Centrum Wiskunde & Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Approximations for the mean sojourn time in a parallel queue

B.P.H. Kemper, M.R.H. Mandjes

REPORT PNA-E0901 MARCH 2009

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Copyright © 2009, Centrum Wiskunde & Informatica
P.O. Box 94079, 1090 GB Amsterdam (NL)
Science Park 123, 1098 XG Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Approximations for the mean sojourn time in a parallel queue

ABSTRACT
This paper considers a parallel queue, which is two-queue network, where any arrival generates a job at both queues. 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 $\max(S_1; S_2)$. It is noted that earlier work has revealed that this class of models is notoriously hard to analyze. We first evaluate a number of bounds developed in the literature, and observe that under fairly broad circumstances these can be rather inaccurate. We distinguish between the homogeneous case, in which the jobs generated at both queue stem from the same distribution, and the heterogeneous case. For the former case we present a number of approximations, that are extensively tested by simulation, and turn out to perform remarkably well. For the latter case, we identify conditions under which S can be accurately approximated by the sojourn time of the queue with the highest load.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: parallel queues, sojourn time

APPROXIMATIONS FOR THE MEAN SOJOURN TIME IN PARALLEL QUEUES

BENJAMIN KEMPER & MICHEL MANDJES

Abstract. This paper considers a parallel queue, which is two-queue network, where any arrival generates a job at both queues. 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 $\max\{S_1, S_2\}$. It is noted that earlier work has revealed that this class of models is notoriously hard to analyze. We first evaluate a number of bounds developed in the literature, and observe that under fairly broad circumstances these can be rather inaccurate. We distinguish between the homogeneous case, in which the jobs generated at both queue stem from the same distribution, and the heterogeneous case. For the former case we present a number of approximations, that are extensively tested by simulation, and turn out to perform remarkably well. For the latter case, we identify conditions under which S can be accurately approximated by the sojourn time of the queue with the highest load.

Both authors are with Korteweg-de Vries Institute for Mathematics, Plantage Muidergracht 24, 1018 TV Amsterdam, the Netherlands. BK is also with the Institute for Business and Industrial Statistics, University of Amsterdam (IBIS UvA). MM is also with Eurandom, P.O. Box 513, 5600 MB Eindhoven, the Netherlands, and CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands. Part of this work was done while MM was at Stanford University, Stanford, CA 94305, US. Email: b.p.h.kemper@uva.nl, m.r.h.mandjes@uva.nl.

1. Introduction

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. One could call the resulting queueing system an 'M/G/1 parallel queue'. 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 parallel queues 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, health care applications, manufacturing systems, and communication networks. With S_i denoting a job's sojourn time in queue i, a particularly interesting object is the *parallel queue's sojourn time* $S := \max\{S_1, S_2\}$, as in many situations the job can be further processed only if service at both queues has been completed. One could think of many specific examples in which parallel queues (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, of which the parallel queue can be a building block. We refer the reader to for instance the process-flow-based modeling framework proposed in [6], featuring metrics such as the arrival rate, process sojourn time, and utilization. In the present paper we focus on the process sojourn time of the parallel queue.

M/G/1 parallel queues have been studied intensively in the past, see for instance the overview article [3], 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 parallel queue, Flatto and Hahn [5] 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 [4]; these provide insight into the interdependence between the two queues. For this M/M/1 parallel queue, 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 [8], 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 parallel queue (and in fact for the GI/G/1 parallel queue), upper and lower bounds on the mean sojourn time were derived by Baccelli and Makowski [1], relying on stochastic comparison techniques. 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.* [2]. The authors of [2] 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 [10]; in that 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 [7].

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 [1] 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). We observe that in many situations, the bounds presented in [1] are rather far apart (and sometimes even outperformed by trivial bounds). By investigating the mean sojourn time for a broad range of loads, and for various coefficients of variations, we empirically determine a relation between these quantities. It turns out that the ratio of $\mathbb{E}S$ and $\mathbb{E}S_1 = \mathbb{E}S_2$ just mildly depends on the load, in line with the observations in [8] for the case exponential service times.
- We then consider 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. For the cases this approach does not lead to accurate results, we present alternative rules of thumb.

The structure of the paper is as follows. In Section 2 we sketch the model, and present some preliminaries. We also review the bounds of [1], and explicitly calculate them for specific service-time distributions. In Section 3 we consider the homogeneous case, i.e., $B_1 =_{\rm d} B_2$, and identify under which conditions the bounds of [1] are far apart. We then present a number of approximations, which turn out to be highly accurate. Section 4 covers the heterogeneous case. The paper is concluded by a brief summary and discussion.

2. Model, preliminaries, and bounds

In this section we formally introduce the parallel queue (or: *fork-join network*), see Fig. 1. This system consists of two queues (or: workstations, nodes) that work in parallel. The jobs arrive according a Poisson process with parameter λ; without loss of generality, we can renormalize time such that $\lambda = 1$ (which we will do throughout this paper). Upon arrival the job *forks* into two different 'sub-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'. The load of node i is defined as $\varrho_i := \lambda\mathbb{E}B_i \equiv \mathbb{E}B_i < 1$. The systems stability is assured under the, intuitively obvious, condition $\max\{\varrho_1, \varrho_2\} < 1$, see [1].
The queues handle the sub-tasks in a first-come-first-serve fashion. In other words: if the sub-task finds the queue non-empty, it waits in the queue before until service starts. When both sub-tasks (corresponding to the same job) have been performed, they *join* and the job departs the network. Therefore, the total sojourn time of a the n-th job in the network is the *maximum* of two sojourn times of the sub-tasks, that is, in self-evident notation, $S_n = \max_{i=1,2} S_{i,n}$. The goal of this paper is to analyze the *mean sojourn time*, i.e.,

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

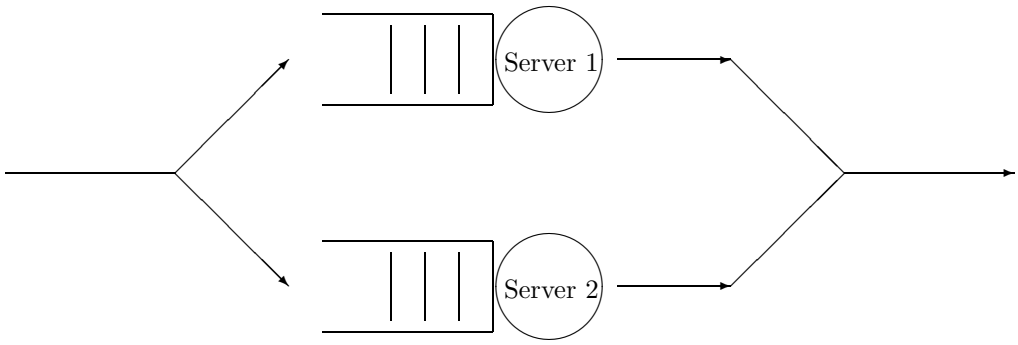

FIGURE 1. A simple fork-join queue

with S_i denoting the sojourn time of an arbitrary customer (in steady-state) in queue i. 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 [8], is recalled in Section 2.1. Relaxing the homogeneity and exponentiality assumptions, upper and lower bounds are known, which will be reviewed in Section 2.2, and made explicit in Section 2.3.

2.1. **The homogeneous M/M/1 parallel queue.** As proven in [8], 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. This result is found by first decomposing the mean sojourn time $\mathbb{E}S$ is the sum of the mean sojourn time m of an $M/M/1$ queue and a mean synchronization delay d, i.e., $\mathbb{E}S = m + d$. Using Little's formula and using the balance equations, one can show that

$$d = \frac{1}{\lambda}\sum_{i=1}^{\infty}\frac{i(i+1)}{2}p_{i0},$$

with p_{i0} the steady-state probability of i jobs in queue 1 and the other queue being empty. The first two moments, that is, $\sum_i ip_{i0}$ and $\sum_i i^2p_{i0}$, are found from the generating function [5]

$$P(z,0) = (1-\varrho)^{3/2}/\sqrt{1-\varrho z},$$

thus yielding $d = m\cdot(4-\varrho)/8$, as desired.
Observe that, when increasing the load from 0 to 1, the ratio of the mean sojourn time $\mathbb{E}S$ and 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 parallel queue.** In this section we discuss a number of bounds on $\mathbb{E}S$ in an M/G/1 parallel queue. It is noted that they in fact apply to the

GI/G/1 parallel queue, but under the assumption of Poisson arrivals explicit computations are possible, see Section 2.3.
An upper and lower bound for the general GI/G/1 case are presented by Baccelli and Makowski [1]; in the sequel we refer to these bounds as the *BM bounds*. The BM bounds for the sojourn time are in fact sojourn times of similar systems of two independent queues:

- 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, explaining that this yields 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, explaining that this yields a lower bound.

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 Section 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, due to Jensen's inequality, we have

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

Because $\max\{a, b\} = a + b - \min\{a, b\} \leq a + b$, 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 by the assumption that $\varrho_i = \lambda \mathbb{E}B_i < 1$ for both $i = 1$ and 2, which is identical to the stability condition of our parallel queueing system.
The idea behind the BM bounds is that the level of the variability of the parallel queueing system's waiting time should be increasing in the level of variability of the stochastic arrival process of the system. Replacing the 'parallel arrivals' by mutually independent homogeneous arrival processes, magnifies the level of variability in the arrival process, and therefore increases the level of variability in the system's sojourn time. Replacing the 'parallel arrivals' by identical deterministic arrival processes, as in the D/G/1 system, reduces the level of variability in the arrival process, and therefore reduces the level of variability in the system's sojourn time. This intuitive reasoning leads to the following bounds, which are rigorously proven in [1].

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, in self-evident notation, $\mathbb{E}\bar{S}$

equals

$$\int_0^\infty\int_0^y y\mathrm{d}\mathbb{P}_{\mathrm{GI/G/1}}(S_1\le x)\mathrm{d}\mathbb{P}_{\mathrm{GI/G/1}}(S_2\le y)+\int_0^\infty\int_y^\infty x\mathrm{d}\mathbb{P}_{\mathrm{GI/G/1}}(S_1\le x)\mathrm{d}\mathbb{P}_{\mathrm{GI/G/1}}(S_2\le y)$$
$$=\int_0^\infty y\mathbb{P}_{\mathrm{GI/G/1}}(S_1\le y)\mathrm{d}\mathbb{P}_{\mathrm{GI/G/1}}(S_2\le y)+\int_0^\infty x\mathbb{P}_{\mathrm{GI/G/1}}(S_2\le x)\mathrm{d}\mathbb{P}_{\mathrm{GI/G/1}}(S_1\le x);$$

we call the latter expression from now on 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 $\mathbb{E}\underline{S}$ equals

$$\int_0^\infty\int_0^y y\mathrm{d}\mathbb{P}_{\mathrm{D/G/1}}(S_1\le x)\mathrm{d}\mathbb{P}_{\mathrm{D/G/1}}(S_2\le y)+\int_0^\infty\int_y^\infty x\mathrm{d}\mathbb{P}_{\mathrm{D/G/1}}(S_1\le x)\mathrm{d}\mathbb{P}_{\mathrm{D/G/1}}(S_2\le y)$$
$$=\int_0^\infty y\mathbb{P}_{\mathrm{D/G/1}}(S_1\le y)\mathrm{d}\mathbb{P}_{\mathrm{D/G/1}}(S_2\le y)+\int_0^\infty x\mathbb{P}_{\mathrm{D/G/1}}(S_2\le x)\mathrm{d}\mathbb{P}_{\mathrm{D/G/1}}(S_1\le x),$$

which we denote in the sequel by L.

2.3. **BM bounds for an number of M/G/1 parallel systems.** We now present a number of explicit expression for the bounds u, U, ℓ, and L in the case of Poisson arrivals and various service time distributions. In Section 3 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 hyperexponential times (useful to approximate service times with coefficient of variation larger than 1). In the sequel we will denote by SCV the *squared* coefficient of variation, defined by the ratio of the variance and the squared mean.

M/M/1 case. Here we let the service times in the first and second queue be both exponentially distributed, with means ϱ_1 and ϱ_2 respectively; recall that the exponential distribution has SCV equal to 1. It is well-known 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

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

It is seen that if m_1 is considerably larger than m_2 (i.e., ϱ_1 considerably smaller than ϱ_2), then $\mathbb{E}S\approx m_1$. This is done as follows. Let m_2 be m_1/M for some $M>1$. Recall that $\ell=m_1\le\mathbb{E}S\le U$, and also

$$U=m_1\left(1+\frac{1}{M}\right)-\left(\frac{1}{m_1}+\frac{M}{m_1}\right)^{-1}\to m_1,$$

as $M \to \infty$. This indicates that, if the loads of both queues are highly asymmetric, the bottleneck queue essentially determines the parallel queue's sojourn time.

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

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

this leads to

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

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

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

where m_i is the mean sojourn time in queue i, which equals

$$m_i = \frac{\varrho_i^2}{2(1-\varrho_i)}(\text{SCV}_i + 1) + \varrho_i, \tag{2}$$

see for instance [9, Eq. (2.55)], 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. We now consider the situation of the service times being 'generalized Erlang' [9, 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 [9, 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 (1) 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 (2). 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 $\mathrm{D}/\mathrm{E}_{1,2}/1$ 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 μ_{i1}^{-1} with probability p_i, and from an exponential distribution with mean μ_{i2}^{-1} with probability $1-p_i$. We fix the mean service times, leading to the requirement

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

and, under the additional condition of 'balanced means' [9, Eq. (A.16)], the SCVs, leading to

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

and $\mu_{i1} = 2p_i\mu_i$ and $\mu_{i2} = 2(1-p_i)\mu_i$. It is obvious that we again have that S_i has mean (2), 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 Equations (1) 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 $\mathbb{P}(S_i \le x)$ for a $\mathrm{D}/\mathrm{H}_2/1$ 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 [8], 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 Section 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 [8] for the case of exponentially distributed service times.

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

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

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 $U \approx \frac{11}{8}m = 1.375\,m$, and for a system with large load $\varrho \uparrow 1$ gives $U \approx \frac{3}{2}m = 1.5\,m$. 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 and m is sensitive to the service load ϱ. For a system with a small load, $\varrho = 0.1$, we have $U \approx 1.45\, m$, whereas for a system with large load, $\varrho = 0.9$, we have $U \approx 1.49\, m$. 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 and 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 and 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 in a single queue, m. For this we have exact expressions (following from 'Pollaczek-Khinchine'), see [9, Eq. (2.55)].
(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 SCV s we obtained a value through simulation.
(iii) The ratio of $\mathbb{E}S$ and 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 and m, denoted by α_U(SCV).
(vi) The lower bound L, obtained through simulation (for SCV $= 1$ we have an exact expression).
(vii) The ratio of L and m, denoted by α_L(SCV).
(viii) The 'BM-spread', that is, the ratio of $(U - L)$ and $\mathbb{E}S$.

The service times with SCV equal to 0.25 and 0.33 are obtained by using E_4 and E_3 distributions, respectively. For SCV s larger than 1 we use hyperexponentional distribution, with the additional condition of 'balanced means' [9, 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 the bounds L and U are relatively close, but the difference can be substantial for higher SCV s. For higher loads, however, L and U tend to be far apart, particularly for low SCV s.
- 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.
- 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 as well.

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%

ϱ	SCV	m	$\mathbb{E}S$	α(SCV)	U	α_U(SCV)	L	α_L(SCV)	BM-Spread
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%

TABLE 1. Simulated sojourn times and the corresponding BM bounds.

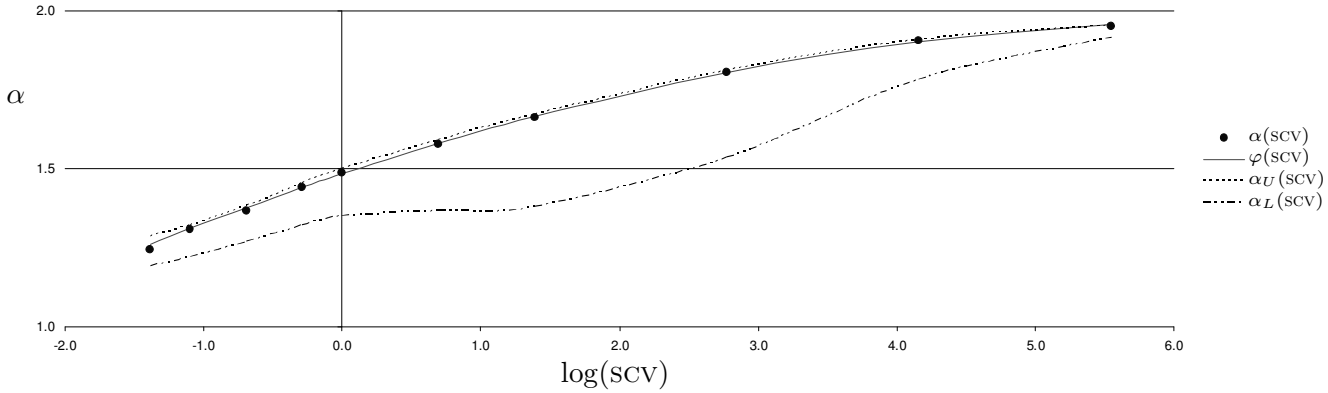

FIGURE 2. Graph with BM bounds, simulated values and approximated values for load $\varrho = 0.1$.

In view of the tables presented above and illustrated in Figures 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-Khintchine formula, see for example [9, 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 $m \cdot \varphi(\text{SCV})$. Note that m, i.e., the mean sojourn time of a single queue

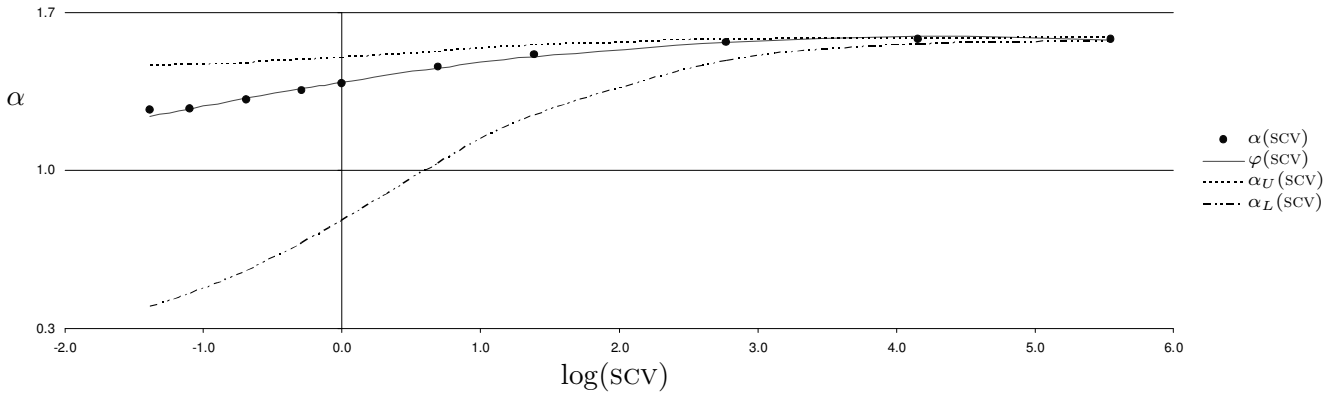

FIGURE 3. Graph with BM bounds, simulated values and approximated values for load $\varrho = 0.9$.

is known explicitly. The function $\varphi(\cdot)$ shown in Figures 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 $\alpha(\text{SCV})$ of the Weibull service-time distribution.

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

ϱ	SCV	m	$\mathbb{E}S$	$\alpha(\text{SCV})$	$\mathbb{E}S_{\text{W}}$	$\alpha(\text{SCV})_{\text{W}}$
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

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

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 SCV < 1 we observe that $\mathbb{E}S$ and $\alpha(\text{SCV})$ are nearly equal to their Weibullian counterparts; for SCV > 1 the difference is modest, that is, up to 3.5%. For $\varrho = 0.9$ the fit is accurate up to SCV $= 4$, whereas for SCV > 4 the difference is modest, that, is about 5%. The results of other numerical experiments give the same impression. These findings justify our two-moment approach.

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 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.

SCV	log(SCV)	$\varrho = 0.1$	$\varrho = 0.3$	$\varrho = 0.5$	$\varrho = 0.7$	$\varrho = 0.9$
0.25	-1.3863	1.2690	1.2603	1.2523	1.2462	1.2449
0.33	-1.0987	1.3070	1.2961	1.2858	1.2773	1.2733
0.50	-0.6931	1.3676	1.3526	1.3381	1.3251	1.3154
0.75	-0.2877	1.4401	1.4170	1.3948	1.3650	1.3568
1.00	0.0000	1.4874	1.4626	1.4374	1.4124	1.3875
2.00	0.6931	1.5792	1.5662	1.5447	1.5114	1.4607
4.00	1.3863	1.6634	1.6658	1.6423	1.5942	1.5148
16.0	2.7726	1.8048	1.8155	1.7685	1.6886	1.5682
64.0	4.1589	1.9062	1.8828	1.8143	1.7175	1.5831
256	5.5452	1.9527	1.8999	1.8207	1.7217	1.5840

TABLE 3. Simulated values of $\alpha(\text{SCV})$ of several SCV s and several loads ϱ.

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(SCV) of degree two, based on 10 datapoints. The coefficients are estimated by applying ordinary least squares. As can be seen in the left part of Table 4 and from Figure 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 $i = 0, \ldots, 4$, but our experiments indicate that for other values values of ϱ nice fits can be achieved by interpolating estimates for $\alpha(\text{SCV})$ linearly.

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

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

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(SCV) of degree one. The results are reported in the rightmost columns 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 [10]. It is first noted that their approach gives expressions that are in line with limiting results for heavy and light loads. Their idea is to interpolate these 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.

4. The heterogeneous case

Having dealt with the case of homogeneous servers in the previous section, we now focus on the situation that the servers are heterogeneous. We restrict ourselves to the case that the service times B_1 and B_2 stem from the same distribution, but with different parameters, as in the setting of Section 2. 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) 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 $\mathbb{E}S$.
Balsamo *et al.* [2] describe a numerical scheme for finding accurate upper and lower bounds for the situation of heterogeneous *exponentially distributed* service times. In this section we further explore this issue by studying the impact of heterogeneity on the mean sojourn time for a broader set of service-time distributions. As in previous section we will use 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).

M/M/1 case. In [2] the numerical experiments are such that the load ϱ_1 of queue 1 (which is the 'bottleneck') is in the interval $(0.1, 0.9)$, whereas the load of queue 2 is $\varrho_2 = b\varrho_1$, with the 'heterogeneity factor' $b = \{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\}$. In these experiments the bounds presented in [2] are rather tight, but the reader should bear in mind that the impact of heterogeneity is modest anyway for $b \in (\frac{1}{3}, \frac{2}{3})$ and a relatively high load in the bottleneck queue (so that, for these situations, $\mathbb{E}S$ can be approximated by the mean sojourn time of queue 1). The most substantial impact occurs in the range of ratios $b \in (0.7, 1)$, as will be shown in Table 5.
Table 5 gives for $\varrho = 0.1 \cdot i$, with $i = 1, \ldots, 9$, simulated values of $\alpha(\text{SCV})$ for various b. As seen from the table, $\alpha(\text{SCV})$ increases in b, as could be expected. We also observe that $\alpha(\text{SCV}) \downarrow 1$ if $b \downarrow 0$, since the sojourn times in the parallel queueing system then mimic those in the bottleneck queue.
We emphasize that Table 5 shows that in a considerable part of the parameter space $\alpha(\text{SCV})$ can be accurately approximated by 1. It is observed that the mean sojourn time of the bottleneck queue plus an increment of about 10% can be a good (conservative) approximation for all levels of heterogeneity b up to, say, 0.4.

ϱ_1	$b = 0.1$	$b = 0.2$	$b = 0.3$	$b = 0.4$	$b = 0.5$	$b = 0.6$	$b = 0.7$	$b = 0.8$	$b = 0.9$	$b = 1.0$
0.1	1.0074	1.0278	1.0590	1.0993	1.1475	1.2033	1.2656	1.3340	1.4081	1.4875
0.2	1.0060	1.0227	1.0490	1.0842	1.1281	1.1805	1.2414	1.3107	1.3885	1.4750
0.3	1.0047	1.0178	1.0395	1.0695	1.1084	1.1567	1.2153	1.2851	1.3672	1.4626
0.4	1.0034	1.0135	1.0305	1.0553	1.0885	1.1320	1.1872	1.2566	1.3432	1.4499
0.5	1.0024	1.0098	1.0225	1.0417	1.0688	1.1061	1.1568	1.2246	1.3156	1.4374
0.6	1.0015	1.0064	1.0153	1.0290	1.0495	1.0798	1.1236	1.1877	1.2824	1.4250
0.7	1.0007	1.0035	1.0090	1.0177	1.0317	1.0534	1.0880	1.1445	1.2405	1.4124
0.8	1.0003	1.0015	1.0043	1.0088	1.0160	1.0285	1.0512	1.0938	1.1828	1.3996
0.9	1.0000	1.0001	1.0009	1.0023	1.0044	1.0087	1.0174	1.0374	1.0969	1.3875

TABLE 5. Simulated values of $\alpha(\text{SCV})$ in case of exponential service-time distribution for various levels of service loads heterogeneity (of type $\varrho_2 = b\varrho_1$).

When increasing b from, say, 0.7 to 1.0 we see that the $\alpha(\text{SCV})$ sharply increases, particularly for the (perhaps more relevant) heavier loads. For these situations the value for $b = 1.0$, which can be determined as described in Section 3, provides us with a conservative estimate.

$M/E_2/1$ case. In a similar way the impact of heterogeneity on values of $\alpha(\text{SCV})$ is presented in case of an E_2 service-time distribution. In Table 6 we observe the same behavior of $\alpha(\text{SCV})$ for the various loads and levels of heterogeneity. The impact in the range $b \in (0.7, 1)$ for the relatively high loads is less severe compared to the M/M/1 case.

ϱ_1	$b = 0.1$	$b = 0.2$	$b = 0.3$	$b = 0.4$	$b = 0.5$	$b = 0.6$	$b = 0.7$	$b = 0.8$	$b = 0.9$	$b = 1.0$
0.1	1.0013	1.0086	1.0243	1.0490	1.0826	1.1247	1.1748	1.2323	1.2967	1.3676
0.2	1.0010	1.0072	1.0204	1.0419	1.0721	1.1113	1.1596	1.2172	1.2840	1.3601
0.3	1.0008	1.0057	1.0166	1.0348	1.0613	1.0970	1.1429	1.2002	1.2697	1.3526
0.4	1.0006	1.0044	1.0130	1.0278	1.0502	1.0817	1.1244	1.1806	1.2531	1.3453
0.5	1.0005	1.0032	1.0096	1.0211	1.0390	1.0656	1.1038	1.1578	1.2333	1.3381
0.6	1.0003	1.0021	1.0065	1.0146	1.0280	1.0489	1.0812	1.1312	1.2089	1.3313
0.7	1.0001	1.0013	1.0039	1.0089	1.0177	1.0324	1.0569	1.0996	1.1769	1.3251
0.8	1.0001	1.0004	1.0018	1.0044	1.0089	1.0170	1.0323	1.0627	1.1322	1.3197
0.9	1.0000	1.0001	1.0004	1.0014	1.0023	1.0051	1.0106	1.0238	1.0665	1.3154

TABLE 6. Simulated values of $\alpha(\text{SCV})$ in case of E_2 service-time distribution for various levels of service loads heterogeneity (of type $\varrho_2 = b\varrho_1$).

$M/H_2/1$ case. Finally, the impact of heterogeneity on values of $\alpha(\text{SCV})$ is presented in case of an H_2 service-time distribution with $\text{SCV} = 4$. In Table 7 we observe a similar behavior of $\alpha(\text{SCV})$, for the various service loads and levels of heterogeneity.

From the experiments above a few, more general, conclusions can be drawn:

- Restricting ourselves to cases with $\text{SCV} \leq 4$ (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 $\varrho_1 \in [0.8, 0.9]$.
- Similarly, for the same range of SCVs, but b smaller than 0.3 and all $\varrho_1 \leq 0.9$, the same statement applies.

ϱ_1	$b=0.1$	$b=0.2$	$b=0.3$	$b=0.4$	$b=0.5$	$b=0.6$	$b=0.7$	$b=0.8$	$b=0.9$	$b=1.0$
0.1	1.0178	1.0551	1.1042	1.1621	1.2284	1.3014	1.3821	1.4691	1.5630	1.6634
0.2	1.0131	1.0426	1.0840	1.1354	1.1974	1.2698	1.3522	1.4458	1.5510	1.6682
0.3	1.0097	1.0325	1.0664	1.1112	1.1679	1.2369	1.3196	1.4172	1.5320	1.6658
0.4	1.0065	1.0235	1.0509	1.0885	1.1387	1.2026	1.2832	1.3831	1.5058	1.6569
0.5	1.0047	1.0164	1.0374	1.0678	1.1100	1.1671	1.2427	1.3419	1.4720	1.6423
0.6	1.0030	1.0108	1.0255	1.0481	1.0817	1.1296	1.1970	1.2923	1.4275	1.6215
0.7	1.0018	1.0065	1.0155	1.0308	1.0539	1.0907	1.1461	1.2315	1.3674	1.5942
0.8	1.0008	1.0029	1.0073	1.0152	1.0292	1.0515	1.0900	1.1567	1.2839	1.5592
0.9	1.0000	1.0004	1.0006	1.0041	1.0087	1.0171	1.0339	1.0680	1.1581	1.5148

TABLE 7. Simulated ratios $\alpha(\text{SCV})$ in case of H_2 service-time distribution for various levels of service loads heterogeneity (of type $\varrho_2 = b\varrho_1$).

- 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.

5. Concluding remarks

The parallel queue is a well known generic building block of more complex service systems in industry, services, and healthcare. The fact that these systems have proven to be highly complex, even in the very simple case of just two servers, is undisputably true. This makes the analysis challenging, and explains the need for simple heuristics.

This paper first discussed the bounds suggested by Baccelli and Makowski [1]. 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 analyzed the heterogeneous parallel queue.

In more detail, the conclusions are as follows:

- A trivial lower on the parallel 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 [1]. We do so for various phase-type service-time distributions. The lower bound suggested in [1], 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 parallel queue itself as well.
- For a substantial part of the parameter space both bounds from [1] 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 of the parallel queue, that is $\mathbb{E}S$, and the mean sojourn time of a single queue, that is m, is depends 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 parallel queue, 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 parallel queueing system insensitive with respect to higher moments of the service-time distribution?
- The study on the effect of heterogeneity, see Section 4, 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).

References

[1] F. Baccelli and A. M. Makowski. Simple computable bounds for the fork-join queue. In *Proc. Johns Hopkins Conf. Information Science*, Johns Hopkins University, Baltimore, 1985.

[2] S. Balsamo, L. Donatiello, and N. M. van Dijk. Bound performance models of heterogeneous parallel processing systems. *IEEE Transactions on Parallel and Distributed Systems*, 9:1041–1056, 1998.

[3] O. Boxma, G. Koole, and Z. Liu. Queueing-theoretic solution methods for models of parallel distributed systems. In *Performance Evaluation of Parallel and Distributed Systems*, pages 1–24, CWI Tract 105, Amsterdam, 1994.

[4] L. Flatto. Two parallel queue created by arrivals with two demands II. *SIAM Journal on Applied Mathematics*, 45:861–878, 1985.

[5] L. Flatto and S. Hahn. Two parallel queue created by arrivals with two demands I. *SIAM Journal on Applied Mathematics*, 44:1041–1053, 1984.

[6] B. P. H. Kemper, J. de Mast, and M. R. H. Mandjes. Modelling process flow using diagrams. *Submitted*, 2009.

[7] S. Ko and R. F. Serfozo. Sojourn times in G/M/1 fork-join networks. *Naval Research Logistics*, 55:432–443, 2008.

[8] R. Nelson and A. N. Tantawi. Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 37:739–743, 1988.

[9] H. Tijms. *Stochastic modelling and analysis — a computational approach*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1986.

[10] S. Varma and A. M. Makowski. Interpolation approximations for symmetric fork-join queues. *Performance Evaluation*, 20:245–265, 1994.