First-Come-First-Served Queues with Multiple Servers and Customer Classes

Alexandre Brandwajn, Thomas Begin

First-Come-First-Served Queues with Multiple Servers and Customer Classes

Alexandre Brandwajn

PALLAS International Corporation, San Jose, CA, USA

Thomas Begin

Université de Lyon, UCBL, ENS Lyon, INRIA, CNRS, LIP UMR 5668, France

Abstract

We present a simple approach to the solution of a multi-server FCFS queueing system with several classes of customers and phase-type service time distributions. The proposed solution relies on solving a single two-class model in which we distinguish one of the classes and we aggregate the remaining customer classes. We use a reduced state approximation to solve this two-class model. We propose two types of aggregation: exact, in which we merge the phase-type service time distributions exactly, and approximate, in which we simplify the phase-type distribution for the aggregated class by matching only its first two moments. The proposed approach uses simple mathematics and is highly scalable in terms of the number of servers, the number of classes, as well as the number of phases per class. Our approach applies both to queues with finite and infinite buffer space.

Keywords: Multiple servers, Multiple classes, Phase-type distribution, First-Come First-Served discipline, Reduced-state approximation, Exact class aggregation, Approximate class aggregation.

1. Introduction

The First-Come-First-Served (FCFS) queueing order is perhaps the most "natural" service discipline in queueing systems. Multi-server queues with several customer classes and such FCFS queueing discipline can be found in many areas of life, including computer systems and computer networks. Despite the wide-spread use of these queues, there appears to be a limited number of results available in the literature.

Chow [1] proposed an analytical solution in the considerably simpler case of a single server queue with multiple customer classes and Poisson arrivals. However, his solution is limited to the case of exponential service times and it becomes tedious when the number of classes exceeds 2. More recently, Takine [2]

developed a solution for a single server queue with multiple customer classes, general service times (different for each class) and general inter-arrival time distributions.

Few studies seem to exist in the case of a FCFS queueing system with multiple customer classes and multiple servers. In fact, the presence of multiple servers, multiple customer classes with non-memoryless service times and no service priorities has led authors like Federgruen and Groenevelt [3] to state that "Exact evaluation of the performance vector of even a simple priority rule like FIFO is not possible in the general *MI/GI/c* model.", due to the explosion of the classical state description used in such systems. Nonetheless, in 2000, Van Harten and Sleptchenko [4] proposed a solution based on the classical state description for the case of a multi-server queue with multiple customers classes, exponentially distributed service times distinct for each class and Poisson arrivals. In 2004, Raz et al. [5], presented an analysis of the fairness in queueing systems under various priority policies (including the FCFS discipline) in which they used the "tagged customer" approach to derive the solution in the case of 2 servers and 2 classes with exponentially distributed service times and Poisson arrivals. In the same year, Van Houdt and Blondia [6] derived the delay distribution for a FCFS queue with MMAP arrivals and multiple customer classes with distinct phase-type service distributions in the case of 1 or 2 servers. Compared to previous work, this paper presents a significantly improved method to compute the delay distribution in such queues, albeit limited to systems with unrestricted buffer space and the number of servers not exceeding 2.

This relative paucity of results seems to be due to the intrinsic complexity of the classical state description in FCFS queues, which requires a vector whose elements are the classes of customers at each queue position (cf. [7]). Naturally, such a state description leads to a combinatorial explosion of the number of states as the number of customer classes increases. Unless the service times are exponentially distributed, this complexity is on top of the complexity inherent in the description of the state of the servers themselves (cf. [8]).

The contribution of this paper is to present a mathematically simple approach to the computation of the steady-state queue length distribution in FCFS queues with potentially large numbers of homogeneous servers and arbitrary number of customer classes with distinct general service times. We use a novel simplified state description to allow us to circumvent the complexity of such a queueing system and obtain an accurate approximate solution in which the number of equations to solve grows linearly with the number of servers and the number of classes. Although much of our work is devoted to memoryless arrivals, we present also an extension of our work to a specific class of phase-type arrivals.

Our paper is organized as follows. Section 2 is devoted to the solution of a multi-server FCFS queue with 2 customer classes with general (phase-type) service times and memoryless arrivals. In Section 3, we use the solution derived in the preceding section as a building block to handle an arbitrary number of customer classes. Section 4 presents an extension of our approach to a class of phase-type arrivals. Finally, Section 5 concludes this paper.

2. Solution with two customer classes

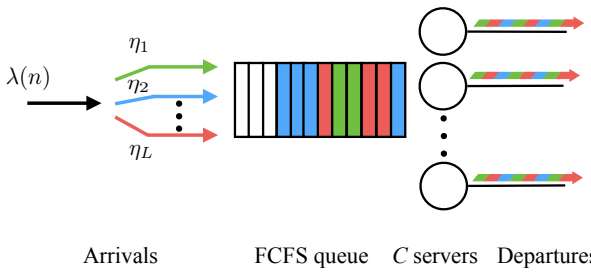


Figure 1: Multi-server FCFS queue with multiple classes of customers.

We consider the queueing system shown in Figure 1. There are C homogeneous servers serving a single queue of customers. Times between customer arrivals are distributed according to a memoryless distribution with rate $\lambda(n)$ where n is the current number of customers in the system. For systems with finite buffer capacity, we denote by N the maximum total number of customers in the system (queued and in service). There is a total of L customer classes and η_ℓ is the probability that an arriving customer is of class $\ell = 1, \ldots, L$, independently of the current state of the system. Customers queue in the order of their arrival and any available server starts serving the customer at the head of the queue (FCFS queueing discipline). We assume that each customer class has its own phase-type service time distribution. As is well known, any distribution can be represented arbitrarily closely by a phase-type distribution [9, 10]. Figure 2 illustrates such a phase-type distribution for customers of class ℓ ($\ell = 1, \ldots, L$). There a total of b_ℓ exponential phases, each with rate (intensity) $\mu_{\ell,i} (i = 1, \ldots, b_\ell)$. Referring to class ℓ, we denote by $\sigma_{\ell,i}$ the probability that a customer service starts in phase i, and by $q_{\ell,ij}$ the probability that the service continues in phase j following the completion of phase i. $\hat{q}_{\ell,i}$ is the probability that the service ends after the completion of phase i and the customer leaves the system. We denote by T_ℓ the mean service time for a class ℓ customer. Table 1 summarizes the main notation used in this paper.

In this section, we consider the case where there are only $L = 2$ customer classes. Let $m = \min(n, C)$ be the current number of busy servers. Following the idea of the reduced state description for *M/Ph/c* queues [8], we represent in detail the phase-type service at a single arbitrarily selected server. Thus, we describe the state of the system by the vector (m_1, ℓ, i, n) where $m_1 = 0, \ldots, m-1$ is the current number of class 1 customers in service at servers other than the selected server, ℓ is the class of the current customer at the selected server, i describes the current service phase of the latter and n is the current number of

Table 1: Notation used.

Symbol	Description
C	Number of servers
N	Maximum total number of customers in the system
L	Number of customer classes
η_ℓ	Probability that an arriving customer is of class ℓ
n	Current number of customers in the system
m	Current number of busy servers
m_1	Current number of class 1 customers in service at servers other than the selected server
$\lambda(n)$	Arrival rate given the current number of customer is n
ℓ	Class of the current customer at the selected server
i	Current service phase of the selected server
b_ℓ	Number of exponential phases in the service time distribution for a class ℓ customer
$\mu_{\ell,i}$	Rate of phase i for a class ℓ customer
$\sigma_{\ell,i}$	Probability of entering phase i upon starting serving a class ℓ customer
$q_{\ell,ij}$	Probability of following to phase j upon completing phase i of a class ℓ customer
$\hat{q}_{\ell,i}$	Probability of ending service upon completing phase i of a class ℓ customer
T_ℓ	Mean service time for a class ℓ customer
$p(m_1,\ell,i,n)$	Probability that the current state of the system is (m_1,ℓ,i,n)
$p(m_1,\ell,i\|n)$	Conditional probability that there are m_1 class 1 customers in service at servers other than the selected server and that the current state of the selected server is (ℓ,i) given that the total number of customers in the system is n
$p(n)$	Probability that the current number of customers in the system is n
$\nu_k(m_1,\ell,i,n)$	Rate of departures of class k customers from servers other than the selected server given that the current system state is (m_1,ℓ,i,n)
$\gamma(n)$	Rate of departures at the selected server given that the current total number of customers in the system is n
$u(n)$	Total rate of customer departures when there are n customers in the system
$\xi_\ell(n)$	Conditional rate of departures from the selected server given that it is busy serving a class ℓ customer and there is a total of n customers in the system
$\overline{n}$	Mean number of customers in the system (regardless of the customer class)
$\overline{U}$	Mean number of busy servers (regardless of the customer class)
θ	Attained throughput (regardless of the customer class)
W	Mean response time (regardless of the customer class)
Q	Mean waiting time (regardless of the customer class)
$\overline{n}_k$	Mean number of class k customers in the system
$\overline{U}_k$	Mean number of busy servers serving class k customers
θ_k	Attained throughput for class k customers
W_k	Mean response time for class k customers
Q_k	Mean waiting time for class k customers

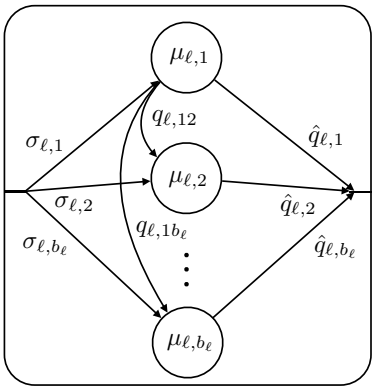


Figure 2: Phase-type distribution for the service times of class ℓ customers.

customers in the system (queued and in service). The possible values for ℓ are

$$\ell = \begin{cases} 0 & \text{if the selected server is idle} \\ 1 & \text{if the selected server is serving a customer of class 1} \\ 2 & \text{if the selected server is serving a customer of class 2,} \end{cases}$$

and $i = 1, \ldots, b_\ell$ where by convention we set $b_0 = 1$.

We consider the system in its steady state (if it exists) and we denote by $p(m_1, \ell, i, n)$ the probability that the current state of the system is (m_1, ℓ, i, n). We also denote by $p(m_1, \ell, i|n)$ the corresponding conditional probability that there are m_1 class 1 customers in service at servers other than the selected server and that the current state of the selected server is (ℓ, i) given that the total number of customers in the system is n. We have $p(m_1, \ell, i, n) = p(m_1, \ell, i|n)p(n)$ where $p(n)$ is the marginal probability for n. We must have $\sum_{m_1=0}^{m-1} \sum_{\ell=0}^{2} \sum_{i=1}^{b_\ell} p(m_1, \ell, i|n) = 1$ for all values of n. Note that the value $\ell = 0$ is possible only when $m < C$ and it is the only value possible when $m = 0$.

It is not difficult to derive the balance equations for the system considered. As an example, for $n > C$, $0 < m_1 < C - 1$, $\ell = 1, 2$ and $i = 1, \ldots, b_\ell$ we have

$$\begin{aligned} &p(m_1, \ell, i, n)[\lambda(n) + \mu_{\ell,i} + \nu_1(m_1, \ell, i, n) + \nu_2(m_1, \ell, i, n)] \\ &\quad = p(m_1, \ell, i, n-1)\lambda(n-1) + \sum_{k=1}^{2} \sum_{i=1}^{b_k} p(m_1, k, i, n+1)\mu_{k,i}\hat{q}_{k,i}\eta_\ell\sigma_{\ell,i} \\ &\qquad + \sum_{j=1}^{b_\ell} p(m_1, \ell, j, n)\mu_{\ell,j}q_{\ell,ji} + \sum_{k=1}^{2} p(m_1, \ell, i, n+1)\nu_k(m_1, \ell, i, n)\eta_k \\ &\qquad\qquad + p(m_1+1, \ell, i, n+1)\nu_1(m_1+1, \ell, i, n+1)\eta_2 \\ &\qquad\qquad\qquad + p(m_1-1, \ell, i, n+1)\nu_2(m_1-1, \ell, i, n+1)\eta_1. \quad (1) \end{aligned}$$

Note that in our equations the probability that a departing customer of any class is replaced by a customer of class ℓ is given simply by η_ℓ (here $\ell = 1, 2$) . The quantities $\nu_k(m_1, \ell, i, n)$ represent the rate of departures of class k customers ($k = 1, 2$) from servers other than the selected server given that the current system state is (m_1, ℓ, i, n). These quantities are not given a priori and will be approximated in our solution.

Let $\gamma(n)$ be the rate of departures (completions) at the selected server given that the current total number of customers in the system is n. We have

$$\gamma(n) = \sum_{m_1=0}^{m-1} \sum_{\ell=1}^{2} \sum_{i=1}^{b_\ell} p(m_1, \ell, i|n) \mu_{\ell,i} \hat{q}_{\ell,i}. \tag{2}$$

Since the servers are homogenous and thus statistically identical, the total rate of customer departures when there are n customers in the system, denoted by $u(n)$, is given by $u(n) = C\gamma(n)$. If it exists, the steady-state probability $p(n)$, can be expressed as

$$p(n) = \frac{1}{G} \prod_{j=1}^{n} \frac{\lambda(j-1)}{u(j)} \qquad \text{for } n = 0, 1, \ldots \tag{3}$$

where G is a normalizing constant such that $\sum_{n\geq 0} p(n) = 1$.

Denote by $\xi_\ell(n)$ the conditional rate of departures from the selected server given that it is busy serving a class ℓ customer ($\ell = 1, 2$) and there is a total of n customers in the system. We have

$$\xi_\ell(n) = \frac{\sum_{m_1=0}^{m-1} \sum_{i=1}^{b_\ell} p(m_1, \ell, i|n) \mu_{\ell,i} \hat{q}_{\ell,i}}{\sum_{m_1=0}^{m-1} \sum_{i=1}^{b_\ell} p(m_1, \ell, i|n)}. \tag{4}$$

We approximate the unknown rates of departures from other servers $\nu_k(m_1, \ell, i, n)$ as

$$\nu_k(m_1, \ell, i, n) \simeq m_k \xi_k(n) \quad \text{for } k = 1, 2 \quad \text{where } m_2 = \begin{cases} m - 1 - m_1 & \text{if } \ell > 0 \\ m - m_1 & \text{if } \ell = 0. \end{cases} \tag{5}$$

We believe that the approximate computation of the departure rates from servers other than the selected server is the only approximation in our solution of a FCFS multi-server queue with two classes of customers. Refer to the Appendix for an additional discussion of this approximation. Using the identity $p(m_1, \ell, i, n) = p(m_1, \ell, i|n)p(n)$ together with formula (3) relating the probabilities $p(n)$ to the arrival and conditional completion rates $\lambda(n)$ and $u(n)$, we can transform the balance equations for $p(m_1, \ell, i, n)$ into equations for the conditional probabilities $p(m_1, \ell, i|n)$.

The resulting system of equations for $p(m_1, \ell, i|n)$ can be solved using a straightforward fixed-point iteration (see Appendix). We do not have a formal proof of convergence of our iterative scheme to a unique solution. However, in the many examples we ran, we never encountered any convergence problems.

For systems with finite buffer capacity, the number of equations to solve is determined by the maximum number of customer in the system, N. For systems with infinite buffer, the number of equations to solve at each iteration is determined by the speed of convergence of the probabilities $p(m_1, \ell, i|n)$ to their asymptotic values as n increases. This is analogous to the approach used by the authors for *M/Ph/c* queues (cf. [11]). Like in the latter, the number of equations to solve in our case grows linearly with the number of servers and the number of phases in service time distributions.

Having obtained the conditional probabilities $p(m_1, \ell, i|n)$, we readily get the conditional rate of customer departures for the selected server $\gamma(n)$ from formula (2), the overall conditional rate of completions $u(n)$ and the steady-state probabilities $p(n)$ from formula (3). Hence, we can obtain the following performance indices for the system as a whole

- attained throughput $\theta = \sum_{n>0} u(n)p(n)$
- mean number of busy servers $\overline{U} = \sum_{n>0} \min(n, C)p(n)$
- mean number of customers in the system $\overline{n} = \sum_{n>0} np(n)$
- mean response time $W = \overline{n}/\theta$
- mean time in the queue waiting for service $Q = (\overline{n} - \overline{U})/\theta$,

as well as specifically for customers of class 1

- attained throughput for class 1, $\theta_1 = \theta\eta_1$
- mean number of class 1 customers in service, $\overline{U}_1 = \theta_1 T_1$
- mean number of class 1 customers in the system (queued and in service), $\overline{n}_1 = Q\theta_1 + \overline{U}_1$
- mean class 1 response time, $W_1 = \overline{n}_1/\theta_1$.

Relationships between mean numbers and mean times are derived using Little's law [12]. We can readily obtain analogous performance indices for class 2 customers. The reduced state description used in our approach yields directly the steady-state distributions for the overall queue length and for the number of class 1 (and hence class 2) customers in service. We can obtain also the steady-state distribution that there are k_1 ($k_1 = 0, 1, \dots$) customer of class 1 queued for service as $P\{k_1\} = \sum_{n \geq C+k_1} p(n)\eta_1^{k_1}\eta_2^{n-C-k_1}\frac{(n-C)!}{k_1!(n-C-k_1)!}$. Similarly, we can get the steady-state distribution of the number of customers of class 2 waiting for service. Although we have focused on mean values, from the steady-state queue length distribution we can readily obtain higher moments of the number of customers queued.

In the next section we consider a FCFS multi-server queue with an arbitrary number of customer classes.

3. Solution with more than two customer classes

3.1. Exact class aggregation

Consider again the system represented in Figure 1, this time with $L > 2$ customer classes. Let's select any class ℓ, e.g. $\ell = 1$. We keep the selected class separate and we aggregate the remaining $L - 1$ customer classes into a single class whose service time distribution is the result of a merger of these customer classes. In practice, the phase-type distributions of the classes merged are simply combined as branches of the resulting phase-type distribution. The initial phase selection probabilities in the resulting distribution are modified as follows: $\sigma_{k,i}$ for class $k \neq \ell$ becomes $\sigma_{k,i}\eta_k / \sum_{j \neq \ell} \eta_j$. This is illustrated in Figure 3.

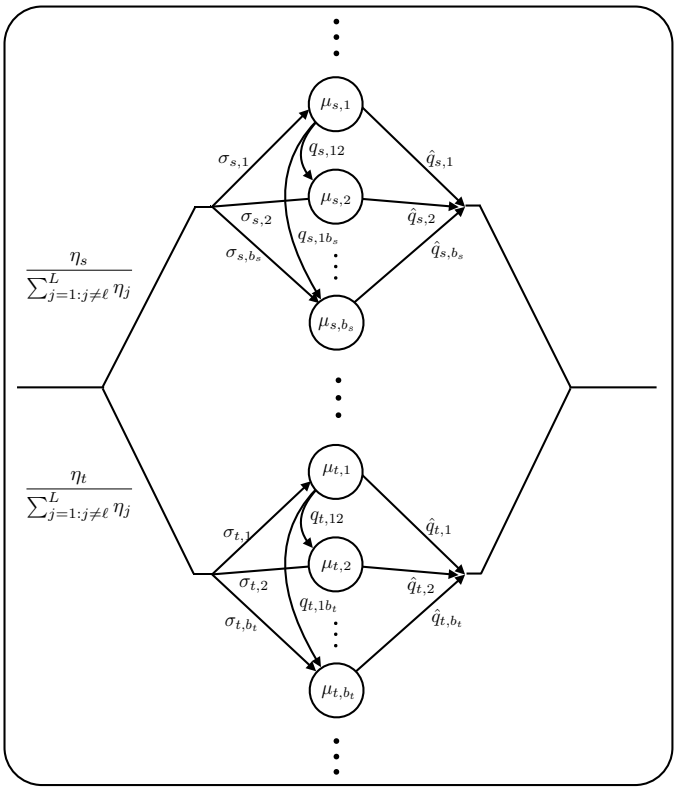


Figure 3: Service time distribution resulting from the aggregation of all classes other than class ℓ.

The performance indices obtained from the solution of the resulting two-class queue (cf. Section 2) include the attained throughput for the system as a whole θ, as well as the mean time in the queue waiting for service Q. From these two

quantities we readily derive desired performance indices for any customer class k, $k = 1, \ldots, L$:

- attained throughput for class k, $\theta_k = \theta\eta_k$
- mean number of class k customers in service, $\overline{U}_k = \theta_k T_k$
- mean number of class k customers in the system, $\overline{n}_k = Q\theta_k + \overline{U}_k$
- mean class k response time, $W_k = \overline{n}_k/\theta_k$.

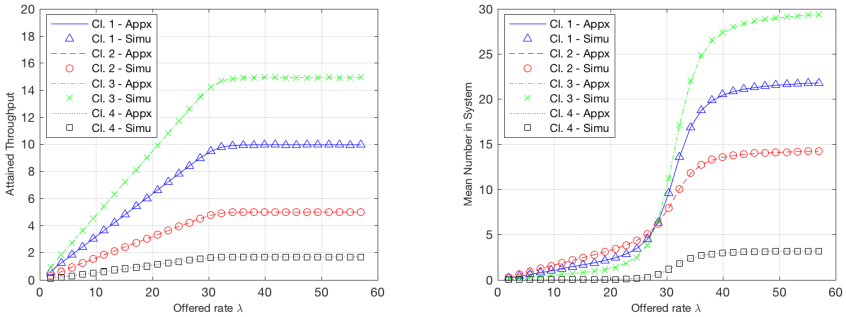
Note that in this approach the customer classes other than the arbitrarily selected class ℓ are aggregated exactly and no approximation is involved in this step.

First example

As an example, consider a queue with $C = 10$ servers, $L = 4$ customer classes and the total number of customers in the system limited to $N = 70$. The customer service times are exponentially distributed with means $T_1 = 1/\mu_{1,1} = 1/3$, $T_2 = 1/\mu_{2,1} = 1/1$, $T_3 = 1/\mu_{3,1} = 1/9$ and $T_4 = 1/\mu_{4,1} = 1/30$, respectively. In this example, arrivals come from a Poisson source with rate λ and the following probabilities for each customer class: $\eta_1 = 6/19$, $\eta_2 = 3/19$, $\eta_3 = 9/19$ and $\eta_4 = 1/19$. This is equivalent to each class having its own Poisson stream of arrivals with a rate $\lambda\eta_\ell$. We solve this system for several load levels ranging from $\lambda = 1.9$ to $\lambda = 57$ (the arrival rate is independent of n in our example). This set of offered load levels covers a spectrum of server utilization values from less than 10% to nearly 100%. For a given load level, we solve a single two-class model, corresponding to class $\ell = 1$ kept separate and the three other customer classes combined into one. Thus, for $\ell = 1$, the first customer class in the two-class model has an exponentially distributed service time with mean 1/3. The service time distribution of the second class is a hyperexponential with three phases (H-3) whose intensities are given by 1, 9 and 30, respectively. The corresponding phase selection probabilities in this H-3 distribution are given by 3/13, 9/13 and 1/13.

Figure 4 shows the numerical results obtained for the four-class queue considered using the proposed approach (with the approximate solution for the two-class model described in Section 2). We show the attained throughput and the mean total number of customers for each class of customers as the selected performance metrics. For comparison, we have also included the results of discrete-event simulations of the same FCFS queue with four customer classes. The simulation results were obtained using the independent replications method [13] with 14 replications of 2,000,000 completions each. The estimated confidence intervals at 95% confidence level being quite narrow, only the middle points of the confidence intervals are shown in our figure. We observe the very close agreement between simulation and the proposed solution in the case studied. The mean and the median relative errors for the mean number of customers in the system were 0.6% and 0.3%, respectively. Table 2 shows the corresponding distribution of relative errors. Relative errors for the attained throughputs

were negligible. As mentioned before, the choice of $\ell = 1$ for the class kept separate in the two-class model is arbitrary. Because the solution presented in Section 2 is approximate, a choice of a different value for ℓ might yield slightly different results. In our example, the relative differences in results were on the order of 0.1%.



(a) Attained throughput per class. (b) Mean number of customers per class.

Figure 4: Accuracy of the proposed solution using the "Exact class aggregation" (Section 3.1) for the first example (exponentially distributed service times, finite buffer).

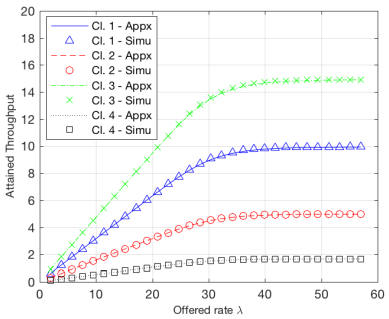
Table 2: Distribution of the relative errors for the mean number of customers per class using the "Exact class aggregation" (Section 3.1) for the first example.

Mean	Median	<1%	<5%	<10%	<25%	≥25%
0.59%	0.28 %	84.17%	99.17%	100.00%	0.00%	0.00%

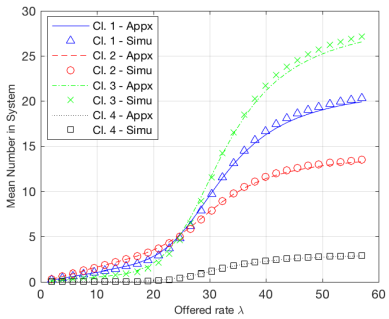
Second example

In our second example, we consider a similar FCFS queue except that this time the service time of each customer class is no longer exponential but represented by H-2 distributions (two exponential branches). The mean service times are as before given by 1/3, 1, 1/9 and 1/30. The squared coefficients of variation of the class service times (defined as the ratio of the variance to the square of the mean) are given by 16, 9, 4 and 2, respectively. The parameters of the corresponding H-2 distributions are as follows: $\sigma_{1,1}$ =0.115, $\sigma_{1,2}$ =0.885, $\mu_{1,1}$ =0.349, $\mu_{1,2}$ =267.706, $\sigma_{2,1}$ =0.392, $\sigma_{2,2}$ =0.608, $\mu_{2,1}$ =0.396, $\mu_{2,2}$ =61.0, $\sigma_{3,1}$ =0.196, $\sigma_{3,2}$ =0.804, $\mu_{3,1}$ =1.782, $\mu_{3,2}$ =729.0 and $\sigma_{4,1}$ =0.653, $\sigma_{4,2}$ =0.347, $\mu_{4,1}$ =19.802, $\mu_{4,2}$ =1030.0. The two-class model solved as part of the proposed solution procedure comprises one class with the original H-2 service time distribution for the selected customer class and one class with an H-6 service time distribution (six exponential branches which represent the combination of the remaining three classes in the original system).

Figure 5 illustrates the numerical results obtained using our approach for the four-class queue with H-2 service time distributions. As before, class $\ell = 1$ was



(a) Attained throughput per class.



(b) Mean number of customers per class.

Figure 5: Accuracy of the proposed solution using the "Exact class aggregation" (Section 3.1) for the second example (non-exponentially distributed service times, finite buffer).

Table 3: Distribution of the relative errors for the mean number of customers per class using the "Exact class aggregation" (Section 3.1) for the second example.

Mean	Median	<1%	<5%	<10%	<25%	≥25%
3.02%	1.75 %	28.33%	84.17%	91.67%	100.00%	0%

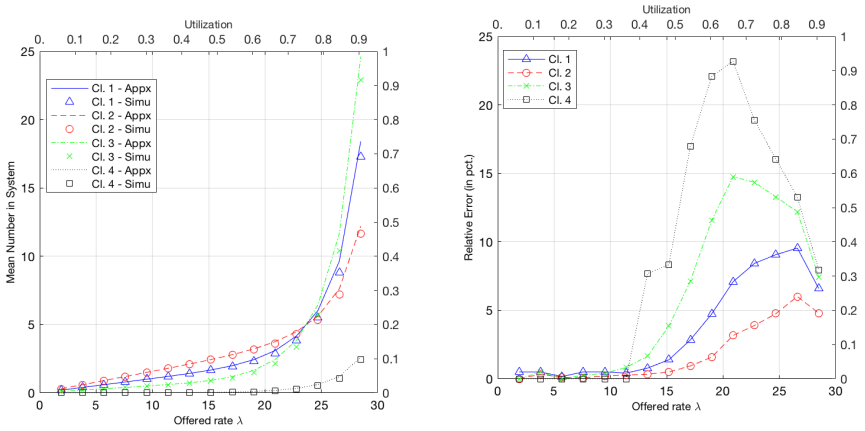
kept separate. We show the throughputs and the mean numbers of customers of each class in the system for the same load levels as in Figure 4. We also include the results of discrete-event simulation for comparison. Again, we observe a close agreement between the results obtained using the proposed approach and simulation. Here, the mean and the median relative errors for the mean number of customers in the system were 3% and 1.7%, respectively. The corresponding distribution of relative errors in Table 3 shows that for this example almost 92% of errors were below 10%. Relative errors for the attained throughput were below 1%.

Comparing Figures 5 and 4 provides an example of the effect of the service time distributions in a FCFS queue with multiple servers. Depending on the load level, relative difference in the mean number of customers may exceed 30% in the example considered. Not surprisingly, the relative differences appear most pronounced for medium load levels. Attained throughputs seem less sensitive to service time distributions in our example.

Third example

In our third example, we consider again the system with four customer classes with non-exponentially distributed service time studied in our second example except that this time the buffer capacity (queueing room) is infinite. We solve this system for values of arrival rate λ ranging from 1.9 to 29 i.e., for server utilization values ranging from some 6% to over 90% (we don't believe that models with infinite buffers are of practical interest at higher server utilization). Figure 6 shows the mean number of customers for each class obtained with

our exact aggregation method, as well as the corresponding relative errors. We observe the generally good agreement with simulation results, even at high server utilization levels. In this example, the mean and the median relative errors for the mean numbers of customers in the system were 5% and 1.6%, respectively. The distribution of relative errors shown in Table 4 indicates that in this example the relative errors are below 10% in over 80% of cases. We notice in Figure 6 that the largest relative errors occur for a class with the smallest mean number of customers at a server utilization of some 60%.



(a) Mean number of customers per class.

(b) Relative errors for the mean number of customers per class.

Figure 6: Accuracy of the proposed solution using the "Exact class aggregation" (Section 3.1) for the third example (non-exponentially distributed service times, infinite buffer).

Table 4: Distribution of the relative errors for the mean number of customers per class using the "Exact class aggregation" (Section 3.1) for the second example.

Mean	Median	<1%	<5%	<10%	<25%	≥25%
5.04%	1.62%	46.67%	63.33%	81.67%	100.00%	0.00%

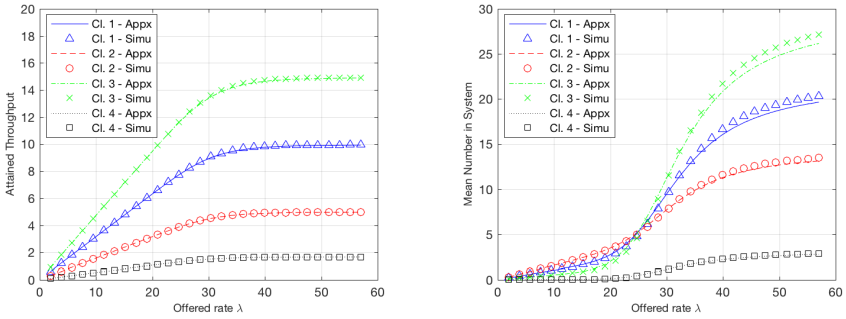
The exact aggregation method is the recommended approach unless the number of phases in the aggregated class becomes intractably large.

3.2. Approximate class aggregation

If the number of phases in the aggregated second class (corresponding to $L-1$ classes in the original L class model) becomes too large to handle, we propose to simplify the aggregated class by replacing it by a simpler distribution with the same first two moments. In practice, if the coefficient of variation of the aggregated second class is greater than $1/\sqrt{2}$, two phases suffice to match the first two moments [14]. We propose to solve the resulting simplified two-class model and use the attained throughput for the system as a whole θ as well as

the mean time in the queue waiting for service Q to derive desired performance indices for each class as described in Section 3.1. This is similar in spirit to the solutions to multiclass problems proposed over the years by several authors (cf. [15]).

Since in our approximate aggregation we are matching only the first two moments of the merged classes and the performance of $M/Ph/c$ queues is known to potentially exhibit higher-order distributional dependencies [16], we expect some loss of accuracy. This is indeed the case and the relative differences versus simulation results seem to increase somewhat compared to exact aggregation. This is illustrated in Figure 7 for our second example, i.e. for the same set of four classes as in Figure 5. The mean and relative errors for the mean number of customers in the system are 3.7% and 3%, respectively (versus 3% and 1.7% previously). The distribution of relative errors in Table 5 indicates that for this example in 92% of cases the relative errors were below 10%. The corresponding values for the attained throughput remain below 1%. We observe that, while the approximation results deviate somewhat from simulation midpoints, they do stay sufficiently close to the latter to be a good approximation. And, of course, the complexity of the resulting two-class model is generally greatly reduced, especially as the number of customer classes increases.



(a) Attained throughput per class. (b) Mean number of customers per class.

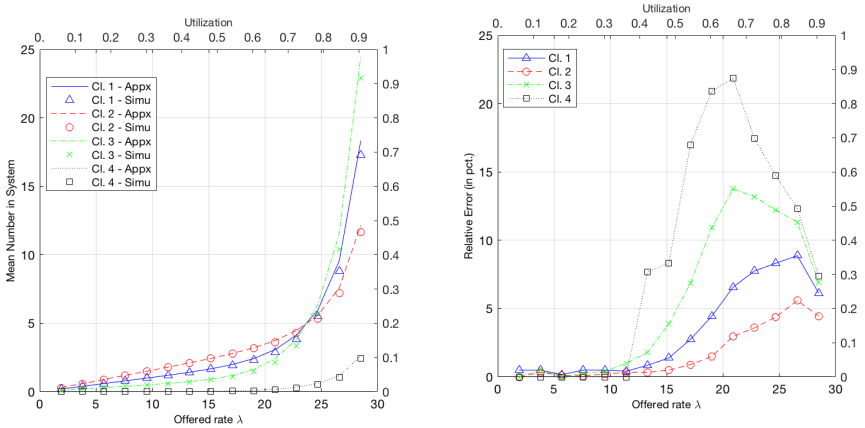
Figure 7: Accuracy of the proposed solution using the "Approximate class aggregation" (Section 3.2) for the second example (non-exponentially distributed service times, finite buffer).

Table 5: Distribution of the relative errors for the mean number of customers per class using the "Approximate class aggregation" (Section 3.2) for the second example.

Mean	Median	<1%	<5%	<10%	<25%	≥25%
3.66%	3.00%	25.00%	84.17%	92.50%	100.00%	0.00%

In Figure 8, we show the numerical results obtained with our approximate aggregation for the model with infinite buffer considered in our third example. We observe again the close agreement between our approximation and simulation results. The mean and median relative errors for the mean number of

customers in the system remain around 5% and 2%, respectively, in the example considered. As shown in Table 6, in this example, the relative errors were below 10% in over 80% of cases. Figure 8b indicates that the largest relative errors occur for a class with a small mean number of customers in the system at a server utilization of close to 60%.



(a) Mean number of customers per class.

(b) Relative errors for the mean number of customers per class.

Figure 8: Accuracy of the proposed solution using the "Approximate class aggregation" (Section 3.2) for the third example (non-exponentially distributed service times, infinite buffer).

Table 6: Distribution of the relative errors for the mean number of customers per class using the "Approximate class aggregation" (Section 3.2) for the third example.

Mean	Median	<1%	<5%	<10%	<25%	≥25%
4.76%	1.64%	46.67%	63.33%	81.67%	100.00%	0.00%

Note that our approximate aggregation approach keeps one customer class intact while preserving only the first two moments of the aggregate of remaining customer classes. In our examples, the resulting accuracy appears sufficient for practical purposes. Clearly, if desired, one can match the first three (or more) moments of the aggregated class using available algorithms [9] (see Appendix).

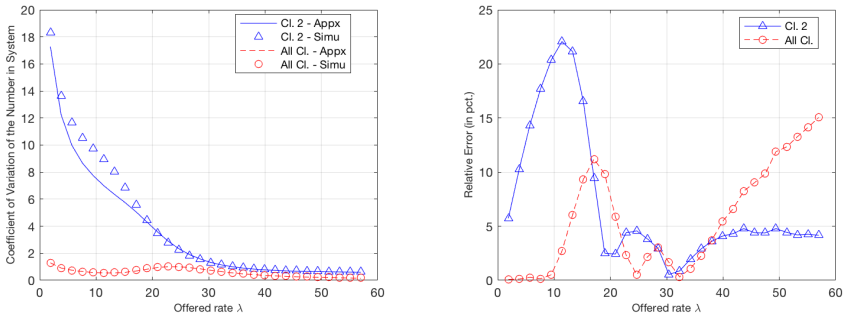
With both exact and approximate aggregation, we solve the two-class model only once, so that the resulting overall complexity in terms of the number of equations to solve grows no more than linearly with the number of classes.

3.3. Coefficients of variation of the number of customers

Formula (3) gives us $p(n)$, the steady-state probability that there are n customers in the system. Thus, we can compute the second moment of the number of customers in the system as $\sum_n n^2 p(n)$. As shown in Section 2, the

state description chosen readily produces the conditional probabilities for the first class in our two-class system that there are m_1 customers in service, as well as k_1 customers in the queue given the total number of customers n. Hence, we obtain the conditional second moments of the number in the queue and in service for the selected customer class given n and the corresponding conditional variances. As an approximation, we assume that the conditional variance of the number of class 1 customers in the system given n can be evaluated as the sum of the latter conditional variances. We then use the law of total moments to evaluate the non-conditional second moment of the total number of customers of the selected class.

As an example, we show in Figure 9 the results obtained for the coefficients of variation for the number of class 2 customers, as well as for the total number of customers in the system computed from our approximate solution for Example 2.



(a) Coefficient of variation of the number of customers.

(b) Relative errors for the coefficient of variation of the number of customers.

Figure 9: Accuracy of the proposed solution for the coefficient of variation of the number of customers in the system using the "Exact class aggregation" (Section 3.1) for the second example (non-exponentially distributed service times, finite buffer)

In this example, while some of the relative errors for smaller values of offered load can approach 25%, the mean and the median relative errors for the approximate coefficient of variation of class 2 remain below 10%. The mean and median relative errors for the coefficient of variation of the total number of customers in the system are below 6%.

In the next section we consider an extension of our approach to a class of phase-type arrivals.

4. Extension to phase-type arrivals

In this section, we briefly discuss an extension of our model to a more general arrival process. Specifically, we consider a queue in which the times between consecutive customer arrivals are distributed according to a phase-type distribution with a exponential phases. As before, we assume that the probability that an arriving customer is of class ℓ is given by η_ℓ, $\ell = 1, \ldots, L$, regardless of

the state of the system. Obviously, unlike in the case of Poisson arrivals, a superposition of individual phase-type times between arrivals, in general, does not correspond to the same type of arrival process unless the class arrival processes are phase synchronized.

Since our reduced-state solution of Section 2 produces explicitly the conditional rate of completions $u(n)$, we propose to apply the simple approach presented in a recent paper on modeling of cloud systems [17]. This approach decomposes the solution of a system with non-memoryless arrivals into two solutions of simpler models: a model in which the arrivals are represented as memoryless with a state-dependent rate $w(n)$ and a model with the original phase-type arrivals and the service process represented by a state-dependent service rate $u(n)$. This is represented in Figure 10.

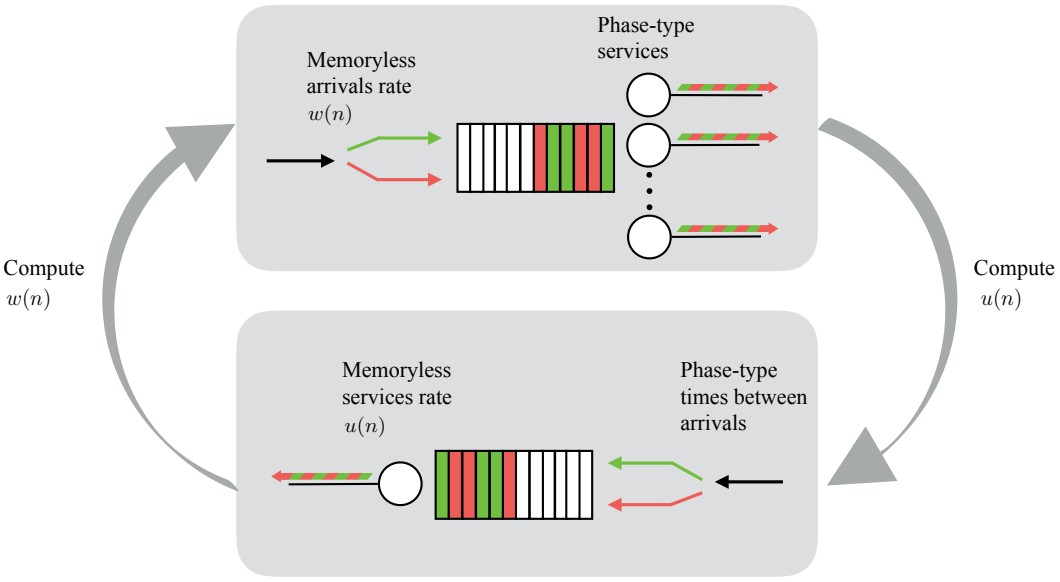


Figure 10: Schematic view of the solution in case of phase-type times between arrivals.

Note that the fixed-point iteration between models shown in Figure 10 needs to happen only for a single two-class model in our solution procedure. The next section concludes this paper.

5. Conclusions

We have presented a simple approach to the solution of a FCFS queueing system with memoryless arrivals, multiple servers and several classes of customers with distinct phase-type service time distributions. The proposed solution procedure relies on solving a single two-class model in which one of the classes has the same parameters as in the original multiclass model while the other is the

result of a merger (aggregation) of the remaining customer classes. With exact aggregation, we merge the phase-type service time distributions exactly into a phase-type distribution with a higher number of phases. If the resulting number of phases becomes unmanageable, we propose to use approximate aggregation in which we simplify the phase-type distribution for the aggregated class by matching only the first two moments of the aggregated customer class. Our solution of the two-class model involves a fixed-point iteration. Although we do not have a theoretical proof of convergence of the fixed point to a unique solution, in practice, in the many examples we ran, it never failed to converge.

The proposed approach uses only simple mathematics together with basic flow conservation ideas and fundamental queueing properties such as Little's law. Because it relies on the reduced state description for the solution of the two-class FCFS queue, the proposed approach is highly scalable in terms of the number of servers, the number of classes, as well as the number of phases per class. Our approach applies to queues with finite and infinite buffer space. It readily yields class performance measures such as the attained throughput and the mean response time, as well as distributions for the overall number of customers in the system or in the queue.

Acknowledgements

The authors wish to thank the anonymous referees for their thorough and constructive review of an earlier version of this paper.

Appendix A. Solving the balance equations

The example equation in Section 2 for $n > C$, $0 < m_1 < C-1$, $\ell = 1, 2$ and $i = 1, \ldots, b_k$ can be transformed into the following equation for conditional probabilities

$$
\begin{aligned}
&p(m_1,\ell,i|n)[\lambda(n)+\mu_{\ell,i}+\nu_1(m_1,\ell,i,n)+\nu_2(m_1,\ell,i,n)]\\
&= p(m_1,\ell,i|n-1)u(n) + \sum_{k=1}^{2}\sum_{i=1}^{b_k} p(m_1,k,i|n+1)\mu_{k,i}\hat{q}_{k,i}\eta_\ell\sigma_{\ell,i}\lambda(n)/u(n+1)\\
&\qquad + \sum_{j=1}^{i-1} p(m_1,\ell,j|n)\mu_{\ell,j}q_{\ell,ji} + \sum_{j=i}^{b_\ell} p(m_1,\ell,j|n)\mu_{\ell,j}q_{\ell,ji}\\
&\qquad + \sum_{k=1}^{2} p(m_1,\ell,i|n+1)\nu_k(m_1,\ell,i,n)\eta_k\lambda(n)/u(n+1)\\
&\quad + p(m_1+1,\ell,i|n+1)\nu_1(m_1+1,\ell,i,n+1)\eta_2\lambda(n)/u(n+1)\\
&\quad + p(m_1-1,\ell,i|n+1)\nu_2(m_1-1,\ell,i,n+1)\eta_1\lambda(n)/u(n+1). \qquad \text{(A.1)}
\end{aligned}
$$

In a similar way, we can obtain equations for values of $n = 1, \ldots, C$.

If we sum equation (A.1) over all values of m_1, ℓ and i, we get

$$\lambda(n) \sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n) + \gamma(n)$$
$$+ \sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n)[\nu_1(m_1, \ell, i, n) + \nu_2(m_1, \ell, i, n)]$$
$$= u(n) \sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n-1) + \lambda(n)\{\gamma(n+1)$$
$$+ \sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n+1)[\nu_1(m_1, \ell, i, n+1) + \nu_2(m_1, \ell, i, n+1)]\}/u(n+1) \tag{A.2}$$

Recall that $u(n) = C\gamma(n)$. For values of $n > C$ considered in equation (A.1), if we have, as must be, $\sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n-1) = 1$, then requiring that $\sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n)[\nu_1(m_1, \ell, i, n) + \nu_2(m_1, \ell, i, n)] = (C-1)\gamma(n)$ and $\sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n+1)[\nu_1(m_1, \ell, i, n+1) + \nu_2(m_1, \ell, i, n+1)] = (C-1)\gamma(n+1)$ allows us to normalize the probabilities $p(m_1, \ell, i|n)$ (so that $\sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n) = 1$).

In the particular case when the two customer classes in the model of Section 2 have exponentially distributed service times with parameters $\mu_{1,1}$ and $\mu_{2,1}$, respectively, formulas (4) and (5) yield simply $\xi_\ell(n) = \mu_{\ell,1}$ and $\nu_k(m_1, \ell, i, n) = m_k \mu_{k,1}$. This is the expected result and the relationship $\sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n)[\nu_1(m_1, \ell, i, n) + \nu_2(m_1, \ell, i, n)] = (C-1)\gamma(n)$ happens to hold exactly. For general service times, formulas (4) and (5) introduce an approximation that results in a slight violation of this relationship. Therefore, in actual computation we simply scale the values of $\nu_k(m_1, \ell, i, n)$ computed using formulas (4) and (5) so as to have the proper values for $\sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n)[\nu_1(m_1, \ell, i, n) + \nu_2(m_1, \ell, i, n)]$.

In equation (A.1), if we treat $u(n)$ as an independent parameter, we notice that the values of $p(m_1, \ell, i|n)$ are increasing functions of $u(n)$. It is not difficult to show that the sum $\sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n)$ as a function of $u(n)$ increases from a value less than 1 for $u(n) = 0$ to arbitrarily large values for $u(n) \to \infty$. This suggests that there can be only one value of $u(n)$ such that $\sum_{m_1} \sum_{\ell} \sum_{i} p(m_1, \ell, i|n) = 1$.

A simple fixed-point solution of the equations for the conditional probabilities $p(m_1, \ell, i|n)$ might proceed as follows. Consider system states in the order of increasing $n = 1, 2, \ldots$, for each value of n, enumerate the states in the order of increasing m_1, for each value of m_1, in the order of increasing values of ℓ, and for each value of ℓ, in the order of increasing $i = 1, \ldots, b_\ell$. Denote by the superscript t the current iteration number. We start with a feasible initial distribution $p^0(m_1, \ell, i|n)$ and the corresponding set of $u^0(n)$, $\xi_\ell^0(n)$ and hence $\nu_k^0(m_1, \ell, i, n)$ for $n = 1, 2, \ldots$. In the case of the equations given above for $n > C$, $0 < m_1 < C-1$, $\ell = 1, 2$ and $i = 1, \ldots, b_k$, we can compute values at iteration $t = 1, 2, \ldots$ as

$$p^t(m_1,\ell,i|n) = 1/[\lambda(n) + \mu_{\ell,i} + \nu_1^{t-1}(m_1,\ell,i,n) + \nu_2^{t-1}(m_1,\ell,i,n)].$$

$$[p^t(m_1,\ell,i|n-1)u^{t-1}(n) + \sum_{k=1}^{2}\sum_{i=1}^{b_k} p^{t-1}(m_1,k,i|n+1)\mu_{k,i}\hat{q}_{k,i}\eta_\ell\sigma_{\ell,i}\lambda(n)/u^{t-1}(n+1)$$

$$+ \sum_{j=1}^{i-1} p^t(m_1,\ell,j|n)\mu_{\ell,j}q_{\ell,ji} + \sum_{j=i}^{b_\ell} p^{t-1}(m_1,\ell,j|n)\mu_{\ell,j}q_{\ell,ji}$$

$$+ \sum_{k=1}^{2} p^t(m_1,\ell,i|n+1)\nu_k^{t-1}(m_1,\ell,i,n)\eta_k\lambda(n)/u^{t-1}(n+1)$$

$$+ p^{t-1}(m_1+1,\ell,i|n+1)\nu_1^{t-1}(m_1+1,\ell,i,n+1)\eta_2\lambda(n)/u^{t-1}(n+1)$$

$$+ p^{t-1}(m_1-1,\ell,i|n+1)\nu_2^{t-1}(m_1-1,\ell,i,n+1)\eta_1\lambda(n)/u^{t-1}(n+1)]. \quad \text{(A.3)}$$

For each value of n (in the case where $n > C$), we must have $\sum_{m_1=0}^{m-1}\sum_{\ell=1}^{2}\sum_{i=1}^{b_\ell} p^t(m_1,\ell,i|n) = 1$. We use this relationship to normalize the values obtained for a given n. Having normalized the probabilities $p^t(m_1,\ell,i|n)$, we compute new values for $u^t(n)$ and $\xi_\ell^t(n)$ using formulas (2) and (4) (and hence $\nu_k^t(m_1,\ell,i,n)$ from formula (5)). Then, we move on to the next value of n. The fixed-point iteration for values of $n = 1,\ldots,C$ proceeds in an analogous manner.

We stop the iteration when, for instance, $|\frac{u^{t-1}(n)}{u^t(n)} - 1| < \epsilon$ for all $n = 1, 2, \ldots,$ where $\epsilon > 0$ is the desired convergence stringency (e.g. $\epsilon = 10^{-6}$).

Appendix B. Number of equations solved at each iteration

With the state description chosen, the number of equations solved at each iteration is given by:

$$NC(b_1 + b_2) + (C-1) - (C-1)C(b_1 + b_2 - 1)/2 \quad \text{(B.1)}$$

where b_1 is the number of phases of the first class in our two-class model. For the exact aggregation approach, b_2 is the sum of the numbers of phases of all the other customer classes. For the approximate aggregation, b_2 is the number of phases chosen to represent the aggregate of all the other classes (in our case, it is most often 2). In the case of infinite buffer capacity, N is replaced by the value of the number of customers n at which the conditional probabilities become sufficiently close to their asymptotic values (typically, between 100 and 1000).

Thus, the complexity in terms of the number of equations to solve increases linearly with the number of servers and at most linearly with the number of service phases.

Appendix C. Approximate aggregation matching three moments

As mentioned in Section 3.2, one can use existing algorithms to match more than the first two moments of the aggregated user class. For our examples, the results of matching the first three moments tend to be marginally better than using only two moments, but the improvement is not always uniform. For instance, the mean and the median relative errors may be lower but there may be higher maximum errors. This is likely due to the fact that errors introduced in limiting the approximate aggregation to the first two moments sometimes compensate errors introduced by the reduced state approximation.

References

[1] W.-M. Chow, Central server model for multiprogrammed computer systems with different classes of jobs, IBM Journal of Research and Development 19 (3) (1975) 314–320.

[2] T. Takine, Queue length distribution in a FIFO single-server queue with multiple arrival streams having different service time distributions, Queueing Systems 39 (4) (2001) 349–375.

[3] A. Federgruen, H. Groenevelt, M/G/c queueing systems with multiple customer classes: characterization and control of achievable performance under nonpreemptive priority rules, Management Science 34 (9) (1988) 1121–1138.

[4] A. Van Harten, A. Sleptchenko, On multi-class multi-server queueing and spare parts management, Queueing systems 43 (4) (2003) 307–328.

[5] D. Raz, B. Avi-Itzhak, H. Levy, Classes, priorities and fairness in queueing systems, RUTCOR, Rutgers University, Tech. Rep. RRR-21-2004.

[6] B. Van Houdt, C. Blondia, The waiting time distribution of a type k customer in a discrete-time MMAP[K]/PH[K]/c (c= 1, 2) queue using QBDs, Stochastic models 20 (1) (2004) 55–69.

[7] F. Baskett, K. M. Chandy, R. R. Muntz, F. G. Palacios, Open, closed, and mixed networks of queues with different classes of customers, Journal of the ACM (JACM) 22 (2) (1975) 248–260.

[8] A. Brandwajn, T. Begin, Reduced complexity in M/Ph/c/N queues, Performance Evaluation 78 (2014) 42–54.

[9] A. Bobbio, A. Horváth, M. Telek, Matching three moments with minimal acyclic phase type distributions, Stochastic models 21 (2-3) (2005) 303–326.

[10] T. Osogami, M. Harchol-Balter, Closed form solutions for mapping general distributions to quasi-minimal PH distributions, Performance Evaluation 63 (6) (2006) 524–552.

[11] A. Brandwajn, T. Begin, Breaking the dimensionality curse in multi-server queues, Computers & Operations Research 73 (2016) 141–149.

[12] S. Stidham Jr, A last word on L= λ W, Operations Research 22 (2) (1974) 417–421.

[13] M. H. MacDougall, Simulating computer systems: techniques and tools, MIT press, 1987.

[14] A. O. Allen, Probability, Statistics and Queueing Theory with Computer Science Applications, Second Edition, Elsevier, 1990.

[15] R. Puigjaner, D. Potier, Modeling techniques and tools for computer performance evaluation, Springer Science & Business Media, 2012.

[16] V. Gupta, M. Harchol-Balter, J. Dai, B. Zwart, On the inapproximability of M/G/K: why two moments of job size distribution are not enough, Queueing Systems 64 (1) (2010) 5–48.

[17] T. Atmaca, T. Begin, A. Brandwajn, H. Castel-Taleb, Performance evaluation of cloud computing centers with general arrivals and service, IEEE Transactions on Parallel and Distributed Systems 27 (8) (2016) 2341–2348.