson or quasi-Poisson. Therefore, in this paper we present an extension of the reduced state approximation to systems with phase-type distributions of the time between arrivals. We also extend the approximation to open queues i.e., queues with unlimited buffer size. While no human-made system possesses a truly unrestricted buffer size, such models are of practical interest when the physical buffer size is relatively large and the server utilization is not too close to saturation. In such cases the use of an open model may result in computational saving over a finite-buffer-size model. Interestingly, in open queues, our approximation happens to tend to the correct asymptotic rates of request arrivals and service rates as the number of requests tends to infinity.

To avoid arbitrary truncation in the open model, we transform the balance equations for the reduced state into equations for the conditional probabilities of the state of the arrival process and the reduced state of the service given the current number of requests. We then exploit the convergence of such conditional probabilities to their asymptotic values so as to enumerate the states only up to the practical asymptotic convergence point.

The use of conditional probabilities partitions the state space into independently normalized subspaces, which may contribute to numerical stability, and we employ them also in the case of finite buffers. We propose a simple fixed-point iteration to solve the conditional probability equations. Although we do not have a theoretical proof of convergence to a unique solution, the proposed iteration has never failed in the large number of cases explored.

Thus, the contributions of this paper include:

- An approximate solution of the *Ph*/*Ph*/*c* queue with unrestricted buffer,
- An approximate solution of the *Ph*/*Ph*/*c*/*N* queue with finite buffer and possible state dependencies.

Besides the addition of the state description to account for non-Poisson (or quasi-Poisson) arrival process, this paper extends the work presented in [8] by using conditional probability equations in the solution, which simplifies the treatment of the queue with unrestricted buffer.

This paper is organized as follows. Section 2 is devoted to the approximate solution of the *Ph*/*Ph*/*c* queue with an infinite buffer. In Section 3 we consider a queue with a finite buffer and state-dependent distributions of interarrival times and service times. Section 4 presents numerical results to illustrate the accuracy of the proposed approximation. Finally, Section 5 concludes this paper.

2. Open model and its solution

We start by considering a classical *Ph*/*Ph*/*c* queue with an infinite buffer [11]. We assume that the c servers are homogeneous, i.e., statistically identical, but not synchronized. As shown in Fig. 1, the distribution of the times between arrivals comprises a exponential phases and the service time distribution has b exponential phases. We denote by σ_i the probability that service starts in phase i, by μ_i the completion rate for phase i ($i = 1, \ldots, b$) of the service process, and by $\hat{q}_i$ the probability that the service process completes after phase i. We denote by τ_j, λ_j and $\hat{r}_j$ the corresponding quantities for phase j ($j = 1, \ldots, a$) of the arrival process.

The classical approach to derive the steady-state probability of the number of requests (customers) in such a system is to consider a state description that includes the current number of requests in the system (n), the current phase of the arrival process (j), and the vector of the current number of requests in each phase of the service process ($\vec{m} = m_1, \ldots, m_b$). It is clear that (for each value of n) such a full state description results in a combinatorial explosion of the number of balance equations one has to solve as the number of servers and service phases increases, compounded by the number of arrival phases.

As we are focusing on systems with large numbers of servers, to escape this issue, we use a reduced state description comprising the current number of users (n), the current phase of the arrival process (j) and the current phase of the service process for an arbitrarily selected server (i). Since for $n < c$ the selected server may be idle, we use the value $i = 0$ to denote its idle state. Let $p(n, j, i)$ be the corresponding steady-state probability where