An Approximate Solution for *Ph/Ph/1* and *Ph/Ph/1/N* Queues

Alexandre Brandwajn
Baskin School of Engineering
University of California Santa Cruz
USA
alexb@soe.ucsc.edu

Thomas Begin
LIP UMR CNRS - ENS Lyon -
UCB Lyon 1 - INRIA 5668
France
thomas.begin@ens-lyon.fr

ABSTRACT

We propose a simple approximation to assess the steady-state probabilities of the number of customers in *Ph/Ph/1* and *Ph/Ph/1/N* queues, as well as probabilities found on arrival, including the probability of buffer overflow for the *Ph/Ph/1/N* queue. The phase-type distributions considered are assumed to be acyclic. Our method involves iteration between solutions of an *M/Ph/1* queue with state-dependent arrival rate and a *Ph/M/1* queue with state-dependent service rate. We solve these queues using simple and efficient recurrences. By iterating between these two simpler models our approximation divides the state space, and is thus able to easily handle phase-type distributions with large numbers of stages (which might cause problems for classical numerical solutions). The proposed method converges typically within a few tens of iterations, and is asymptotically exact for queues with unrestricted queueing room. Its overall accuracy is good: generally within a few percent of the exact values, except when both the inter-arrival and the service time distributions exhibit low variability. In the latter case, especially under moderate loads, the use of our method is not recommended.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Queueing theory; D.4.8 [**Performance**]: Queueing theory

Keywords

Ph/Ph/1 and *Ph/Ph/1/N* queue, steady-state probabilities, buffer overflow probability, large number of phases, approximate solution, numerical stability.

1. INTRODUCTION

Despite the recent proliferation of multi-server facilities in numerous application areas, e.g. [GAN03, GEP06], many situations remain where the processing of requests (customers) is performed by a single server. This is the case, for instance, for packets at a network interface [BOL93] or requests at a database lock. Clearly, the distributions of the times between customer arrivals, as well as of the request service times are dependent on the particular application and, in general, need not be close to the exponential distribution. In many cases, there may be a high variability, in both the inter-arrival and service times. We use acyclic phase-type distributions (e.g. [BOB05]) to represent the time between arrivals and the service time so that the resulting model is a *Ph/Ph/1* queue. As is well known, any distribution can be approximated arbitrarily closely by a phase-type distribution [OCI90]. Since in all human-made systems the queueing room is finite, the unrestricted *Ph/Ph/1* queue may not be an adequate model for higher traffic intensity as the buffer overflow probability becomes of interest in many applications. Hence, we also explicitly consider a queue in which the total number of customers cannot exceed a given value N, i.e. the *Ph/Ph/1/N* queue.

Although there is a considerable body of literature devoted to the single-server queue, e.g. [CHAU92, COH82, OTT87, JAG88, ABA93], no explicit easily usable solution exists in the general case, not even for the average number of customers with unrestricted queueing room [BOL05, page 265]. There are established numerical methods to solve *Ph/Ph/1* queues (e.g. matrix-geometric methods [LAT99, BIN05]), however, due to the cardinality of the resulting state space, they may not scale well for large numbers of phases needed to adequately represent empirical distributions.

A number of approximations exist for the mean waiting time [BUZ93, KIM91, KUE79, SHA80, BOL05, RAO99], however, with few exceptions [WHI89], they are limited to the first two moments of the service and inter-arrival times, and none seems readily applicable to the evaluation of buffer overflow probabilities.

Recently, a simple numerically stable recurrent solution has been proposed to compute the steady-state probabilities for the number of customers in the *M/Ph/1* queue with state-dependent arrivals [BRA08], and an analogous recurrent approach to the computation of the steady-state probabilities in a *Ph/M/c* queue with state-dependent service [BRA12]. We propose to use these recurrent solutions to obtain an approximation for the steady-state distribution of the number of customers in the *Ph/Ph/1* and the *Ph/Ph/1/N* queues. The resulting approximation has the advantage of taking into account the actual form (as opposed to only the first two moments) of the service and inter-arrival distributions. The knowledge of the stationary probability for the number of customers in the system allows us to assess the state of

ICPE'12, April 22–25, 2012, Boston, Massachusetts, USA.

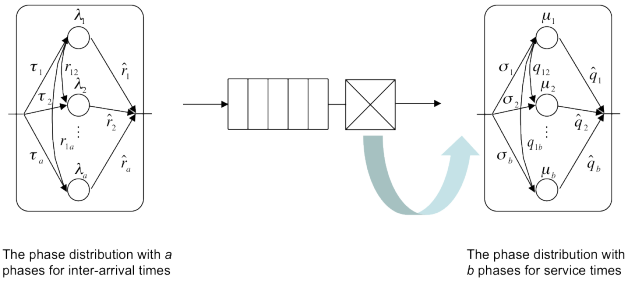

Figure 1. *Ph/Ph/1* queue.

the queue found by arriving requests, including the probabilities of buffer overflow.

In the following section, we derive our approximation. Section 3 gives numerical examples to illustrate the typical performance of this approximation. Section 4 concludes this paper.

2. APPROXIMATION

As stated before, we assume that the times between arrivals and the service times are represented as acyclic phase-type distributions [OCI90]. Figure 1 shows the corresponding *Ph/Ph/1* queue. We denote by a the number of phases in the distribution of the times between arrivals, and by b the number of phases in the distribution of the service times. The total current number of requests in the system is denoted by n. The steady-state of this queue can be described by the current phase of the arrival process j, the current phase of the service process (if the queue is nonempty) i, and by the total number of requests in the system, viz. (j,i,n). In the case of a finite queueing room, we consider that the arrival process continues unperturbed when the buffer is full and arriving customers are then simply lost. Other assumptions on the arrival process (e.g. blocking of the request source) are possible. Table 1 summarizes the notation used in our paper.

If we consider a marginal state description (i,n), our queue can be represented as Queue 1 in Figure 2 where the state-dependent rate of customer arrivals $\alpha(n,i)$ is given by

$$\alpha(n,i) = \sum_{j=1}^{a} \lambda_j \hat{r}_j \, p(j \mid n,i) \tag{1}$$

Analogously, if we consider the marginal state description (j,n), our queue can be represented as Queue 2 shown in Figure 2 where the state-dependent rate of service $u(n,j)$ is given by

$$u(n,j) = \sum_{i=1}^{b} \mu_i \hat{q}_i \, p(i \mid n,j) \tag{2}$$

To derive our approximation, we assume that $p(j \mid n,i) \approx p(j \mid n)$ and $p(i \mid n,j) \approx p(i \mid n)$. Consequently, we have $\alpha(n,i) \approx \alpha(n)$ and $u(n,j) \approx u(n)$. Queue 1 in Figure 2 then becomes an *M/Ph/1* queue with a state-dependent arrival rate $\alpha(n)$, and Queue 2 becomes a *Ph/M/1* queue with a state-dependent service rate $u(n)$. A simple recurrence can be used to obtain an efficient and numerically stable solution of the *M/Ph/1* queue [BRA08] yielding the state-dependent service rate $u(n)$. Similarly, an analogous simple recurrence can be used to obtain the state-dependent arrival rate $\alpha(n)$ [BRA12]. Hence, the obvious idea

Table 1. Principal notation used in this paper

τ_j	Probability that arrival process starts in phase j, $j = 1,...,a$
λ_j	Completion rate for phase j of arrival process
r_{jl}	Probability that arrival process continues in phase l upon completion of phase j, $j,l = 1,...,a,\ l > j$
$\hat{r}_j$	Probability that arrival process ends (new request generated) upon completion of phase j, $j = 1,...,a$
σ_i	Probability that service of a request starts in phase i, $i = 1,...,b$
μ_i	Completion rate for phase i of service process
q_{ih}	Probability that service process continues in phase h upon completion of phase i, $i,h = 1,...,b, h > i$
$\hat{q}_i$	Probability that service process ends (request departs the system) upon completion of phase i, $i = 1,...,b$
$p(i \mid n,j)$	Conditional probability that the service stage is i given that the number in the system is n and the current arrival stage is j
$p(j \mid n,i)$	Conditional probability that the arrival stage is j given that the number in the system is n and the current stage of the service process is i

to iterate between the solutions of these two queues until a fixed point is reached for the arrival and service rates.

Having obtained the arrival and service rates $\alpha(n)$ and $u(n)$, we can compute the steady-state probability for the number of customers in the system $p(n)$ as

$$p(n) \approx \frac{1}{G} \prod_{m=1}^{n} \alpha(m-1)/u(m), \ n = 0,1,... \tag{3}$$

G is a normalization constant chosen so that $\sum_n p(n) = 1$. The mean number of customers in the system is $\bar{n} = \sum_n np(n)$. The probability that an arriving customer finds n customers already present in the system, $P_A(n)$, can be expressed as

$$P_A(n) \approx \frac{\alpha(n)p(n)}{\sum_{i \geq 0} \alpha(i)p(i)}, \quad n = 0,1,... \tag{4}$$

The proposed approximate solution can be described as follows (superscripts denote the iteration number)

(a) Set the initial value of the arrival rate $\alpha^0(n)$ to the inverse of the mean time between arrivals for all values of $n = 0,1,...$

(b) Solve the simple recurrence for the *M/Ph/1* queue [BRA08] with arrival rate $\alpha^{k-1}(n)$ ($k = 1,2,...$ is the iteration number) to produce the state-dependent service rate $u^k(n)$ for $n = 1,...,n^k_{\max}$. With a finite buffer, $n^k_{\max}$ is the maximum number of customers in the system N, and with unrestricted buffer, it is the value of the number of customers n for which the rate $u^k(n)$ is close enough to its asymptotic value (cf. [BRA08].) Compute also the expected number of customers in the system $\bar{n}^k_a$ from the *M/Ph/1* model.

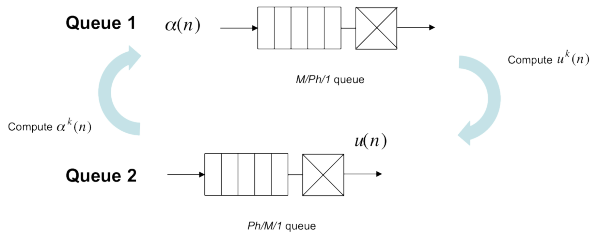

Figure 2. Iterations between *M/Ph/1* and *Ph/M/1* queues.

(c) Solve the *Ph/M/1* queue with the service rate $u^k(n)$ from Step (b), using the simple recurrence given in [BRA12], to produce the state-dependent arrival rate $\alpha^k(n)$ for $n = n^k_{\max},...,0$. Compute also the expected number of customers in the system $\overline{n}^k_b$ from the *Ph/M/1* model.

(d) If $\left|1-\overline{n}^k_a/\overline{n}^k_b\right| < \varepsilon$, where ε is the desired convergence stringency, go to Step (e), otherwise perform another step of the iteration, i.e., go to Step (b).

(e) Compute $p(n)$ and $P_A(n)$ from formulae (3) and (4), respectively.

In the next section we discuss the accuracy and speed of convergence of the proposed approximate solution.

3. ACCURACY AND SPEED OF CONVERGENCE

We performed a large number of tests of the proposed approximation comparing its results to those of an exact numeric solution. In addition to the mean number of customers in the system the test quantities included the probability that a customer has to wait before service, as well as the general shape of the steady-state probability distribution $p(n)$. It is interesting to note that our approximation produces the correct server utilization in the case of an unrestricted queueing room. For queues with restricted queueing room, we examined also the server utilization and the probability of buffer overflow. The typical accuracy tends to be good, within a few percent of the exact values. In virtually all cases a low number of iterations (a few tens) is sufficient to attain a fixed-point convergence of our approximate solution (in all examples, we used $\varepsilon = 10^{-7}$).

Four examples illustrate the behavior of the proposed method.

Example A: Small coefficients of variation (less than 1) for arrivals and service

We start by the case when the times between arrivals and the service times both exhibit low variability, viz., we consider an unrestricted $E_k/E_k/1$ queue with the same number of stages in the arrival and service distributions. In Figure 3 we show the relative error in the mean number of customers in such a system as a function of the server utilization for the number of stages varying from 2 to 10 , i.e., the squared coefficient of variation varying from 0.5 to 0.1. The approximation is, of course, exact for the *M/M/1* queue. We notice that the largest relative errors tend to occur in the range of moderate to moderately high server utilizations (say, 0.6 to 0.9). In this range, the accuracy of the approximation tends to degrade with the number of stages in the Erlang distribution, and exceeds 20% with 7 or more stages.

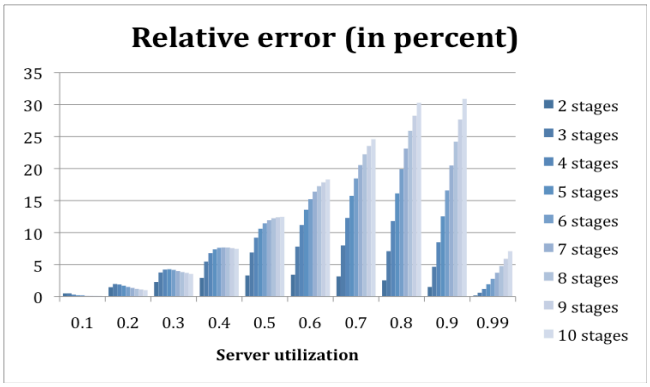

Figure 3. Relative error for mean number in system for a range of numbers of stages in arrival and service Erlang distributions (same number of stages for both) in Example A.

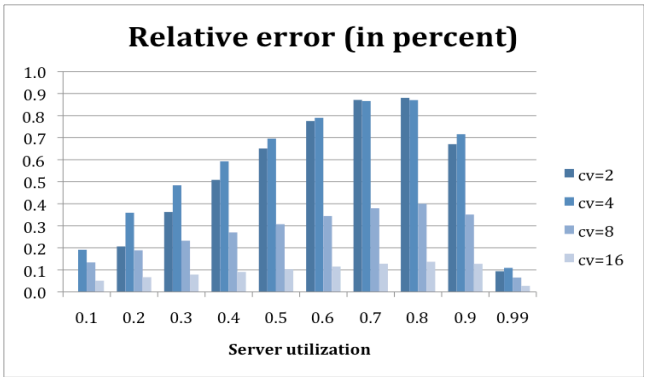

Figure 4. Relative error for mean number in system for a range of coefficients of variation in Example B.

Interestingly, as the server approaches saturation, the approximation accuracy improves. In fact, one can show that our approximation is asymptotically exact as $n \to \infty$ (see Appendix A.) This explains the improved accuracy near server saturation seen in Figure 3.

Intuitively, the lower accuracy when both arrivals and service exhibit high regularity, appears to be due to the fact that with hypo-exponential distributions, when the number of customers in the system is low (especially, just one user), the knowledge of the current stage of the service distribution provides non-negligible information on the possible stage of the arrival process (and vice versa). In particular, when there is a single user in the system and it is in its first stage of service, it is very likely that the arrival process is also in its first stage. This knowledge is lost in our approximation. Because the method tends to be inaccurate when both the time between arrivals and the service times exhibit low variability (say, coefficients of variation less than 0.3), especially for moderate loads, our approximation is not recommended in this case.

Example B: Coefficients of variation greater than 1

In our second example, we consider an unrestricted queue with times between arrivals represented by a two-phase hyper-exponential distribution (H_2). The service times are represented by a different two-phase H_2 distribution with mean 1 and the same coefficient of variation as the distribution of the times between

Table 2. Accuracy and convergence of speed in Example C

Server utilization	Mean number in the system		No wait probability		Number of iterations
	Exact	Appr.	Exact	Appr.	
0.1	0.1079	0.1079	0.8834	0.8834	3
0.2	0.2365	0.2365	0.7669	0.7669	3
0.3	0.3968	0.3968	0.6503	0.6503	3
0.4	0.6097	0.6097	0.5337	0.5337	3
0.5	0.9191	0.9191	0.4172	0.4172	4
0.6	1.4374	1.4373	0.3007	0.3007	4
0.7	2.5600	2.5599	0.1842	0.1843	4
0.8	7.3596	7.3590	0.0682	0.0682	5
0.9	173.623	173.861	0.0001	0.0001	15

arrivals (see Appendix.) Figure 4 shows the relative error for the mean number of users in the system for a range of traffic intensities and for coefficients of variation ranging from 2 to 16.

We observe that, in this example, the relative errors of the proposed approximation are small, on the order of a percent, and remain below 1% even for a traffic intensity of 0.99 and coefficients of variation of 16.

Example C: Large coefficient of variation for time between arrivals and small coefficient of variation for service

Our next example is a queue with a finite queueing room of 200 ($N = 200$). The time between arrivals is represented by a two-phase hyper-exponential distribution with a coefficient of variation of 20, and the service time is represented by an Erlang-5 distribution (squared coefficient of variation of 0.2). We show in Table 2 the results obtained for this example, including the number of iterations needed to achieve the convergence, for a range of server utilization values.

We observe that the relative errors of the proposed approximation remain below one percent despite the small coefficient of variation of the service time distribution. We also observe that only a few iterations are required to attain convergence. In our next example we take a closer look at the convergence pattern of our approximation in the context of a larger total number of phases.

Example D: Pareto-like distribution of the time between arrivals with 16 phases and four-phase service time distribution

In our last example, we consider a *Ph/Ph/1/N* queue with a buffer size of $N = 15$. The arrival process is represented by a Pareto-like distribution with a total of 16 phases, 10 of which are used in the heavy-tail part of the distribution [BRA11]. The service time is represented as a mixture of two Erlang-2 distributions with overall mean 1 and coefficient of variation 3. The rate of a single stage in the first Erlang-2 distribution is 4.4064 and this distribution is selected with probability 0.95. With probability 0.05 the second Erlang-2 distribution is selected. The rate of a single stage for this distribution is 0.1758.

Figure 5 illustrates the convergence pattern of our approximation to its fixed-point solution for this example. We observe the evolution of the relative difference in mean number of customers between the *M/Ph/1* and *Ph/M/1* models as the iteration progresses for server utilizations of 0.5, 0.8 and 0.9. In all three

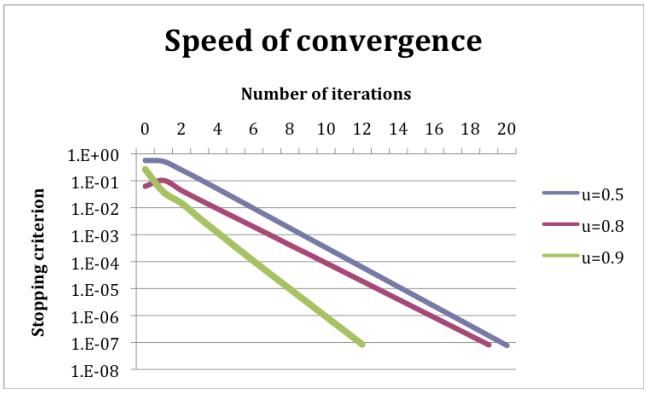

Figure 5. Speed of convergence of proposed approximation for various levels of server utilization for Example D.

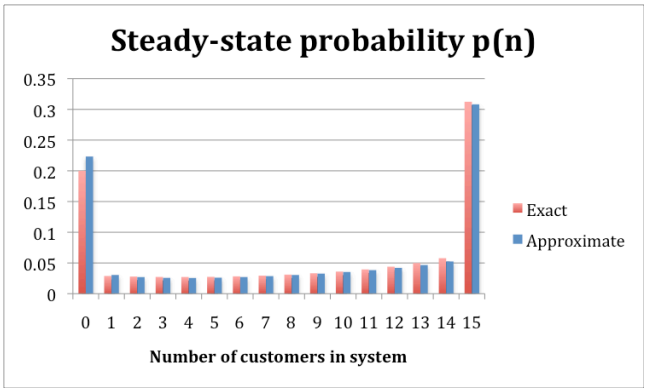

Figure 6. Exact and approximate probabilities for server utilization of 0.8 in Example D.

cases, the decrease in the relative difference appears to be geometric after the first few iterations. Such a geometric decrease seems typical for the convergence of the proposed method.

Figure 6 illustrates the ability of our approximation to reproduce the shape of the steady-state distribution of the number of customers in the system $p(n)$ (for server utilization of 0.8). We observe that the general shape of the steady-state distribution is well represented and the relative errors for individual state probabilities are moderate.

Summary discussion of empirical results

In the last three examples, the relative errors of the proposed approximation tend to be within just a few percent of the exact values and the method converges within a limited number of iterations. Loss probabilities (not reported in this paper) similarly tend to be with a few percent of the exact values. The above behavior appears typical for the method. As shown by example A, the proposed method tends to be less accurate when both the time between arrivals and the service times exhibit low variability (say, coefficients of variation less than 0.3), especially for moderate loads. Therefore, our approximation is not recommended in this case. Note that, as illustrated by example C, if only one of the inter-arrival or service time distributions exhibits low variability the method's accuracy does not seem affected.

Overall, the accuracy of the proposed approximation varies somewhat with the shape of the inter-arrival and service distributions (not just their first two moments). It tends to be particularly good when these distributions are skewed (e.g. unbalanced hyper-exponentials). Highly skewed distributions tend to be characteristic of the traffic in computer networks. It is important to note that these cases happen to be most difficult for some numerical methods [CHAU92] and discrete-event simulation alike [ASM00]. As illustrated by examples B and C, our method can easily handle problems with very large coefficients of variation.

It is worthwhile noting that the speed and numerical stability advantage of the proposed approximation over an exact numerical solution is particularly glaring with higher numbers of stages in the phase distributions. In our experimentation, the method easily handled distributions with hundred phases.

4. CONCLUSIONS

We have proposed an approximation to obtain steady-state probabilities of the number of customers in *Ph/Ph/1* and *Ph/Ph/1/N* queues, as well as related probabilities "seen" by an arriving customer, including the probability of buffer overflow in the case of the *Ph/Ph/1/N* queue. The phase-type distributions considered are assumed to be acyclic. Our method iterates between solutions of an *M/Ph/1* queue with state-dependent arrival rate and a *Ph/M/1* queue with state-dependent service rate. Each of these queues is solved using an efficient numerically stable recurrence. The resulting method is simple and easy to implement.

Although we don't have a theoretical proof of convergence of our method, in practice it converges typically within a few tens of iterations. The results produced by our approximation tend to be within a few percent of the exact values, except when both the inter-arrival and the service time distributions exhibit low variability. In the latter case, especially under moderate loads, the use of our method is not recommended.

Compared to an exact numerical solution of a *Ph/Ph/1* queue, by dividing the state space (through the iteration between the *M/Ph/1* and *Ph/M/1* queues) the proposed method affords a significant reduction in computational complexity. The resulting speed advantage is particularly significant with a larger number of phases possibly needed to represent empirical distributions. Additionally, numerical problems due to floating point underflow issues for very small state probabilities are reduced owing to the partitioning of the state space into normalized subsets.

Future work includes the extension of the proposed method to the *Ph/Ph/c* queue.

5. REFERENCES

[ABA93] Abate, J., Choudhury, G. L. and Whitt, W. 1993. Calculation of the *GI/G/1* waiting time distribution and its cumulants from Pollaczek's formulas. *Arch. Elektr. Uebertragung 47*, 311-321.

[ASM00] Asmussen, K., Binswanger, K. and Hojgaard B. 2000. *Rare events simulation for heavy-tailed distributions.* Bernoulli. 6, 2, 303-322.

[BIN05] Bini, D. A., Latouche, G., Meini, B., 2005. *Numerical Methods for Structured Markov Chains.* Oxford University Press, Inc.

[BOB05] Bobbio, A., Horváth, A. and Telek, M. 2005. Matching Three Moments with Minimal Acyclic Phase Type Distributions. *Stochastic Models*. 21, 2, 303-326.

[BOL93] Bolot, J. 1993. End-to-end packet delay and loss behavior in the Internet. *In SIGCOMM Computer Communication Review*. 23, 4 (Oct. 1993), 289-298.

[BOL05] Bolch, G., Greiner, S., Meer, H. d. and Trivedi, K. S. 2005. *Queueing Networks and Markov Chains*. Second Edition, Wiley-Interscience.

[BUZ93] Buzacott, J.A. and Shanthikumar, J.G. 1993. *Stochastic models of manufacturing systems*. Prentice Hall, Englewood Cliffs.

[BRA08] Brandwajn, A. and Wang, H. 2008. A conditional probability approach to *M/G/1*-like queues. Performance Evaluation. 65, 5 (May. 2008), 366-381.

[BRA11] Brandwajn, A. and Begin, T. 2011. Performance evaluation of a single node with general arrivals and service. *In ASMTA 2011*.

[BRA12] Brandwajn, A. and Begin, T. 2012. A Recurrent Solution of *Ph/M/c/N*-like and *Ph/M/c*-like Queues. In INRIA Research Report 7321. (June 2010). To appear in J. of App. Probability. 49, 1 (March 2012).

[CHAU92] Chaudhry, M. L., Agarwal, M. and Templeton, J. G. 1992. Exact and approximate numerical solutions of steady-state distributions arising in the queue *GI/G/1*. *Queueing Systems Theory Applications*. 10, 1-2 (Jan. 1992), 105-152.

[COH82] Cohen, J.W. 1982. *The single server queue*. North- Holland (second edition).

[GAN03] Gans, N., Koole, G. and Mandelbaum, A. 2003. Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5, 79-141.

[GEP06] Gepner, P. and Kowalik, M. F. 2006. Multi-Core Processors: New Way to Achieve High System Performance. *In Proceedings of PARELEC* (September 13 - 17, 2006). Washington, DC, 9-13.

[JAG88] Jagerman, D. 1988. Approximations for waiting time in *GI/G/1* systems. *Queueing Systems Theory Applications*. 2, 4 (Feb. 1988), 351-361.

[KIM91] Kimura, T. 1991. Approximating the Mean Waiting Time in the *GI/G/s* Queue. *The Journal of the Operational Research Society*. 42, 11 (Nov. 1991), 959- 970.

[LAT99] Latouche , G., Ramaswami, V., 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA, 1999.

[OCI90] O'Cinneide, C.A. 1990. Characterization of phase-type distributions. *Communications in Statistics: Stochastic Models*. 6, 1, 1-57.

[OTT87] Ott, T.J. 1987. On the Stationary Waiting-Time Distribution in the *GI/G/1* Queue, I: Transform Methods and Almost-Phase-Type Distributions. *Advances in Applied Probability*. 19, 1 (Mar. 1987), 240-265.

[RAO99] Rao, B. V. and Feldman, R. M. 1999. Numerical approximations for the steady-state waiting times in a *GI/G/1* queue. *Queueing Systems Theory Applications.* 31, 1/2 (Jan. 1999), 25-42.

[SHA80] Shanthikumar, J.G. and Buzacott, J.A. 1980. On the approximations of the single-server queue. *International Journal Production Research*. 18 (1980), 761-773.

[WHI89] Whitt, W. 1989. An Interpolation Approximation for the Mean Workload in a *GI/G/1* Queue. *Operations Research*. 37, 6 (Nov. - Dec. 1989), 936-952.

APPENDIX

A. Solution asymptotically exact

We will now show that our method produces results that are asymptotically exact as $n \to \infty$ in the case of unrestricted queueing room.

Consider the original *Ph/Ph/1* queue described in Section 2. In steady state, the queue can be described by the probability $p(j,i,n)$ where j ($j = 1,...,a$) is the current phase of the arrival process, i ($i = 1,...,b$) is the current phase of the service process, and n is the current number of customers in the system. It is easy to show that $p(n)$, the marginal steady-state probability for the number of customers in the system, can be expressed as

$$p(n) = \frac{1}{H}\prod_{m=1}^{n}\beta(m-1)/\nu(m) \tag{5}$$

where

$$\beta(n) = \sum_{i=1}^{b}\sum_{j=1}^{a}\lambda_j \hat{r}_j p(j,i \mid n), \tag{6}$$

$$\nu(n) = \sum_{i=1}^{b}\sum_{j=1}^{a}\mu_i \hat{q}_i p(j,i \mid n), \tag{7}$$

H is a normalization constant chosen so that $\sum_n p(n) = 1$, and $p(j,i \mid n)$ denotes the steady-state conditional probability of the current arrival and service phases given the number in system. Using the identity $p(j,i,n) = p(j,i \mid n)p(n)$ in the balance equations we obtain explicit equations for $p(j,i \mid n)$. Letting $n \to \infty$, and denoting by $\tilde{p}(j,i)$ the limit of $p(j,i \mid n)$ as $n \to \infty$ in these equations, we get the following equations for the asymptotic probability $\tilde{p}(j,i)$

$$\begin{aligned}\tilde{p}(j,i)[\lambda_j + \mu_i] &= \sum_{l=1}^{j-1}\lambda_l r_{lj}\tilde{p}(l,i) + \tau_j\tilde{\nu}\sum_{l=1}^{a}\lambda_l\hat{r}_l\tilde{p}(l,i)/\tilde{\beta} \\ &+ \sum_{l=1}^{i-1}\mu_l q_{li}\tilde{p}(j,l) + \sigma_i\tilde{\beta}\sum_{l=1}^{b}\mu_l\hat{q}_l\tilde{p}(j,l)/\tilde{\nu}\end{aligned} \tag{8}$$

where

$$\tilde{\beta} = \sum_{i=1}^{b}\sum_{j=1}^{a}\lambda_j\hat{r}_j\tilde{p}(j,i) \tag{9}$$

and

$$\tilde{\nu} = \sum_{i=1}^{b}\sum_{j=1}^{a}\mu_i\hat{q}_i\tilde{p}(j,i). \tag{10}$$

Note that there is no approximation involved in the above derivation.

Consider now the particular case of an $E_a/E_b/1$ queue. As discussed in Section 3, because of the sequential nature of the Erlang distribution, when there is just one customer and its service is in its first phase, it is very likely that the arrival process is also in its first stage. However, as the number of customers in the queue increases, there is less and less link between the current stage of service and process.

Hence, it is intuitively clear that, for an arbitrary *Ph/Ph/1* queue, as $n \to \infty$, the knowledge of the current phase of the service process provides less and less information on the current phase of the arrival process (and vice versa). Therefore, the probabilities of the current phases of arrival and service processes must become independent in the limit so that

$$\tilde{p}(j,i) = \tilde{f}(j)\tilde{g}(i) \tag{11}$$

where $\tilde{f}(j)$ is the limiting probability that phase of the arrival process is j, and $\tilde{g}(i)$ is the limiting probability that the phase of the service process is i.

Using the product form of the asymptotic probabilities $\tilde{p}(j,i)$ in (8), (9) and (10), and summing over all values of the current phase of the arrival process j ($j = 1,...,a$), we readily obtain after simple manipulation

$$\tilde{g}(i)[\mu_i + \tilde{\beta}] = \tilde{g}(i)\tilde{\nu} + \sum_{l=1}^{i-1}\mu_l q_{li}\tilde{g}(l) + \sigma_i\tilde{\beta}, \tag{12}$$

for $i = 1,...,b$.

Similarly, summing over all values of the current phase of the service process i ($i = 1,...,b$), we obtain after simple manipulation

$$\tilde{f}(j)[\lambda_j + \tilde{\nu}] = \tilde{f}(j)\tilde{\beta} + \sum_{l=1}^{j-1}\lambda_l r_{lj}\tilde{f}(l) + \tau_j\tilde{\nu}, \tag{13}$$

for $j = 1,...,a$.

Equation (12) turns out to be identical to the asymptotic equation for the *M/Ph/1* queue, and equation (13) is identical to the asymptotic equation for the *Ph/1/M* queue. Thus, by iterating between the solutions of these two queues, we are in effect solving iteratively the exact asymptotic equations for the *Ph/Ph/1* queue.

B. H2 distributions used in Examples B and C in Section 3

The mean service time is kept at 1. The parameters of the H_2 distributions for the service time are given in the following table.

Table 3. Parameters of the service time distributions in Example B

cv	μ_1	μ_2	σ_1	$\sigma_2 = 1-\sigma_1$
2	8.00e-002	1.150e+000	1.121e-002	9.8879e-01
4	2.353e-002	1.2206e+000	4.340e-003	9.9566e-01
8	6.150e-003	1.2313e+000	1.206e-003	9.9879e-01
16	1.556e-003	1.2480e+000	3.097e-004	9.9969e-01

For a mean time between arrivals of 1, the parameters of the H_2 distribution for the inter-arrival time are given in Table 4.

Table 4. Parameters of the arrival distributions in Examples B and C

cv	λ_1	λ_2	τ_1	$\tau_2 = 1-\tau_1$
2	5.714e-002	1.11e+000	5.480e-003	9.9452e-01
4	1.681e-002	1.1471e+000	2.187e-003	9.9781e-01
8	4.396e-003	1.1615e+000	6.136e-004	9.9939e-01
16	1.112e-003	1.16537e+000	1.579e-004	9.9984e-01
20	7.125e-004	1.16584e+000	1.014e-004	9.9990e-01

For higher times between arrivals, the rates λ_1, λ_2 are scaled down proportionately, e.g., for a mean time between arrivals of 2, these rates are doubled. Other parameters, such as τ_1, τ_2, remain unchanged.