UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Mean sojourn time in two-queue fork-join systems: bounds and approximations

Kemper, B.; Mandjes, M.

DOI
10.1007/s00291-010-0235-y
Publication date
2012
Document Version
Final published version
Published in
OR Spectrum

Link to publication

Citation for published version (APA):
Kemper, B., & Mandjes, M. (2012). Mean sojourn time in two-queue fork-join systems: bounds and approximations. *OR Spectrum*, *34*(3), 723-742. https://doi.org/10.1007/s00291-010-0235-y

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:06 Dec 2021

OR Spectrum
DOI 10.1007/s00291-010-0235-y

REGULAR ARTICLE

Mean sojourn times in two-queue fork-join systems: bounds and approximations

Benjamin Kemper · Michel Mandjes

© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract This paper considers a fork-join system (or: parallel queue), which is a two-queue network in which any arrival generates jobs at *both* queues and the jobs synchronize before they leave the system. The focus is on methods to quantify the mean value of the 'system's sojourn time' S: with S_i denoting a job's sojourn time in queue i, S is defined as $\max\{S_1, S_2\}$. Earlier work has revealed that this class of models is notoriously hard to analyze. In this paper, we focus on the homogeneous case, in which the jobs generated at both queues stem from the same distribution. We first evaluate various bounds developed in the literature, and observe that under fairly broad circumstances these can be rather inaccurate. We then present a number

Part of this work was done while M. Mandjes was at Stanford University, Stanford, CA 94305, USA.

B. Kemper (✉)
Institute for Business and Industrial Statistics, University of Amsterdam,
Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands
e-mail: b.p.h.kemper@uva.nl

B. Kemper
Department of Quantitative Economics, Faculty of Economics and Business,
University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands

M. Mandjes
Korteweg-de Vries Institute for Mathematics, University of Amsterdam,
Science Park 904, 1090 GE Amsterdam, The Netherlands
e-mail: m.r.h.mandjes@uva.nl

M. Mandjes
EURANDOM, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

M. Mandjes
CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Published online: 15 January 2011

of approximations, that are extensively tested by simulation and turn out to perform remarkably well.

Keywords Queueing · Fork-join network · Simulation · Parallel processing · Synchronization · Throughput time

1 Introduction

Fork-join systems (or: parallel queues) are service systems in which every arrival generates input in multiple queues. One could for example consider a Poissonian arrival stream (with rate λ) that generates jobs in two queues. The service times in queue i (for $i = 1, 2$) constitute an i.i.d. sequence of non-negative random quantities $(B_{i,n})_{n\in\mathbb{N}}$ (distributed as a generic random variable B_i), where in addition both sequences $(B_{1,n})_{n\in\mathbb{N}}$ and $(B_{2,n})_{n\in\mathbb{N}}$ are assumed to be mutually independent. After their service the two jobs synchronize before leaving the system. One could call the resulting queueing system an 'M/G/1 fork-join system'. To ensure that the system is stable, one imposes the obvious condition that $\lambda\,\mathbb{E}B_i$ be smaller than 1 for both $i = 1$ and 2.

While the distribution of the sojourn time of both individual queues, which behave as M/G/1 queues, is explicitly known (albeit in terms of its Laplace transform, through the celebrated Pollaczek–Khinchine formula), considerably less is known about the joint distribution of the workload in both queues of the parallel queue. It is clear that these workloads are positively correlated: if the workload of one of the queues is larger than usual, a potential reason for this is that there were temporarily unusually many arrivals, such that the workload in the other queue is probably larger than average as well. The level of correlation is primarily caused by the shape of the distributions of B_1 and B_2; as can be seen easily the correlation is maximal if both B_1 and B_2 equal the same deterministic number (as then both queues evolve 'synchronically').

The rationale behind studying fork-join systems of the type described above lies in the fact that they are a natural model for several relevant real-life systems, for instance in service systems, healthcare applications, manufacturing systems, and communication networks. With S_i denoting a job's sojourn time in queue i, a particularly interesting object is the *fork-join system's sojourn time* $S := \max\{S_1, S_2\}$. This sojourn time is relevant, as in many situations the job can be further processed only if service at both queues has been completed, which explains the terminology 'fork-join'. One could think of many specific examples in which fork-join systems (and the sojourn time S) play a crucial role, such as:

- a request for a mortgage is handled simultaneously by a loan division and a life insurance division of a bank; the mortgage request is finalized when the tasks at both divisions have been completed.
- a laboratorial request of several blood samples is handled simultaneously by several lab employees of a hospital; the patient's laboratorial report is finalized when all the blood samples have been analyzed.
- a computer code runs two routines in parallel; both should be completed in order to start a next routine.

We here remark that on a generic level, many service systems can be modeled as networks of queues, see for instance the process-flow-based modeling framework proposed in Kemper et al. (2010), of which fork-join systems can be an important building block.

M/G/1 fork-join systems have been studied intensively in the past, see for instance the overview article (Boxma et al. 1994) and the references therein, and have turned out to be notoriously hard to analyze. We now give a brief account of the literature, where we restrict ourselves to the papers that are relevant in the scope of our work.

In general, no explicit expressions are known for the joint steady-state workload distribution of both queues, nor for the mean sojourn time. For the specific case of an M/M/1 fork-join system, Flatto and Hahn (1984) derive the probability generating function of the joint queue-length (in terms of numbers of jobs), thus defining the steady-state probabilities p_{ij}, where i and j represent the number of jobs in the two queues. The asymptotics of this distribution are analyzed in Flatto (1985); these provide insight into the dependence between the two queues. For this M/M/1 fork-join system, under the additional assumption that the service times at both queues stem from the *same* exponential distribution, the mean sojourn time can be derived explicitly from the system's balance equations, see Nelson and Tantawi (1988), and obeys a simple closed-form expression. It is noted, however, that the underlying argument breaks down as soon as we depart from the exponentiality and homogeneity assumptions.

For the general M/G/1 fork-join system (and in fact for the GI/G/1 variant), upper and lower bounds on the mean sojourn time were derived by Baccelli and Makowski (1985), relying on stochastic comparison techniques; see also Baccelli et al. (1989). These bounds are not always easy to compute, as they require the availability of explicit expressions or accurate approximations of the distribution function of the workload in related single-node M/G/1 and D/G/1 queues. In addition, the bounds are in many cases quite far apart, as observed from the numerical results on the heterogeneous exponential case by Balsamo et al. (1998). In their paper, Balsamo et al. (1998) present considerably more accurate bounds, but their approach is restricted to the situation of heterogeneous exponential service times; also, their method is of relatively high computational complexity. An elegant approximation technique for the homogeneous case was proposed in Varma and Makowski (1994). In their work, special attention is paid to the impact of the number of servers operating in parallel (which we assume to be 2 throughout this paper). We finally note that results on the corresponding G/M/1 queue are given in Ko and Serfozo (2008).

The above literature overview underscores the need for accurate methods to approximate the mean sojourn time $\mathbb{E}S$ that work for a broad set of service-time distributions. In this paper we present a set of such approximations and heuristics, that are of low computational complexity, yet remarkably accurate. In more detail, our contributions are the following:

- We explicitly compute the upper bound of Baccelli and Makowski (1985) for a set of frequently used service-time distributions. We also note that the accompanying lower bound can be evaluated for a limited set of service-time distributions only.
- We systematically assess the homogeneous case (i.e., B_1 and B_2 having the same distribution, say that of a random variable B). The approach followed is the following.

- We first observe that in many situations, the bounds presented in Baccelli and Makowski (1985) are rather far apart (and sometimes even outperformed by trivial bounds).
- The approximations we develop are based on a two-moment characterization of the service times; after scaling the arrival intensity to 1, the only relevant parameters in the model are then the load ϱ and the squared coefficient of variation (SCV) of B. This approach essentially assumes an insensitivity: $\mathbb{E}S$ depends only on the first two moments of the service-time distribution. This claim is justified by simulation results (where we sample from various service-time distributions with the same first two moments, some of which have heavy tails). The reason why we restrict ourselves to two-moments-based approximations is that in the *single* M/G/1 queue the mean sojourn time, say m, depends on B through its first two moments only, due to the celebrated Pollaczek–Khinchine formula.
- We argue, based on theoretical as well as empirical arguments, that approximations of the type

$$\mathbb{E}S = \frac{3}{2}m,$$

  with m denoting the mean sojourn time in one of the individual queues, work surprisingly well for a broad set of parameters.
- We then refine this crude approximation to fits of the type

$$\frac{\mathbb{E}S}{m} \approx a(\varrho) + b(\varrho)\log \text{SCV},$$

  and

$$\frac{\mathbb{E}S}{m} \approx a(\varrho) + b(\varrho)\log \text{SCV} + c(\varrho)(\log \text{SCV})^2;$$

  particularly the latter type turns out to have an excellent fit.

- We briefly touch upon heterogeneous scenarios. If the loads of both queues are different, $\mathbb{E}S$ could be approximated by the mean sojourn time of the queue with the highest load. We assess under what conditions such a bottleneck approach works well.

Our semi-empirical approach, that was sketched above, may be considered as somewhat unconventional by the operations research (OR) community. The striking accuracy of the fit, however, makes our findings interesting and practically relevant. In addition, we hope that our empirical results trigger new research, so that they will eventually be justified by theoretical arguments. As an aside, we mention that empirically derived approximations constitute an important subject within OR—think for instance of the classical approximations (de Kok and Tijms 1985; Kühn 1979; Whitt 1983). Evidently, such approximations gain credibility when they are backed by theoretical justification (for instance if they are exact for certain special cases, or in certain asymptotic regimes).

The structure of the paper is as follows. In Sect. 2 we sketch the model, and present some preliminaries. We also review the bounds of Baccelli and Makowski (1985), and explicitly calculate them for specific service-time distributions. In Sect. 3 we consider the homogeneous case, i.e., $B_1 =_{\mathrm{d}} B_2$, and identify under which conditions the bounds of Baccelli and Makowski (1985) are far apart. We then present the approximations, which turn out to be highly accurate. The paper is concluded in Sect. 4 by a brief summary and discussion.

2 Model, preliminaries, and bounds

In this section we formally introduce the fork-join system (or: parallel queue), see Fig. 1. This system consists of two queues (or: workstations, nodes) that work in parallel. The jobs arrive according to a Poisson process with parameter λ; we renormalize time by setting $\lambda \equiv 1$ (we return to this issue later). Upon arrival the job *forks* into two different 'tasks' that are directed simultaneously to both workstations. The service times in workstation i (for $i = 1, 2$), which can be regarded as a *queue*, are an i.i.d. sequence of non-negative random quantities $(B_{i,n})_{n\in\mathbb{N}}$ (distributed as a generic random variable B_i); we also assume $(B_{1,n})_{n\in\mathbb{N}}$ and $(B_{2,n})_{n\in\mathbb{N}}$ to be mutually independent. As mentioned before, one could call the resulting queueing system an 'M/G/1 parallel queue'. In this system we denote by B_i the generic service time, and by S_i the stationary sojourn time of an arbitrary customer in queue i, for $i = 1, 2$ (where it is noted that S_1 and S_2 are *not* independent).

With $\lambda = 1$, the load of node i is defined as $\varrho_i := \lambda\mathbb{E}B_i \equiv \mathbb{E}B_i$. The systems stability is assured under the, intuitively obvious, condition $\max\{\varrho_1, \varrho_2\} < 1$, see Baccelli and Makowski (1985). Under the stability condition and with $\lambda = 1$, the Pollaczek–Khinchine mean formula for the mean sojourn time in each node, $m_i = \mathbb{E}S_i$, yields

$$\begin{aligned} m_i &= \frac{\lambda\mathbb{E}[B_i^2]}{2(1-\varrho_i)} + \mathbb{E}B_i \\ &= \frac{\varrho_i^2}{2(1-\varrho_i)}(\mathrm{SCV}_i + 1) + \varrho_i, \end{aligned} \tag{1}$$

see for instance Tijms (1986, Eq. (2.55)).

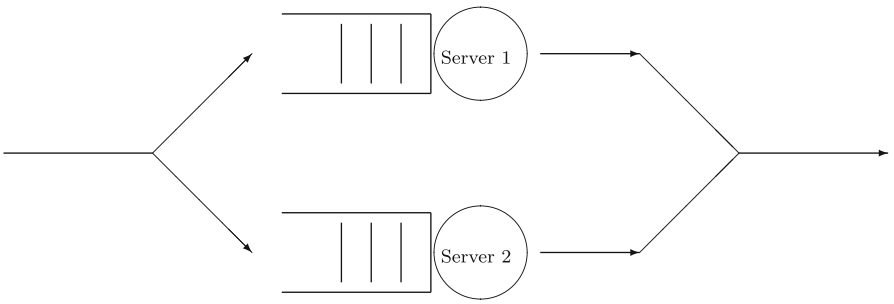

Fig. 1 A simple fork-join queue

In the sequel we will denote by SCV the *squared* coefficient of variation, defined by the ratio of the variance to the squared mean. Our approach is validated over large range of values for SCV. Note however that the SCV in most applications is in the range SCV $\in [0.5, 2]$, see for example Brown et al. (2005) and Cayirli and Veral (2003) and references therein.

Each queue handles the tasks in a first-come first-serve fashion. In other words: if the task finds the queue non-empty, it waits in the queue until service starts. When both tasks (that correspond to the same job) have been performed, they *join* and the job departs the network (thus explaining the terminology 'fork-join system'). Therefore, the total sojourn time of a job in the network is the *maximum* of the two individual sojourn times. The goal of this paper is to devise ways to approximate the *mean stationary sojourn time*, i.e.,

$$\mathbb{E}S = \mathbb{E}\left[\max\{S_1, S_2\}\right].$$

As mentioned above, without loss of generality, we may renormalize time such that $\lambda = 1$ (which we will do throughout this paper). Note that the general case $\lambda > 0$ can be derived from the special case $\lambda = 1$, since we have for $i = 1, 2$, in self-evident notation,

$$S_i(\lambda, B_i) =_{\mathrm{d}} \frac{S_i(1, \lambda B_i)}{\lambda},$$

so that

$$S(\lambda, B_1, B_2) =_{\mathrm{d}} \frac{S(1, \lambda B_1, \lambda B_2)}{\lambda}.$$

In general, the mean sojourn time cannot be explicitly calculated, the only exception being the case that B_1 and B_2 correspond to the same exponential distribution, as mentioned in the introduction. This result, by Nelson and Tantawi (1988), is recalled in Sect. 2.1. Relaxing the homogeneity and exponentiality assumptions, upper and lower bounds are known, which will be reviewed in Sect. 2.2, and made explicit in Sect. 2.3.

2.1 The homogeneous M/M/1 fork-join system

As proven by Nelson and Tantawi (1988), in case of two homogeneous servers with exponentially distributed service times, the mean sojourn time obeys the strikingly simple formula

$$\mathbb{E}S = \left(\frac{12 - \varrho}{8}\right) \cdot m,$$

where $m := \varrho/(1 - \varrho)$ is the mean sojourn time of a M/M/1 queue by virtue of (1).

Observe that, when increasing the load from 0 to 1, the ratio of the mean sojourn time $\mathbb{E}S$ to the mean sojourn time of a *single* workstation, i.e., $\mathbb{E}S/m$, varies just mildly: for $\varrho \uparrow 1$ it is $11/8 = 1.375$, whereas for $\varrho \downarrow 0$ it is $12/8 = 3/2 = 1.5$, i.e., about 8% difference. This entails that an approximation of the type $\mathbb{E}S \approx \frac{3}{2}m$ is conservative, yet quite accurate.

2.2 Bounds for the M/G/1 fork-join system

In this section we discuss a number of bounds on $\mathbb{E}S$ in an M/G/1 fork-join system. It is noted that they in fact apply to the GI/G/1 fork-join system, but under the assumption of Poisson arrivals often explicit computations are possible, see Sect. 2.3.

An upper and lower bound for the general GI/G/1 case are presented by Baccelli and Makowski (1985), see also Baccelli et al. (1989); in the sequel we refer to these bounds as the *BM bounds*. The idea behind these bounds is that the level of the variability of the fork-join system's waiting time should be increasing in the level of variability of the stochastic arrival process of the system. The BM bounds for the sojourn time are in fact sojourn times of related two-queue systems, but, importantly, in these systems the queues are *independent*:

- in the BM upper bound one does as if two queues are independent. Informally, by making the queues independent, the stochasticity increases, and therefore the mean of the maximum of $\mathbb{E}S_1$ and $\mathbb{E}S_2$ increases, and therefore this approach results in an upper bound.
- in the BM lower bound one considers two D/G/1 queues (with the same loads as in the original parallel queue). Informally, by assuming deterministic arrivals, one reduces the system's stochasticity, and therefore the mean of the maximum of $\mathbb{E}S_1$ and $\mathbb{E}S_2$ decreases, and therefore this approach results in a lower bound.

This intuitive reasoning leads to bounds, which are rigorously proven in Baccelli and Makowski (1985) and Baccelli et al. (1989). Below we discuss these BM bounds, and in addition also a number of trivial (but useful) bounds. Then we show how to compute these bounds explicitly in a number of practically relevant cases in Sect. 2.3.

2.2.1 Trivial bounds

We first present a trivial lower bound. Using that $x \mapsto \max\{0, x\}$ is a convex function and due to Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}S &= \mathbb{E}S_1 + \mathbb{E}\max\{0, S_2 - S_1\} \\ &\geq \mathbb{E}S_1 + \max\{0, \mathbb{E}(S_2 - S_1)\} = \max\{\mathbb{E}S_1, \mathbb{E}S_2\} =: \ell.\end{aligned}$$

Since $\max\{a, b\} = a + b - \min\{a, b\} \leq a + b$, we also have the upper bound

$$\mathbb{E}S \leq \mathbb{E}S_1 + \mathbb{E}S_2 =: u.$$

Notice that these bounds are in some sense *insensitive*, as they depend on the distribution of S_1 and S_2 only through their respective means.

2.2.2 BM bounds

The BM bounds for the GI/G/1 parallel queue are 'explicit' in the sense that they reduce to standard formulas in terms of the distribution of the sojourn times of *single* GI/G/1 systems for the upper bound, and *single* D/G/1 systems for the lower bound (with the same load as the original system). Recall that the stability of these systems is ensured if $\lambda\,\mathbb{E}B_i < 1$ for both $i = 1$ and 2, which is identical to the stability condition of our fork-join system. The bounds, as established in Baccelli and Makowski (1985) and Baccelli et al. (1989), are then as follows.

Upper bound. We do as if the queues are actually independent, that is, fed by independent processes (but identical in law). As a consequence, S_1 and S_2 are independent as well; call the maximum of S_1 and S_2 under this assumption $\bar{S}$. Then it is elementary that, in self-evident notation, $\mathbb{E}\bar{S}$ equals

$$\mathbb{E}\bar{S} = \int_0^\infty y\mathbb{P}_{\mathrm{GI/G/1}}(S_1 \le y)\mathrm{d}\mathbb{P}_{\mathrm{GI/G/1}}(S_2 \le y) + \int_0^\infty x\mathbb{P}_{\mathrm{GI/G/1}}(S_2 \le x)\mathrm{d}\mathbb{P}_{\mathrm{GI/G/1}}(S_1 \le x) =: U.$$

Lower bound. Now we do as if both queues are fed by deterministic arrival processes. Call the maximum of S_1 and S_2 under this assumption $\underline{S}$. Then

$$\mathbb{E}\underline{S} = \int_0^\infty y\mathbb{P}_{\mathrm{D/G/1}}(S_1 \le y)\mathrm{d}\mathbb{P}_{\mathrm{D/G/1}}(S_2 \le y) + \int_0^\infty x\mathbb{P}_{\mathrm{D/G/1}}(S_2 \le x)\mathrm{d}\mathbb{P}_{\mathrm{D/G/1}}(S_1 \le x) =: L.$$

It thus holds that

$$\max\{\ell, L\} \le \mathbb{E}S \le U \le u.$$

In our numerical experiments, in Sect. 3, we have included the trivial bounds ℓ and u to offer a comprehensive view. In addition, we will show that in many situations the trivial lower bound ℓ is actually tighter than the BM lower bound L, and, given the computational advantages, one could consider ℓ and u as an approximation instead of L and U. Also, the BM bounds L and U cannot be explicitly computed for all M/G/1 fork-join systems. Hence, if they have to be determined numerically, then their advantage over estimating $\mathbb{E}S$ by simulation is unclear (in Sect. 2.3 we present a few examples in which U and L *can* be computed, though).

As a final remark, we mention that if m_1 is considerably larger than m_2 (i.e., ϱ_1 considerably larger than ϱ_2), then $\mathbb{E}S \approx m_1$. This is proven for the M/G/1 case as

follows. Suppose the load of the second queue is $\epsilon < 1$. Then, relying on (1),

$$m_1 \leq \mathbb{E}S \leq m_1 + m_2 = m_1 + \frac{\epsilon^2}{2(1-\epsilon)}(\text{SCV}_2 + 1) + \epsilon,$$

so that $\mathbb{E}S \to m_1$ as $\epsilon \downarrow 0$. This indicates that, if the loads of both queues are highly asymmetric, the bottleneck queue essentially determines the parallel queue's sojourn time.

2.3 BM bounds for specific M/G/1 fork-join systems

We now present a number of explicit expressions for the bounds u, U, ℓ, and L in the case of Poisson arrivals and various service-time distributions. In Sect. 3 we approximate the service-time distribution by a so-called *phase-type distribution* (with appropriate mean and variance), and therefore we focus on a number of phase-type service-time distributions, viz. exponential service times, Erlang service times (useful to approximate service times with coefficient of variation smaller than 1), and hyper-exponential times (useful to approximate service times with coefficient of variation larger than 1). The use of phase-type distributions make models tractable, but one can also view them as a semi-parametric density, see Asmussen et al. (1996). The sensitivity of the approach with respect to the service-time distribution is discussed in Sect. 3.

M/M/1 case. Here we let the service times in both queues be exponentially distributed, with means ϱ_1 and ϱ_2 respectively; recall that the exponential distribution has SCV equal to 1. From (1) follows that S_i has an exponential distribution with mean $m_i := \varrho_i/(1-\varrho_i)$. Trivially,

$$\ell = \max\{m_1, m_2\}, \quad u = m_1 + m_2.$$

It is now a trivial computation to show that

$$U = m_1 + m_2 - \left(\frac{1}{m_1} + \frac{1}{m_2}\right)^{-1}.$$

In case of deterministic arrivals it is known that S_i has an exponential distribution (in fact any G/M/1 leads to an exponential distribution). Its mean, that is $\mathbb{E}S_i$, reads $\kappa_i := \varrho_i/(1-\omega_i)$, where ω_i is the unique solution to $\omega_i = e^{-(1-\omega_i)/\varrho_i}$, with $0 < \omega_i < 1$. Then computing the integrals yields

$$L = \kappa_1 + \kappa_2 - \left(\frac{1}{\kappa_1} + \frac{1}{\kappa_2}\right)^{-1}.$$

M/E$_2$/1 case. We now consider the case of the service times having an Erlang distribution with two phases. Random variables with an Erlang distribution are known to

be 'less variable' than the exponential distribution; more precisely, an Erlang distribution consisting of k phases has a SCV of $1/k$. In case $k = 2$, these two exponential phases have mean length $\varrho_i/2 = 1/\mu_i$. Using elementary queueing theory, it is readily checked that the Laplace transforms of the sojourn times read, for $i = 1, 2$,

$$\bar{S}_i(s) = \frac{(1-\varrho_i)\mu_i^2}{s^2 + s(2\mu_i - 1) + \mu_i(\mu_i - 2)}.$$

Applying a partial fraction expansion, with $s_{\pm,i}$ denoting the zeros of the denominator

$$s_{\pm,i} := \frac{1}{2}\left(1 - 2\mu_i \pm \sqrt{4\mu_i + 1}\right),$$

and

$$\alpha_{1i} := \frac{s_{-,i}}{s_{-,i} - s_{+,i}}, \quad \alpha_{2i} := -\frac{s_{+,i}}{s_{-,i} - s_{+,i}},$$

leads to

$$\mathbb{P}(S_i \le x) = \alpha_{1i}(1 - \exp(s_{+,i}x)) + \alpha_{2i}(1 - \exp(s_{-,i}x)). \tag{2}$$

This result enables us to evaluate the upper bound U. Tedious computations eventually lead to

$$U = m_1 + m_2 + \frac{1}{(s_{-,1} - s_{+,1})(s_{-,2} - s_{+,2})} \times \left(\frac{s_{+,1}s_{+,2}}{(s_{-,1} + s_{-,2})} - \frac{s_{-,1}s_{+,2}}{(s_{+,1} + s_{-,2})} - \frac{s_{+,1}s_{-,2}}{(s_{-,1} + s_{+,2})} + \frac{s_{-,1}s_{-,2}}{(s_{+,1} + s_{+,2})}\right),$$

where m_i is the mean sojourn time in queue i, which in this case reduces to $\varrho_i(4 - \varrho_i)/(4 - 4\varrho_i)$. The lower bound L is based on $\mathbb{P}(S_i \le x)$ for a D/E$_2$/1 queue, for which no explicit form is known, to the best of our knowledge.

M/E$_{1,2}$/1 case. Let us consider the situation of the service times being 'generalized Erlang', see Tijms (1986, p. 398). More specifically, we consider a mixture of an E$_1$ and an E$_2$ with the *same* scale parameters, which is denoted as an E$_{1,2}$. We here choose the parameters such that the SCV of the service time is $\frac{3}{4}$. This is done by choosing for B_i with probability p_i an exponential distribution with mean $1/\mu_i$, and with probability $1 - p_i$ an E$_2$ distribution with mean $2/\mu_i$. For given ϱ_i and SCV, the parameters p_i and μ_i are uniquely defined, see Tijms (1986, Eq. (A.14)). Standard queueing theory then yields the Laplace transforms of the sojourn times, for $i = 1, 2$,

$$\bar{S}_i(s) = \frac{(1-\varrho_i)(\mu_i^2 + p_i\mu_i s)}{s^2 + s(2\mu_i - 1) + \mu_i(\mu_i + p_i - 2)}.$$

With $s_{\pm,i}$ be the zeros of the denominator, that is,

$$s_{\pm,i} := \frac{1}{2}\left(1 - 2\mu_i \pm \sqrt{4(1-p_i)\mu_i + 1}\right), \tag{3}$$

and

$$\alpha_{1i} := \frac{s_{-,i} + p_i(\mu_i - 2 + p_i)}{s_{-,i} - s_{+,i}}, \quad \alpha_{2i} := 1 - \alpha_{1i}, \tag{4}$$

Equation (2) again applies, but now with $s_{\pm,i}$ given through (3) and α_{ji} through (4). S_i has a $\mathrm{E}_{1,2}$ distribution with mean given through (1). It can then be shown that

$$U = m_1 + m_2 + \frac{\alpha_{11}\alpha_{12}}{s_{+,1} + s_{+,2}} + \frac{\alpha_{21}\alpha_{12}}{s_{-,1} + s_{+,2}} + \frac{\alpha_{11}\alpha_{22}}{s_{+,1} + s_{-,2}} + \frac{\alpha_{21}\alpha_{22}}{s_{-,1} + s_{-,2}}. \tag{5}$$

The lower bound L is based on $\mathbb{P}(S_i \le x)$ for a $\mathrm{D/E}_{1,2}/1$ queue, for which no explicit form is known, to our best knowledge.

M/H$_2$/1 case. Above we concentrated on service times with SCV smaller than 1; we now consider the case of SCVs larger than 1. A hyperexponentially distributed random variable B_i now results from sampling from an exponential distribution with mean $1/\mu_{i1}$ with probability p_i, and from an exponential distribution with mean $1/\mu_{i2}$ with probability $1 - p_i$. We fix the mean service times, leading to the requirement

$$\varrho_i = \frac{p_i}{\mu_{i1}} + \frac{1 - p_i}{\mu_{i2}}.$$

Under the additional condition of 'balanced means' Tijms (1986, Eq. (A.16)), one imposes $\mu_{i1} = 2p_i\mu_i$ and $\mu_{i2} = 2(1 - p_i)\mu_i$, and with fixed SCVs this leads to

$$\mathrm{SCV}_i := \frac{\mathbb{V}\mathrm{ar}\, B_i}{(\mathbb{E}B_i)^2} = \frac{1}{2p_i(1-p_i)} - 1 \quad \Rightarrow \quad p_i = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\mathrm{SCV}_i - 1}{\mathrm{SCV}_i + 1}}.$$

It is obvious that we again have that S_i has mean as in (1), with the SCVs given in the previous display. For $i = 1, 2$ we find, as before, the Laplace transforms of the sojourn times:

$$\bar{S}_i(s) = \frac{4p_i(1-p_i)(\mu_i^2 - \mu_i) + 2s(p_i^2 + (1-p_i)^2)(\mu_i - 1)}{s^2 + s(2\mu_i - 1) + 4p_i(1-p_i)(\mu_i^2 - \mu_i)}.$$

With $s_{\pm,i}$ denoting the zeros of the denominator, i.e.,

$$s_{\pm,i} = \frac{1}{2}\left(1 - 2\mu_i \pm \sqrt{1 - 4\frac{\mathrm{SCV}_i - 1}{\mathrm{SCV}_i + 1}\mu_i + 4\frac{\mathrm{SCV}_i - 1}{\mathrm{SCV}_i + 1}\mu_i^2}\right), \tag{6}$$

and

$$\alpha_{1i} := \frac{1}{2} + \frac{\frac{1}{2} + \frac{\mathrm{SCV}_i - 1}{\mathrm{SCV}_i + 1}(1 - \mu_i)}{\sqrt{1 - 4\frac{\mathrm{SCV}_i - 1}{\mathrm{SCV}_i + 1}\mu_i + 4\frac{\mathrm{SCV}_i - 1}{\mathrm{SCV}_i + 1}\mu_i^2}}, \quad \alpha_{2i} = 1 - \alpha_{1i}, \tag{7}$$

it follows that Eqs. (2) and (5) again apply, but now with $s_{\pm,i}$ given through (6) and α_{ji} through (7). The lower bound L requires knowledge of $\mathbb{P}(S_i \leq x)$ for a D/H_2/1 queue, for which no explicit expression is available.

3 The homogeneous case

In this section we consider the situation of *homogeneous* servers, i.e., B_1 and B_2 are (independently) sampled from the same distribution. As shown by Nelson and Tantawi (1988), the mean sojourn time in case of homogeneous exponentially distributed service times is a simple function of the mean sojourn time of a single queue, say m, and the service load, ϱ, see Sect. 2.1; for other service times, however, no explicit results are known. In this section we assess the accuracy of the bounds u, ℓ, U, and L, by systematic comparison with simulation results. We do this by varying the load ϱ (equal for both queues) imposed on the system, as well as the 'variability' of the service times (in terms of the SCV).

Our analysis indicates that for a substantial set of model instances the upper and lower bounds are far apart, and therefore we have attempted to develop more accurate approximations. We empirically find an approximation with a nearly perfect fit, which gives us the mean sojourn time as a function of the load and SCV. An important by-product of the analysis performed in this section, is a number of explicit expressions for the bounds, for a set of practically relevant service-time distributions (e.g., Erlang and hyperexponential); it is noted that the trivial bounds u and ℓ reduce to $2m$ and m, respectively, in case of homogeneity. Our results once again clearly reveal that the effect of the system's service load ϱ is modest, as was already observed by Nelson and Tantawi (1988) for the case of exponentially distributed service times.

M/M/1 case. As mentioned earlier, in the symmetric case when $m = m_1 = m_2 = \varrho/(1-\varrho)$, the mean sojourn time is explicitly known: $\mathbb{E}S = m \cdot (12 - \varrho)/8$, see Nelson and Tantawi (1988). Also, it is easily seen from the results in Sect. 2 that

$$U = \frac{3}{2} \cdot m;$$

notably, this fraction $\frac{3}{2}$ is insensitive with respect to the load ϱ. The upper bound U is close to the mean sojourn time $\mathbb{E}S$ for small ϱ; one must, however, bear in mind that this scenario is perhaps not so realistic in practice. Also,

$$L = \frac{3}{2} \cdot \kappa,$$

Table 1 Simulated sojourn times and the corresponding BM bounds

ϱ	SCV	m	$\mathbb{E}S$	α(SCV)	U	α_U(SCV)	L	α_L(SCV)	BM-Spread (%)
0.1	0.25	0.1069	0.1357	1.2690	0.1375	1.2861	0.1273	1.1908	7.55
	0.33	0.1074	0.1403	1.3070	0.1421	1.3227	0.1313	1.2220	7.68
	0.5	0.1083	0.1482	1.3676	0.1497	1.3819	0.1375	1.2693	8.23
	0.75	0.1097	0.1580	1.4401	0.1594	1.4531	0.1452	1.3230	9.03
	1	0.1111	0.1653	1.4875	0.1667	1.5003	0.1500	1.3501	10.10
	2	0.1167	0.1842	1.5792	0.1855	1.5902	0.1596	1.3681	14.06
	4	0.1278	0.2126	1.6634	0.2138	1.6730	0.1762	1.3787	17.67
	16	0.1944	0.3509	1.8048	0.3520	1.8105	0.2985	1.5350	15.26
	64	0.4611	0.8790	1.9062	0.8804	1.9093	0.8215	1.7815	6.70
	256	1.5278	2.9833	1.9527	2.9862	1.9546	2.9247	1.9143	2.06
0.9	0.25	5.9600	7.4225	1.2449	8.7203	1.4625	2.3497	0.3941	85.83
	0.33	6.3000	8.0219	1.2733	9.2529	1.4687	2.8561	0.4534	79.74
	0.5	6.9750	9.1751	1.3154	10.3173	1.4792	3.8797	0.5562	70.16
	0.75	7.9875	10.8374	1.3568	11.9037	1.4903	5.4102	0.6773	59.92
	1	9.0000	12.4875	1.3875	13.5000	1.5000	6.9912*	0.7768	52.12
	2	13.050	19.0620	1.4607	19.9568	1.5293	13.4624	1.0316	34.07
	4	21.150	32.0373	1.5148	32.8541	1.5534	26.3568	1.2462	20.28
	16	69.750	109.3820	1.5682	110.1838	1.5797	103.6263	1.4857	6.00
	64	264.15	418.1811	1.5831	419.4601	1.5880	412.2813	1.5608	1.72
	256	1041.75	1650.0856	1.5840	1656.5520	1.5902	1636.7130	1.5711	1.20

with κ the mean sojourn time of a single D/M/1 queue with appropriate load. We will see later on in this section, in Table 1, that U and L substantially differ from the 'real' (i.e., simulated) mean sojourn time.

M/E$_2$/1 case. We consider the case that SCV $= \frac{1}{2}$. Straightforward computations yield

$$U = 2m + \frac{(\mu-1)(-5\mu+1)}{2\mu(\mu-2)(2\mu-1)} = m\frac{11\mu^2-10\mu+3}{8\mu^2-8\mu+2} = m\frac{3\varrho^2-20\varrho+44}{2(\varrho-4)^2}.$$

The fraction clearly is sensitive to the service load ϱ. For a system with small load $\varrho \downarrow 0$ gives $U \approx \frac{11}{8}m = 1.375m$, and for a system with large load $\varrho \uparrow 1$ gives $U \approx \frac{3}{2}m = 1.5m$. This once more implies that a conservative approximation can be of the type $\mathbb{E}S \approx \frac{3}{2}m$.

M/E$_{1,2}$/1 case. We now consider service times following a generalized Erlang distribution with SCV $= \frac{3}{4}$. In this symmetric case straightforward calculus yields, with $s_\pm \equiv s_{\pm,i}$ given by (3) and $\alpha_j \equiv \alpha_{ji}$ by (4), for $i = 1, 2$,

$$U = 2m + \frac{\alpha_1^2}{2s_+} + \frac{2\alpha_1\alpha_2}{1-2\mu} + \frac{\alpha_2^2}{2s_-}, \tag{8}$$

where we have used that $s_- + s_+ = 1 - 2\mu$. It can be seen that the ratio of U to m is sensitive to the service load ϱ. For a system with a small load, $\varrho = 0.1$, we have $U \approx 1.45m$, whereas for a system with large load, $\varrho = 0.9$, we have $U \approx 1.49m$. Again, a conservative approximation can be of type $\mathbb{E}S \approx \frac{3}{2}m$.

M/H$_2$/1 case. We again obtain (8), but now with $s_{\pm,i}$ given through (6) and α_{ij} through (7). Again the ratio of U to m is sensitive to the service load ϱ. For a system with SCV $= 2$ and a small load, $\varrho = 0.1$, we find $U \approx 1.59m$, whereas for a system with large load, $\varrho = 0.9$, it holds that $U \approx 1.53m$; for a system with SCV $= 4$ and small load, $\varrho = 0.1$, we have $U \approx 1.89m$, whereas for a system with large load, $\varrho = 0.9$, we have $U \approx 1.55m$.

Observe that the ratio of U to m is close to $\frac{3}{2}$ in the (perhaps most relevant) situation that the load is relatively high, that is, for loads ϱ higher than, say, 0.9.

The lower bound L cannot be given in closed-form, except in the M/M/1 case, but can of course be determined through simulation. We now verify the accuracy of the bounds L and U, see Table 1. We concentrate on two 'extreme' loads (0.1 and 0.9), and we vary the SCV. The table should be read as follows. The upper part is on the case $\varrho = 0.1$, while the lower part relates to $\varrho = 0.9$. Then we provide, for several values of the SCV:

(i) The mean sojourn time m of a single queue. For this we have exact expressions, see (1).
(ii) The mean sojourn time $\mathbb{E}S$ of the parallel queue. We have an exact expression for this for SCV $= 1$, and for the other SCVs we obtained a value through simulation.
(iii) The ratio of $\mathbb{E}S$ to m, which we call α(SCV). In view of the trivial bounds, it is clear that α lies between 1 and 2.
(iv) The upper bound U, using the expressions derived earlier in this section.
(v) The ratio of U to m, denoted by α_U(SCV).
(vi) The lower bound L, obtained through simulation (for SCV $= 1$ the corresponding phase-type distribution is the exponential distribution, for which we have an exact expression; Nelson and Tantawi 1988).
(vii) The ratio of L to m, denoted by α_L(SCV).
(viii) The 'BM-spread', that is, the ratio of $(U - L)$ to $\mathbb{E}S$.

The service times with SCV equal to 0.25 and 0.33 are obtained by using E_4 and E_3 distributions, respectively. For SCVs larger than 1 we use hyperexponential distribution, with the additional condition of 'balanced means' (Tijms 1986, Eq. (A.16)). In this table we used explicit formulae where possible; we otherwise relied on simulation. Here and in the sequel, the spread of the 95% confidence intervals for the simulated mean sojourn times is less than 0.5%.

The main conclusions from this table (and additional numerical experimentation, on which we do not report here) are the following:

– For low loads, i.e., $\varrho = 0.1$, the bounds L and U are relatively close, the difference can be substantial for values of SCVs between 1 and 16.
– For high loads, i.e., $\varrho = 0.9$, L and U tend to be far apart, particularly for low SCVs.
– In several cases, the lower bound L is even below the trivial lower bound $\ell = m$. It is readily checked that this effect is not ruled out in the construction of the lower bound L.

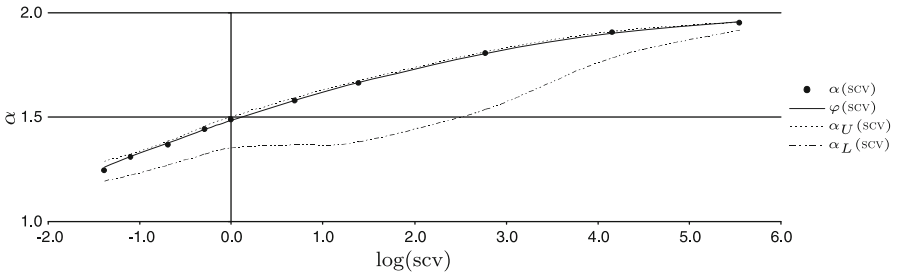

Fig. 2 Graph with BM bounds, simulated values, and approximated values for load $\varrho = 0.1$

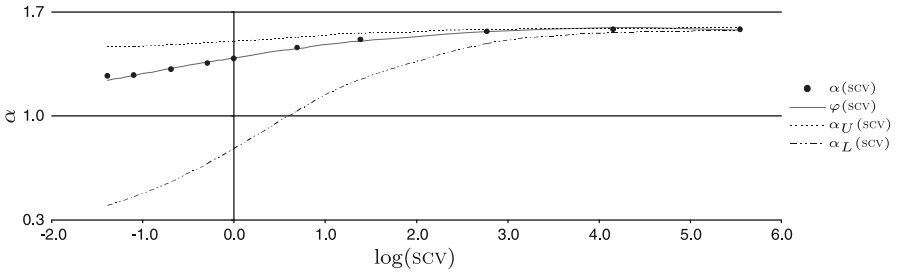

Fig. 3 Graph with BM bounds, simulated values, and approximated values for load $\varrho = 0.9$

– A disadvantage of relying on these bounds is that particularly L is in most cases not known in closed-form. It therefore needs to be obtained by simulation, but then there is no advantage of using this bound anymore: with comparable effort we could have simulated the parallel queue itself as well.

In view of the tables presented above and illustrated in Figs. 2 and 3, there is a clear need for more accurate bounds and/or approximations. The approach followed here is to identify, for any given value of the load ϱ, an elementary function $\varphi(\cdot)$, such that $\varphi(\text{SCV})$ accurately approximates $\alpha(\text{SCV})$. In this approach we parameterize the service-time distribution by its mean and SCV. The underlying idea is that in a single M/G/1 queueing system the mean sojourn time solely depends on its first two moments, as it can be expressed as a function of its mean service time and coefficient of variation through the Pollaczek–Khinchine formula, see for example Tijms (1986, Eq. (2.55)). We expect the mean sojourn time of the parallel queueing system to exhibit (by approximation) similar characteristics, thus justifying the approach followed. Having a suitable function $\varphi(\cdot)$ at our disposal, we can estimate $\mathbb{E}S$ by $m \cdot \varphi(\text{SCV})$. The function $\varphi(\cdot)$ shown in Figs. 2 and 3 refers to the one that will be proposed in the left panel of Table 4.

(*Approximate*) *insensitivity.* In the approach described above, we assume that $\mathbb{E}S$ is (approximately) insensitive, in that it depends on the first two moments of the service-time distribution only. We verified this property by comparing $\mathbb{E}S$ for two different distributions of the service times with identical first and second moments. Table 2 gives a representative illustration of our findings. There we compare the ratio $\alpha(\text{SCV})$ of the phase-type service-time distribution with the ratio $\alpha(\text{SCV})$ of the Weibull service-time distribution.

Table 2 Simulated sojourn times and the corresponding $\alpha(\text{SCV})$ s for phase-type and Weibull service-time distributions

ϱ	SCV	m	$\mathbb{E}S$	$\alpha(\text{SCV})$	$\mathbb{E}S_{\text{W}}$	$\alpha(\text{SCV})_{\text{W}}$
0.1	0.25	0.1069	0.1357	1.2690	0.1363	1.2749
	0.33	0.1074	0.1403	1.3070	0.1411	1.3135
	0.5	0.1083	0.1482	1.3676	0.1488	1.3737
	0.75	0.1097	0.1580	1.4401	0.1579	1.4392
	1	0.1111	0.1653	1.4875	0.1653	1.4875
	2	0.1167	0.1842	1.5792	0.1871	1.6037
	4	0.1278	0.2126	1.6634	0.2184	1.7092
	16	0.1944	0.3509	1.8048	0.3627	1.8651
	64	0.4611	0.8790	1.9062	0.8965	1.9448
	256	1.5278	2.9833	1.9527	3.0227	1.9727
0.9	0.25	5.96	7.4225	1.2449	7.4117	1.2431
	0.33	6.30	8.0219	1.2733	8.0110	1.2715
	0.5	6.98	9.1751	1.3154	9.1639	1.3138
	0.75	7.99	10.8374	1.3568	10.8412	1.3572
	1	9.00	12.4875	1.3875	12.4848	1.3874
	2	13.05	19.0620	1.4607	18.9871	1.4549
	4	21.15	32.0373	1.5148	31.9305	1.5100
	16	69.75	109.3820	1.5682	110.4690	1.5836
	64	264.15	418.1811	1.5831	430.3272	1.6318
	256	1041.75	1650.0856	1.5840	1729.6191	1.6684

In our approach we took phase-type distributions, in the way we explained above: Erlang for SCV smaller than 1 and balanced-means hyperexponential for SCV larger than 1. For values of SCV up to 1, the corresponding Weibull distribution has a shape-parameter larger than 1, meaning that all moments exist and that even the moment generating function (mgf) is finite for some positive arguments—we could then call these distributions 'light tailed'. For larger values of the SCV, however, the shape parameter will lie between 0 and 1, and then the Weibull distribution could be called heavy-tailed: although all moments exist, the moment generating function does not (for any positive argument). For instance for SCV equal to 16 (256) the shape parameter of the Weibull distribution has value 0.35 (0.20, respectively). It is noted that Weibull tails are not as heavy as Pareto tails, but our findings obviously provide support for our operational claim of approximate insensitivity.

The table should be read as follows. The upper part is on $\varrho = 0.1$, while the lower part relates to $\varrho = 0.9$. Then we provide, for a range of values of SCV, the mean sojourn time $\mathbb{E}S$ and the corresponding $\alpha(\text{SCV})$ for the service times having a phase-type distribution, as well as their counterparts $\mathbb{E}S_{\text{W}}$ and the corresponding $\alpha(\text{SCV})_{\text{W}}$ in case of Weibullian service times. The main conclusions from our experiments are the following. For $\varrho = 0.1$ and $\text{SCV} < 1$, we observe that $\mathbb{E}S$ and $\alpha(\text{SCV})$ are nearly equal to their Weibullian counterparts; for $\text{SCV} > 1$ the difference is modest, that is, up to 3.5%. For $\varrho = 0.9$ the fit is accurate up to $\text{SCV} = 4$, whereas for $\text{SCV} > 4$ the difference is modest, about 5%. The results of other numerical experiments give the same impression. These findings justify our two-moment approach.

Numerical experiments. Now that we have justified the use of phase-type distributions, we proceed as follows. To estimate $\alpha(\text{SCV}) = \mathbb{E}S/m$ for various values of SCV

Table 3 Simulated values of $\alpha(\text{SCV})$ of several SCVs and several loads ϱ

SCV	$\varrho = 0.1$	$\varrho = 0.3$	$\varrho = 0.5$	$\varrho = 0.7$	$\varrho = 0.9$
0.25	1.2690	1.2603	1.2523	1.2462	1.2449
0.33	1.3070	1.2961	1.2858	1.2773	1.2733
0.50	1.3676	1.3526	1.3381	1.3251	1.3154
0.75	1.4401	1.4170	1.3948	1.3650	1.3568
1.00	1.4874	1.4626	1.4374	1.4124	1.3875
2.00	1.5792	1.5662	1.5447	1.5114	1.4607
4.00	1.6634	1.6658	1.6423	1.5942	1.5148
16.0	1.8048	1.8155	1.7685	1.6886	1.5682
64.0	1.9062	1.8828	1.8143	1.7175	1.5831
256	1.9527	1.8999	1.8207	1.7217	1.5840

Table 4 Fitted ratios $\alpha(\text{SCV})$ for various loads ϱ based on least squares estimation

Load ϱ	$\varphi(\text{SCV})$	R^2 (%)	$\varphi(\text{SCV})$	R^2 (%)
0.1	$1.484 + 0.1461\log(\text{SCV}) - 0.01099\log(\text{SCV})^2$	100.00	$1.463 + 0.1031\log(\text{SCV})$	96.20
0.3	$1.476 + 0.1527\log(\text{SCV}) - 0.01344\log(\text{SCV})^2$	99.70	$1.451 + 0.1001\log(\text{SCV})$	93.80
0.5	$1.456 + 0.1448\log(\text{SCV}) - 0.01406\log(\text{SCV})^2$	99.50	$1.430 + 0.0898\log(\text{SCV})$	91.70
0.7	$1.427 + 0.1266\log(\text{SCV}) - 0.01323\log(\text{SCV})^2$	99.40	$1.403 + 0.07486\log(\text{SCV})$	89.70
0.9	$1.392 + 0.0950\log(\text{SCV}) - 0.01109\log(\text{SCV})^2$	99.60	$1.372 + 0.05158\log(\text{SCV})$	85.80

and ϱ, we performed simulation experiments, leading to the results shown in Table 3. The table indicates that a rule of thumb of the type $\mathbb{E}S \approx \frac{3}{2}m$ (that is $\alpha \approx \frac{3}{2}$) is a conservative, yet accurate approximation for a broad range of parameter values. We now try to identify a function $\varphi(\cdot)$ with a better fit.

In Table 3 we study the simulated ratios as function of the service-time distribution's SCV. We approximate the ratio $\alpha(\text{SCV})$ with a polynomial of $\log(\text{SCV})$ of degree two, based on 10 data points. The coefficients are estimated by applying ordinary least squares. The performance of the procedure is verified through the R^2, which is the coefficient of determination that indicates how well the model approximates the real data points, i.e., the goodness of fit of a model; see e.g., Stone (1996, Sect. 8.3) for a definition.

As can be seen in the left part of Table 4 and from Figs. 2 and 3, the polynomial regression fits extremely well, with an R^2 of nearly 100%. The table gives fitted curves for $\varrho = 0.1 + 0.2 \cdot i$, with $i = 0, \ldots, 4$. Our experiments indicate that for other values of ϱ, we are able to achieve good approximations by interpolating estimates for $\alpha(\text{SCV})$ linearly.

Note that one could also think of fitting a function of *both* SCV and ϱ (rather than fitting functions of just SCV, for various ϱ). However, it turned out that such a function does not perform significantly better than the interpolation-based approach described above.

We could also try to see how good a fit can be obtained by an even simpler function, for instance by approximating $\alpha(\text{SCV})$ by a polynomial of $\log(\text{SCV})$ of degree one. The results are reported in the rightmost column of Table 4. The model still shows a reasonable fit, but one observes that the R^2 for this polynomial regression analysis is decreasing in the load ϱ. Especially for larger values of ϱ the polynomial of degree one fits considerably worse than the polynomial of degree two.

We conclude this section with a few words on the approximation approach proposed by Varma and Makowski (1994). Their idea is to interpolate heavy- and light-load results to expressions for arbitrary load. The results show a good fit, and the procedures are of modest numerical complexity. In our paper, we took an alternative approach, relying on (i) a two-moment parameterization of the service times (and replacing them by their phase-type counterpart), (ii) an (empirically derived) approximation with a nearly perfect fit. Our approach requires negligible computational effort, and can therefore be used as an easily applicable engineering heuristic.

It is first noted that their approach gives expressions that are in line with limiting results for heavy and light loads. Compared to our approach, our empirically derived approximation is perhaps slightly easier to work with—as in the approach of Varma and Makowski (1994) the approximation needs to be determined case-by-case (see their Examples 1–4 in Sect. 6). The resulting approximation is very accurate (but less accurate than our approximation of Table 4). For instance in case the SCV equals 1/2 (Erlang-2 services), the Varma–Makowski algorithm gives 0.1481 for $\varrho = 0.1$ and 9 for $\varrho = 0.9$, where our simulated values were 0.1482 and 9.1751, respectively.

Heterogeneous case. We end this section with a few short remarks on the heterogeneous case; a detailed account can be found in Kemper and Mandjes (2009). First, two basic observations are in place: (i) in order to obtain a conservative estimate of $\mathbb{E}S$, we can replace the service-time distribution of the most lightly loaded queue by the service-time distribution of the other queue, so that we obtain a homogeneous system to which the theory developed in the previous section applies; (ii) as mentioned above, if one of the queues has a substantially higher load than the other one, one expects that the mean sojourn time of the queue with the heaviest load yields a good approximation for $\mathbb{E}S$.

As in the previous section, our findings are based on the typical phase-type service distributions, namely Erlang-2, exponential, and hyperexponential. As before, we analyze the ratio $\alpha(\text{SCV}) = \mathbb{E}S/m$, where m is now the mean sojourn time of the bottleneck queue (that is, the queue with the heaviest load). From the experiments above a few, more general, conclusions can be drawn:

- Restricting ourselves to cases with SCV ≤ 4 (which is quite realistic in most applications), a rule of thumb of the type $1.10 \cdot m$ always yields a conservative estimate for the system's mean sojourn time $\mathbb{E}S$ for heterogeneity level $b \in (0.1, 0.7)$ and loads $\varrho_1 \in [0.8, 0.9]$; here b is such that $\varrho_2 = b\varrho_1$.
- Similarly, for the same range of SCVs, but b smaller than 0.3 and all $\varrho_1 \leq 0.9$, the same statement applies.
- In all other situations, replacing the service-time distribution of the most lightly loaded queue by the service-time distribution of the other queue yields a conservative estimate; for the resulting homogeneous system the theory developed in the previous section applies.

4 Concluding remarks

The fork-join queue is an important generic building block of more complex service systems in manufacturing, services, and healthcare. The fact that the analysis of these

systems has proven to be highly complex, even in the very simple case of just two servers, is indisputably true. This makes the analysis challenging, and explains the need for simple heuristics.

This paper first discussed the bounds suggested by Baccelli and Makowski (1985). Then these bounds were numerically assessed for the homogeneous parallel queue (i.e., the service times at both queues have the same distribution). As they performed poorly, we developed an alternative approach: we identified a suitable function of the first two moments of the service-time distribution to estimate the mean sojourn time of the homogeneous parallel queue. Finally, we briefly commented on the heterogeneous parallel queue by giving several practical approximation guidelines.

In more detail, the conclusions are as follows:

- A trivial lower on the fork-join queue's mean sojourn time is evidently the largest of the individual mean sojourn times, $\ell := \max\{\mathbb{E}S_1, \mathbb{E}S_2\}$, and an upper bound is the sum of the two mean sojourn times, $u := \mathbb{E}S_1 + \mathbb{E}S_2$.
- Using standard queueing-theoretic methods, we derive explicit expressions for the upper bound developed in Baccelli and Makowski (1985). We do so for various phase-type service-time distributions. The lower bound suggested by Baccelli and Makowski (1985), however, can only be evaluated through simulation for almost all service-time distributions. We stress that when doing so there is no advantage of using this bound anymore: with comparable effort we could have simulated the fork-join system itself as well.
- For a substantial part of the parameter space both bounds from Baccelli and Makowski (1985) are highly inaccurate. In some cases their lower bound is even outperformed by the trivial lower bound.
- In the *homogeneous* case, the ratio of the mean sojourn time $\mathbb{E}S$ in the fork-join system to the mean sojourn time m of a single queue depends *approximately* only on the distribution of the service times mainly through the first two moments, or equivalently, the load ϱ, and the SCV of the service times. This legitimates our approach to express $\mathbb{E}S$ as a function of ϱ and SCV. The resulting function has a nearly perfect fit.
- In case of two *heterogeneous* queues in the parallel queueing system, we identified situations in which $\mathbb{E}S$ is close to the mean sojourn time of the queue with the highest load (the 'bottleneck'). In all other situations, we showed how to conservatively approximate $\mathbb{E}S$ by the mean sojourn time of a suitable homogeneous fork-join system, to which the theory mentioned above applies (see previous bullet).

Possible directions for future research include:

- To what extent is the mean sojourn time of the fork-join system insensitive with respect to higher moments of the service-time distribution?
- The study on the effect of heterogeneity can be extended, for instance by considering scenarios in which the service times stem from two entirely different distributions (e.g., exponentially distributed service times in queue 1, and E_2 service times in queue 2).

Acknowledgments The authors would like to thank the editor and the referees for their useful and valuable comments.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Asmussen S, Nerman O, Olssen M (1996) Fitting phase-type distributions via the EM algorithm. Scand J Stat 23(4):419–441

Baccelli F, Makowski AM (1985) Simple computable bounds for the fork-join queue. In: Proceedings of Johns Hopkins conference information science. Johns Hopkins University, Baltimore

Baccelli F, Makowski AM, Shwartz A (1989) The fork-join queue and related systems with synchronization constraints. Adv Appl Probab 21:629–660

Balsamo S, Donatiello L, van Dijk NM (1998) Bound performance models of heterogeneous parallel processing systems. IEEE Trans Parallel Distrib Syst 9:1041–1056

Boxma O, Koole G, Liu Z (1994) Queueing-theoretic solution methods for models of parallel distributed systems. In: Performance evaluation of parallel and distributed systems. CWI Tract 105, Amsterdam, pp 1–24

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: a queueing-science perspective. J Am Stat Assoc 100(469):36–50

Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. Prod Oper Manag 12(4):519–549

de Kok A, Tijms H (1985) A two-moment approximation for a buffer design problem requiring a small rejection probability. Perform Eval 5:77–84

Flatto L (1985) Two parallel queue created by arrivals with two demands II. SIAM J Appl Math 45:861–878

Flatto L, Hahn S (1984) Two parallel queue created by arrivals with two demands I. SIAM J Appl Math 44:1041–1053

Kemper B, Mandjes M (2009) Approximations for the mean sojourn time in a parallel queue. CWI report. http://oai.cwi.nl/oai/asset/13947/13947A.pdf

Kemper B, de Mast J, Mandjes M (2010) Modelling process flow using diagrams. Qual Reliab Eng Int 26(4):341–349

Ko S, Serfozo RF (2008) Sojourn times in G/M/1 fork-join networks. Nav Res Logist 55:432–443

Kühn P (1979) Approximate analysis of general queueing networks by decomposition. IEEE Trans Commun 27:113–126

Nelson R, Tantawi AN (1988) Approximate analysis of fork/join synchronization in parallel queues. IEEE Trans Comput 37:739–743

Stone C (1996) A course in probability and statistics. Duxbury Press, Belmont

Tijms H (1986) Stochastic modelling and analysis—a computational approach. In: Wiley series in probability and mathematical statistics: applied probability and statistics. Wiley, Chichester

Varma S, Makowski AM (1994) Interpolation approximations for symmetric fork-join queues. Perform Eval 20:245–265

Whitt W (1983) The queueing network analyzer. Bell Syst Tech J 62:2779–2815