The Unobserved Waiting Customer Approximation

Kevin Granville *†‡ Steve Drekic *§

Abstract We introduce a new approximation procedure to improve the accuracy of matrix analytic methods when using truncated queueing models to analyze infinite buffer systems. This is accomplished through emulating the presence of unobserved waiting customers beyond the finite buffer that are able to immediately enter the system following an observed customer's departure. We show that this procedure results in exact steady-state probabilities at queue lengths below the buffer for truncated versions of the classic $M/M/1$, $M/M/1+M$, $M/M/\infty$, and $M/PH/1$ queues. We also present two variants of the basic procedure for use within a $M/PH/1+M$ queue and a N-queue polling system with exhaustive service, phase-type service and switch-in times, and exponential impatience timers. The accuracy of these two variants in the context of the polling model are compared through several numerical examples.

Keywords Matrix analytic methods · Polling model · Quasi-birth-and-death process · Reneging · Phase-type distribution · Truncation

Declarations

Funding

Steve Drekic and Kevin Granville acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada through its Discovery Grants program (RGPIN-2016-03685) and Postgraduate Scholarship-Doctoral program, respectively.

Conflict of interest

The authors declare that they have no conflict of interest.

*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada
†Present address: Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON, Canada
‡kgranvil@uwo.ca (✉)
§sdrekic@uwaterloo.ca

1 Introduction

The process in which a system allocates its resources to customers (or jobs) may be modelled by what is referred to as a queueing system. A typical queueing system involves customers waiting in one or more queues to receive service from one or more servers and is thus defined by the assumptions made about customer and server behaviour. Of particular interest is the case of multiple queues of customers waiting to be served by an individual server that periodically visits each of the queues. These systems are specifically referred to as polling models and have received extensive attention in the queueing literature (e.g., Takagi [32], Levy and Sidi [23], Vishnevskii and Semenova [33], Boon [6], and Boon, van der Mei, and Winands [7]). Our analysis method of choice is matrix analytic methods (MAM) (e.g., He [19]), which relies heavily on Markov chain theory. Some examples of recent queueing papers that make use of MAM are Chakravarthy [10], Kim and Kim [20], Avrachenkov, Perel, and Yechiali [3], Perel and Yechiali [26], Sakuma and Takine [30], and Granville and Drekic [16, 17, 18].

An advantage to using MAM when modelling a queueing system or network is that it is possible to analyze models with very complex features. Oftentimes, systems are modelled as a continuous-time Markov chain (CTMC) whose infinitesimal generator matrix is of quasi-birth-and-death (QBD) form, being only able to move within a level or to an adjacent level in a single transition. Should the matrix blocks not depend on the level of the process (after level 0), we refer to this as a level-independent QBD. Otherwise, it is a level-dependent QBD. Resulting from their structure, convenient algorithms have been developed to numerically analyze the systems at steady state (e.g., Neuts [25], Gaver, Jacobs, and Latouche [14], Bright and Taylor [9], Baumann and Sandmann [4, 5], and He [19]).

However, MAM does come with its share of limitations and restrictions. For one, it is generally unable to accept the assumption of arbitrarily distributed interarrival or service times. Instead, one is restricted to Markovian distributions such as the exponential or phase-type (e.g., Neuts [24]). Fortunately, phase-type distributions can be used to approximate non-negative distributions (e.g., Asmussen, Nerman, and Olsson [2]), although higher levels of accuracy come at the cost of increased number of phases. Additionally, a level-independent QBD may permit the state space of only a single dimension (e.g., observable lengths of a single queue) to be infinite, while a level-dependent QBD typically has finite-many states.

Within this work, we introduce the *unobserved waiting customer* approximation which strives to improve the performance of MAM in a situation where it may struggle. Specifically, it aims to reduce the natural biases incurred from the required use of state truncation on a system that should, in reality, have infinite buffers. In what follows, let IB denote the true *infinite buffer* model of interest, let FB denote the *finite buffer* model obtained through simple truncation, and let UWC denote the truncated model making use of the *unobserved waiting customer* approximation.

For simple queues, such as a single-queue system modelled as a level-independent QBD, we may conduct an exact accurate analysis that considers all possible queue lengths. However, to analyze more complicated queueing systems involving multiple queues and/or level-

dependent QBD structures (e.g., due to reneging), we may be required to truncate the state space. If we say, remove all states beyond a threshold representing a queue length of C customers, then it is typical to interpret the removal of these states as the enforcement of a finite buffer which is not present in the real world system we are trying to model. This inaccuracy will result in the steady-state probabilities of the removed states being redistributed proportionally to lower queue length states.

This was observed by Bright and Taylor [9] within their work on how to numerically solve for the steady-state probabilities of a level-dependent QBD. They stated that elementwise, if the CTMC is positive recurrent, the steady-state probability for a state at a given truncation level is greater than or equal to the true value (which we may recover by letting the truncation level go to infinity). They discussed how to select the truncation level to ensure that the steady-state probability of the QBD being in a state at or above this level is negligible. One method is simply to iteratively increase the level until the sum of steady-state probabilities of all states at the truncation level is below a desired tolerance. This is similar to the approach used by Gertsbakh [15] when modelling a 2-queue system where an arriving customer joins the shortest queue. The level of their process was set to be the length of the shorter queue, while the longer queue is truncated to never be n customers longer than the shorter queue. If the difference in queue length reached n, then it was assumed that a customer would immediately jockey to the shorter queue. In their numerical investigation, a value of n was selected such that the steady-state probabilities for all states below the truncation level would change by less than 10^{-6} when further increasing the threshold level by 1.

Alternatively, Bright and Taylor [9] also investigated how to construct a dominating process which can be used to find an analytic upper bound on the tail probability, making use of normal birth-and-death process results. By ensuring that the upper bound of the tail probability is below a threshold, the true tail probability must also be acceptable. An example of applying their methodology is the work of Krishnamoorthy, Babu, and Narayanan [21], in the context of a queue with self-promoting customers (which resulted in a level-dependent QBD). Rather than simply considering the tail probability, Kim and Kim [20] derived an upper bound for the truncation error in their $M/PH/1$ retrial queue with no waiting room, such that an arriving customer who did not find the server free immediately entered an orbit. Truncation error was defined as the sum of absolute-value differences in the steady-state probabilities for all states between their truncated model and the true IB model. They similarly used this upper bound to select a level at which to truncate their customer orbit such that the truncation error was below a specified tolerance.

Unfortunately, it is not always computationally feasible to use a truncation level of C that is large enough to ensure that the tail probability or truncation error is below a small tolerance. For example, in Section 4, we consider a N-queue polling system with phase-type service and switch-in times. If we suppose that each queue has the same truncation level C, along with one-phase switch-in time and two-phase service time distributions, then Table 1 indicates how the number of states required to model the CTMC (which is explicitly characterized later by Equation (4.3)) increase with N and C. While the increase in the number

Table 1: The number of states in a CTMC which models a N-queue polling system having one-phase switch-in time distributions, two-phase service time distributions, and common truncation level C

	C					
N	2	4	6	8	10	12
2	42	130	266	450	682	962
3	189	975	2793	6075	11253	18759
4	756	6500	26068	72900	165044	325156
5	2835	40625	228095	820125	2269355	5283785
6	10206	243750	1915998	8857350	29955486	82427046

of states (and hence the increase in the computational cost to calculate the steady-state distribution) that results from increasing C is not unreasonable for $N = 2$, the computational complexity rapidly increases as N gets larger. While the equal truncation level assumption does not need to hold in practice, if the queueing system has a high level of traffic, then it would not be surprising to require large truncation levels on multiple queues to obtain accurate results. Furthermore, if we used more complex switch-in and service time distributions having more phases, then the number of required states would be even larger.

Therefore, whether for the purposes of analyzing a N-queue polling system or some other system with high traffic and/or a large number of states, it behooves us to consider alternative modifications to a CTMC that give rise to results that outperform a simple FB model. For example, Diamond and Alfa [12] analyzed a retrial queue which tracked both the number of customers in the queue as well as in the retrial orbit. Similar to a 2-queue polling system, it is impossible to let both the queue and orbit have infinite buffers when using MAM. They elected to put a finite buffer on the queue, taking the number of customers in the orbit as the level of their QBD. They modified their CTMC so that after a certain level (chosen so that the tail probability was below a given tolerance), if their queue is not full, then a customer will immediately enter it from the orbit. This approximation is fairly reasonable since as the level increases, the time between retrial attempts will go to zero. This leads to a level-independent QBD structure beyond this point, resulting in more accurate steady-state probabilities than ones obtained via simple truncation. Shin and Choo [31] used a similar approximation in that for part of their analysis of a $M/M/s$ retrial queue with customer balking and reneging, they assumed that the total effective reneging rate of customers in queue did not change beyond a certain level.

Differing from these adjustments, we propose the use of our UWC approximation to improve the overall numerical accuracy when approximating an infinite buffer system when we are unable to use a large enough C. Our ultimate goal is to reduce the negative bias in the expected value of queue lengths at steady state as well as the error in approximated steady-state probabilities that results from state truncation. As we wish to apply this to polling models with potentially very large state spaces, we will do so without requiring the model to track additional states. Also, rather than altering the behaviour of customers in

the system to create a level-independent structure, we will be approximating events that are unobservable by the model. Suppose that in a given queue we truncate at level C, such that we remove all states corresponding to queue lengths greater than C. Rather than assuming the presence of a finite buffer, we assume that customers may be present in positions $C+1, C+2, \ldots$, but are unobservable. If the observed portion of the queue is full, then following an observed customer departure, an unobserved waiting customer may immediately fill the available observed position.

The goal of the UWC approximation is to aggregate probability mass from the tail to the truncation level, resulting in steady-state probabilities at states below the buffer that are either unbiased or less biased than those in a standard FB model. While not designed for level-dependent QBDs, ETAQA (an acronym for an *efficient technique for the analysis of QBD-processes by aggregation*) is an example of an aggregation method for level-independent QBDs. ETAQA was introduced by Ciardo and Smirni [11] for level-independent QBDs satisfying the restriction that all transitions reducing the level of the QBD must transition into the same sublevel, and was extended to $M/G/1$-type CTMCs by Riska and Smirni [28]. Specifically, their ETAQA method calculated steady-state probability vectors $\underline{\pi}_0$, $\underline{\pi}_1$, and $\underline{\pi}^* = \sum_{i=2}^{\infty} \underline{\pi}_i$, where $\underline{\pi}_0$ and $\underline{\pi}_1$ are unbiased and in the latter vector, all sublevels across higher levels are grouped into individual states (i.e., $\pi^*_j = \sum_{i=2}^{\infty} \pi_{i,j}$). Differing from ETAQA, an advantage to UWC is that it may be used to improve accuracy in more general cases where it does not yield exact results. The ability to apply UWC to level-dependent QBDs is also of more use in general than being limited to level-independent QBDs (we note, however, that the goal of ETAQA is not to circumvent truncation limitations, but rather to provide a quicker alternative to solve for quantities such as linear combinations of queue length moments).

The remainder of the paper is organized as follows. In Section 2, we derive optimal applications of UWC to simple single-queue models having exponentially distributed service times (referred to as $M/M/$-type queues). In Section 3, we broaden our consideration to queues having phase-type distributed service times and investigate two choices of approximation that become necessary when assuming both reneging customers and non-exponential service. We proceed to the application of UWC to polling models in Section 4, specifically to the case of a N-queue exhaustive polling system with phase-type service times and potential reneging, and present multiple numerical examples that investigate the effectiveness of UWC. Lastly, we present our concluding remarks in Section 5.

2 $M/M/$-type Queues

2.1 $M/M/1+M$ Queue

We begin by considering the classic $M/M/1+M$ queueing system. In this queue, customer arrivals are governed by a Poisson process with intensity λ and customers are served individually by a single server. Each customer requires an independent and identically distributed (iid) service time following common distribution $Ser \sim \text{Exp}(\mu)$. Additionally, each arriving

customer who is not actively being served is at risk of reneging from the queue due to their own iid impatience timer. Here, the '$+M$' notation (e.g., Boxma and de Waal [8]) indicates that these times are exponentially distributed and we let their common rate be α. Some other examples of papers that consider exponentially distributed impatience timers include Altman and Yechiali [1], Yechiali [34], Shin and Choo [31], Drekic et al. [13], and Granville and Drekic [16].

Let π_i be the steady-state probability of observing i customers in the IB model, $i \in \mathbb{N}$. The balance equations for the IB model of the $M/M/1+M$ queue are

$$\lambda \pi_i = (\mu + i\alpha)\pi_{i+1}, \ i \in \mathbb{N}.$$

Under the normalization condition $1 = \sum_{i=0}^{\infty} \pi_i$, we obtain the solution

$$\pi_i = \frac{\lambda^i \left(\prod_{j=0}^{i-1} (\mu + j\alpha) \right)^{-1}}{1 + \sum_{k=1}^{\infty} \lambda^k \left(\prod_{j=0}^{k-1} (\mu + j\alpha) \right)^{-1}}, \ i \in \mathbb{N}, \tag{2.1}$$

where we use the convention $\prod_{j=0}^{0-1}(\mu + j\alpha) = 1$.

Let us now briefly consider the simple case where we truncate at level C. Letting π_i^{FB} be the steady-state probability of observing i customers in the FB model, $i = 0, 1, \ldots, C$, the modified balance equations for the FB model become

$$\begin{aligned} \lambda \pi_i^{\mathrm{FB}} &= (\mu + i\alpha)\pi_{i+1}^{\mathrm{FB}}, \ i = 0, 1, \ldots, C-2, \\ \lambda \pi_{C-1}^{\mathrm{FB}} &= (\mu + (C-1)\alpha)\pi_C^{\mathrm{FB}}, \end{aligned}$$

which in combination with the normalization condition $1 = \sum_{i=0}^{C} \pi_i^{\mathrm{FB}}$ results in

$$\pi_i^{\mathrm{FB}} = \pi_i \cdot \frac{1 + \sum_{k=1}^{\infty} \lambda^k \left(\prod_{j=0}^{k-1} (\mu + j\alpha) \right)^{-1}}{1 + \sum_{k=1}^{C} \lambda^k \left(\prod_{j=0}^{k-1} (\mu + j\alpha) \right)^{-1}} > \pi_i, \ i = 0, 1, \ldots, C.$$

That is, the truncated steady-state probabilities are simply equal to the re-normalized steady-state probabilities for states $0, 1, \ldots, C$ from the IB model, where all probability mass from states above level C is proportionately redistributed across the lower states. Therefore, the calculated steady-state probabilities have a positive bias while the expected queue length would have a negative bias.

We will now introduce our UWC approximation to adjust the system so that this negative bias will be reduced. In the FB model, the implication of the buffer is that a customer who observes a queue length of C at their arrival instant will be blocked and be lost. Instead, we

suppose that these customers can still wait in the queue, however they are unobserved by the system. As they are not tracked, we must instead approximate their presence. We do so by introducing a probability p_C^* of there being one or more unobserved customers present in the queue at an observed customer's departure epoch when the queue length immediately prior to the departure was C. In this way, with probability p_C^*, there will be a customer present who will immediately fill the vacant observable position within the queue following the departure, and hence the observed queue length does not decrement from our perspective.

As we are not introducing any new states and must preserve the Markov property within our analytical framework, it is a necessity for the distribution of how many observed customer departures are required to decrement the queue length below the buffer to be geometric with success probability $1-p_C^*$. It follows that the amount of time that the UWC model spends in state C has an $\text{Exp}((1-p_C^*)(\mu+(C-1)\alpha))$ distribution. Letting π_i^{UWC} be the steady-state probability of observing i customers in the UWC model, $i=0,1,\ldots,C$, we modify the FB model balance equations to obtain

$$\begin{aligned}\lambda\pi_i^{\text{UWC}} &= (\mu+i\alpha)\pi_{i+1}^{\text{UWC}},\ i=0,1,\ldots,C-2,\\ \lambda\pi_{C-1}^{\text{UWC}} &= (1-p_C^*)(\mu+(C-1)\alpha)\pi_C^{\text{UWC}},\end{aligned}$$

which, when solved along with the normalization condition $1=\sum_{i=0}^{C}\pi_i^{\text{UWC}}$, yields the solution

$$\pi_0^{\text{UWC}} = \left(1+\sum_{k=1}^{C-1}\frac{\lambda^i}{\prod_{j=0}^{i-1}(\mu+j\alpha)}+\frac{1}{1-p_C^*}\cdot\frac{\lambda^C}{\prod_{j=0}^{C-1}(\mu+j\alpha)}\right)^{-1}, \tag{2.2}$$

$$\begin{aligned}\pi_i^{\text{UWC}} &= \frac{\lambda^i}{\prod_{j=0}^{i-1}(\mu+j\alpha)}\pi_0^{\text{UWC}}\\ &= \frac{\frac{\lambda^i}{\prod_{j=0}^{i-1}(\mu+j\alpha)}}{1+\sum_{k=1}^{C-1}\frac{\lambda^i}{\prod_{j=0}^{i-1}(\mu+j\alpha)}+\frac{1}{1-p_C^*}\cdot\frac{\lambda^C}{\prod_{j=0}^{C-1}(\mu+j\alpha)}},\end{aligned} \tag{2.3}$$

for $i=1,2,\ldots,C-1$, and

$$\begin{aligned}\pi_C^{\text{UWC}} &= \frac{1}{1-p_C^*}\cdot\frac{\lambda^C}{\prod_{j=0}^{C-1}(\mu+j\alpha)}\pi_0^{\text{UWC}}\\ &= \frac{\frac{1}{1-p_C^*}\cdot\frac{\lambda^C}{\prod_{j=0}^{C-1}(\mu+j\alpha)}}{1+\sum_{k=1}^{C-1}\frac{\lambda^i}{\prod_{j=0}^{i-1}(\mu+j\alpha)}+\frac{1}{1-p_C^*}\cdot\frac{\lambda^C}{\prod_{j=0}^{C-1}(\mu+j\alpha)}}.\end{aligned} \tag{2.4}$$

We must now determine an appropriate choice for p_C^*. We elect to equate the number of observed customer departures (either from service completions or reneging from the first C queue positions) during a level-C busy period in the UWC and IB models, where we define a *level-C busy period* as the time between the beginnings of a visit to state C and the next

visit to state $C-1$. For the UWC model, this is simply $(1-p_C^*)^{-1}$. The distribution of a level-C busy period in the IB model will be identically distributed as a standard busy period of a $M/M/1+M$ queue with service rate $\mu+(C-1)\alpha$ and individual customer reneging rate α. That is, we can group the reneging rates of all customers at or before the truncation level with the service rate of the leading customer to get an effective overall service rate, since we only care that a departure occurs, regardless of how a customer left the system (or from what observed queue position). Let this effective service time be represented by the random variable $Ser_C \sim \text{Exp}(\mu+(C-1)\alpha)$.

In order to solve for the mean busy period of a $M/M/1+M$ queueing system, we make use of the theory of alternating renewal processes (e.g., Ross [29], Section 7.5.1), which allows us to express the long run proportion of time that the server is busy by

$$1-\pi_0 = \frac{\text{E}[BP]}{\frac{1}{\lambda}+\text{E}[BP]}. \tag{2.5}$$

Rearranging Equation (2.5), we have

$$\text{E}[BP] = \frac{1-\pi_0}{\lambda\pi_0}. \tag{2.6}$$

Letting $i=0$ and replacing μ by $\mu+(C-1)\alpha$ in Equation (2.1), we apply Equation (2.6) to ultimately obtain

$$\text{E}[BP_C] = \sum_{k=1}^{\infty} \lambda^{k-1} \left(\prod_{j=0}^{k-1} (\mu+(C-1+j)\alpha) \right)^{-1},$$

where we let BP_C denote the IB model level-C busy period. Finally, we set

$$\frac{1}{1-p_C^*} = \frac{\text{E}[BP_C]}{\text{E}[Ser_C]} = \sum_{k=1}^{\infty} \frac{\lambda^{k-1}(\mu+(C-1)\alpha)}{\prod_{j=0}^{k-1}(\mu+(C-1+j)\alpha)}, \tag{2.7}$$

implying that

$$p_C^* = 1 - \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}(\mu+(C-1)\alpha)}{\prod_{j=0}^{k-1}(\mu+(C-1+j)\alpha)} \right)^{-1}. \tag{2.8}$$

Note that the formula on the right-hand side of Equation (2.7) can alternatively be obtained through a direct derivation of the expected services in a $M/M/1+M$ queue's busy period (followed by replacing μ by $\mu+(C-1)\alpha$).

Observe that

$$\begin{aligned} \frac{1}{1-p_C^*} \cdot \frac{\lambda^C}{\prod_{j=0}^{C-1}(\mu+j\alpha)} &= \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}(\mu+(C-1)\alpha)}{\prod_{j=0}^{k-1}(\mu+(C-1+j)\alpha)} \right) \frac{\lambda^C}{\prod_{j=0}^{C-1}(\mu+j\alpha)} \\ &= \sum_{i=C}^{\infty} \frac{\lambda^i}{\prod_{j=0}^{i-1}(\mu+j\alpha)}. \end{aligned} \tag{2.9}$$

If we substitute Equation (2.9) into Equations (2.2)-(2.4), we can recover $\pi_i^{\text{UWC}} = \pi_i$, $i = 0, 1, \ldots, C-1$, and

$$\pi_C^{\text{UWC}} = \sum_{i=C}^{\infty} \frac{\lambda^i \left(\prod_{j=0}^{i-1} (\mu + j\alpha) \right)^{-1}}{1 + \sum_{k=1}^{\infty} \lambda^k \left(\prod_{j=0}^{k-1} (\mu + j\alpha) \right)^{-1}} = \sum_{i=C}^{\infty} \pi_i.$$

Therefore, the UWC model can accurately calculate the steady-state probabilities below the truncation level with no bias, while collecting all excess probability mass into π_C^{UWC}. It immediately follows that the expected queue length of the UWC model will have less negative bias than that of the FB model.

2.2 $M/M/1$ Queue

By setting $\alpha = 0$ in our results from Section 2.1, we obtain the corresponding results for the classic $M/M/1$ queue with patient customers. Assuming that the stability condition $\rho = \lambda/\mu < 1$ holds true, we have

$$\pi_i = \rho^i (1 - \rho), \ i \in \mathbb{N},$$
$$\pi_i^{\text{FB}} = \frac{\pi_i}{1 - \rho^{C+1}}, \ i = 0, 1, \ldots, C,$$

and Equation (2.8) simplifies to give

$$p_C^* = 1 - \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{\mu^{k-1}} \right)^{-1} = 1 - \left(\frac{1}{1-\rho} \right)^{-1} = \rho,$$

which is independent of the truncation level C and must still result in $\pi_i^{\text{UWC}} = \pi_i$, $i = 0, 1, \ldots, C-1$, and $\pi_C^{\text{UWC}} = \sum_{i=C}^{\infty} \pi_i$.

2.3 $M/M/\infty$ Queue

By setting $\alpha = \mu$ in our results from Section 2.1, we immediately recover the analysis for a $M/M/\infty$ queue where every customer immediately begins an iid $\text{Exp}(\mu)$ service time upon entering the system. In summary,

$$\pi_i = \frac{\rho^i}{i!} e^{-\rho}, \ i \in \mathbb{N},$$
$$\pi_i^{\text{FB}} = \pi_i \cdot \frac{e^{\rho}}{\sum_{k=0}^{C} \rho^k / k!} > \pi_i, \ i = 0, 1, \ldots, C,$$
$$p_C^* = 1 - \frac{\rho^C}{C!} \left(e^{\rho} - \sum_{k=0}^{C-1} \frac{\rho^k}{k!} \right)^{-1}, \tag{2.10}$$

and it remains that $\pi_i^{\text{UWC}} = \pi_i$, $i = 0, 1, \ldots, C-1$, and $\pi_C^{\text{UWC}} = \sum_{i=C}^{\infty} \pi_i$.

3 $M/PH/1$-type Queues

3.1 $M/PH/1$ Queue

We now consider an analogous model to the $M/M/1$ queue, however we generalize the customer service time distribution from $Ser \sim \text{Exp}(\mu)$ to $Ser \sim \text{PH}_b(\underline{\beta}, B)$. That is, service times are iid continuous phase-type random variables of order b, and we assume that $\underline{\beta}\underline{e}' = \sum_{i=1}^{b} \beta_i = 1$, indicating that service times must be strictly positive in duration. Here, $\underline{e}'$ denotes a column vector of ones having an appropriate length. We assume that $\lambda \text{E}[Ser] < 1$ to guarantee stability in the model.

The $M/PH/1$ queue is modelled using a CTMC denoted by $\{(X(t), Y(t)), t \geq 0\}$, where $X(t)$ is the number of customers in the system and $Y(t)$ is the current service phase at time t, which has possible values depending on $X(t)$:

$$Y(t) \in \Omega_Y(X(t)) = \begin{cases} \{0\} & , \text{ if } X(t) = 0, \\ \{1, 2, \ldots, b\} & , \text{ if } X(t) \in \mathbb{Z}^+. \end{cases}$$

Letting $X(t)$ denote the level of the process and allowing $Q_{i,j}$ to contain the rates corresponding to transitions where $X(t)$ would change from i to j, the infinitesimal generator matrix for this queue takes on the following QBD form:

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ C-1 \\ C \\ C+1 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & \cdots & C-1 & C & C+1 & \cdots \end{array} \\ \begin{bmatrix} Q_{0,0} & Q_{0,1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & Q_{2,1} & Q_{2,2} & \ddots & \ddots & \mathbf{0} & \mathbf{0} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \cdots \\ \mathbf{0} & \mathbf{0} & \ddots & \ddots & Q_{C-1,C-1} & Q_{C-1,C} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & Q_{C,C-1} & Q_{C,C} & Q_{C,C+1} & \ddots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & Q_{C+1,C} & Q_{C+1,C+1} & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix} \end{array}. \tag{3.1}$$

Defining I as an appropriately dimensioned identity matrix and $\underline{B}_0' = -B\underline{e}'$ as the column vector of absorption rates for rate matrix B, the generator blocks for the $M/PH/1$ queue are

$$\begin{array}{rclrclrcl} & & & Q_{0,0} & = & -\lambda, & Q_{0,1} & = & \lambda\underline{\beta}, \\ Q_{1,0} & = & \underline{B}_0', & Q_{1,1} & = & B - \lambda I, & Q_{1,2} & = & \lambda I, \\ Q_{i,i-1} & = & \underline{B}_0'\underline{\beta}, & Q_{i,i} & = & B - \lambda I, & Q_{i,i+1} & = & \lambda I, \ i = 2, 3, \ldots. \end{array} \tag{3.2}$$

As $Q_{i,j}$, $j = i-1, i, i+1$, do not change with i, $i \geq 2$, this is a level-independent QBD. Letting $\pi_{i,j}$ be the steady-state probability of observing the CTMC in state (i, j) and partitioning the steady-state distribution as $\underline{\pi} = (\underline{\pi}_0, \underline{\pi}_1, \underline{\pi}_2, \ldots)$, where $\underline{\pi}_0 = \pi_{0,0}$ and $\underline{\pi}_i = (\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,b})$,

$i \in \mathbb{Z}^+$, we obtain from Equations (3.1) and (3.2)

$$0 = \pi_{0,0}(-\lambda) + \underline{\pi}_1 \underline{B}'_0, \tag{3.3}$$

$$\underline{0} = \pi_{0,0}(\lambda \underline{\beta}) + \underline{\pi}_1 (B - \lambda I) + \underline{\pi}_2 \underline{B}'_0 \underline{\beta}, \tag{3.4}$$

$$\underline{0} = \underline{\pi}_i (\lambda I) + \underline{\pi}_{i+1} (B - \lambda I) + \underline{\pi}_{i+2} \underline{B}'_0 \underline{\beta}, \ i \in \mathbb{Z}^+. \tag{3.5}$$

We now review how to solve for $\underline{\pi}$ (e.g., He [19], Section 4.3) as a point of reference when solving for $\underline{\pi}^{\mathrm{UWC}}$ and selecting p_C^* in the equivalent UWC model. From Equation (3.3), it immediately follows that

$$\lambda \pi_{0,0} = \underline{\pi}_1 \underline{B}'_0. \tag{3.6}$$

After post-multiplying Equations (3.4) and (3.5) by $\underline{e}'$ and performing some elementary substitutions, we can similarly obtain

$$\lambda \underline{\pi}_i \underline{e}' = \underline{\pi}_{i+1} \underline{B}'_0, \ i \in \mathbb{Z}^+. \tag{3.7}$$

Substituting Equation (3.7) for $i = 1$ into Equation (3.4) and solving for $\underline{\pi}_1$, we find

$$\underline{\pi}_1 = \pi_{0,0} \underline{\beta} \lambda (\lambda I - \lambda \underline{e}' \underline{\beta} - B)^{-1}, \tag{3.8}$$

and from Equations (3.5), (3.7), and (3.8),

$$\underline{\pi}_i = \underline{\pi}_{i-1} \lambda (\lambda I - \lambda \underline{e}' \underline{\beta} - B)^{-1} = \pi_{0,0} \underline{\beta} \lambda^i (\lambda I - \lambda \underline{e}' \underline{\beta} - B)^{-i}, \ i \in \mathbb{Z}^+. \tag{3.9}$$

Letting $R = \lambda (\lambda I - \lambda \underline{e}' \underline{\beta} - B)^{-1}$, we obtain the matrix geometric solution $\underline{\pi}_i = \underline{\pi}_1 R^{i-1}$, $i \in \mathbb{Z}^+$. Finally, the normalization condition is

$$1 = \pi_{0,0} + \sum_{i=1}^{\infty} \underline{\pi}_i \underline{e}' = \pi_{0,0} + \underline{\pi}_1 (I - R)^{-1} \underline{e}' = \pi_{0,0} \left(1 + \underline{\beta} R (I - R)^{-1} \underline{e}'\right). \tag{3.10}$$

We can confirm that

$$R(I - R)^{-1} \underline{e}' = \lambda (1 - \lambda \mathrm{E}[Ser])^{-1} (-B^{-1} \underline{e}'), \tag{3.11}$$

where $\mathrm{E}[Ser] = -\underline{\beta} B^{-1} \underline{e}'$. Substituting Equation (3.11) into Equation (3.10) and solving for $\pi_{0,0}$, we obtain

$$\pi_{0,0} = \left(1 + \frac{\lambda \mathrm{E}[Ser]}{1 - \lambda \mathrm{E}[Ser]}\right)^{-1} = 1 - \lambda \mathrm{E}[Ser].$$

Therefore, by Equation (3.9), the remaining steady-state probabilities for the $M/PH/1$ IB model are

$$\underline{\pi}_i = (1 - \lambda \mathrm{E}[Ser]) \underline{\beta} \lambda^i (\lambda I - \lambda \underline{e}' \underline{\beta} - B)^{-i}, \ i \in \mathbb{Z}^+. \tag{3.12}$$

We now consider the UWC model, and confirm that we can recover unbiased steady-state probabilities for levels $0, 1, \ldots, C-1$. Define the partitioned row vector of steady-state probabilities for the truncated CTMC applying the UWC approximation by

$$\underline{\pi}^{\mathrm{UWC}} = (\underline{\pi}_0^{\mathrm{UWC}}, \underline{\pi}_1^{\mathrm{UWC}}, \ldots, \underline{\pi}_C^{\mathrm{UWC}}),$$

where $\underline{\pi}_0^{\mathrm{UWC}} = \pi_{0,0}^{\mathrm{UWC}}$. The generator blocks $Q_{i,j}^{\mathrm{UWC}}$ are only adjusted for $i \geq C$, such that the blocks which do not contain only zeroes are:

$$\begin{array}{rclrclrcl} & & & Q_{0,0}^{\mathrm{UWC}} & = & -\lambda, & Q_{0,1}^{\mathrm{UWC}} & = & \lambda\underline{\beta}, \\ Q_{1,0}^{\mathrm{UWC}} & = & \underline{B}_0', & Q_{1,1}^{\mathrm{UWC}} & = & B - \lambda I, & Q_{1,2}^{\mathrm{UWC}} & = & \lambda I, \\ Q_{i,i-1}^{\mathrm{UWC}} & = & \underline{B}_0'\underline{\beta}, & Q_{i,i}^{\mathrm{UWC}} & = & B - \lambda I, & Q_{i,i+1}^{\mathrm{UWC}} & = & \lambda I,\ i = 2, 3, \ldots, C-1, \\ Q_{C,C-1}^{\mathrm{UWC}} & = & (1 - p_C^*)\underline{B}_0'\underline{\beta}, & Q_{C,C}^{\mathrm{UWC}} & = & B + p_C^*\underline{B}_0'\underline{\beta}, & & & \end{array} \tag{3.13}$$

Here, we no longer observe arrivals at level C, and with probability p_C^*, there is at least one unobserved customer ready to enter the observed states at the time of a service completion, so we have $Q_{C,C}^{\mathrm{UWC}} = Q_{C,C} + \lambda I + p_C^* Q_{C,C-1}$, while we also set $Q_{C,C-1}^{\mathrm{UWC}} = (1 - p_C^*)Q_{C,C-1}$ and $Q_{C,C+1} = \mathbf{0}$.

From Equations (3.1) and (3.13), we have

$$\begin{aligned} 0 &= \pi_{0,0}^{\mathrm{UWC}}(-\lambda) + \underline{\pi}_1^{\mathrm{UWC}}\underline{B}_0', \\ \underline{0} &= \pi_{0,0}^{\mathrm{UWC}}(\lambda\underline{\beta}) + \underline{\pi}_1^{\mathrm{UWC}}(B - \lambda I) + \underline{\pi}_2^{\mathrm{UWC}}\underline{B}_0'\underline{\beta}, \\ \underline{0} &= \underline{\pi}_i^{\mathrm{UWC}}(\lambda I) + \underline{\pi}_{i+1}^{\mathrm{UWC}}(B - \lambda I) + \underline{\pi}_{i+2}^{\mathrm{UWC}}\underline{B}_0'\underline{\beta},\ i = 1, 2, \ldots, C-3, \\ \underline{0} &= \underline{\pi}_{C-2}^{\mathrm{UWC}}(\lambda I) + \underline{\pi}_{C-1}^{\mathrm{UWC}}(B - \lambda I) + \underline{\pi}_C^{\mathrm{UWC}}(1 - p_C^*)\underline{B}_0'\underline{\beta}, \\ \underline{0} &= \underline{\pi}_{C-1}^{\mathrm{UWC}}(\lambda I) + \underline{\pi}_C^{\mathrm{UWC}}(B + p_C^*\underline{B}_0'\underline{\beta}), \end{aligned} \tag{3.14}$$

from which we can obtain

$$\lambda\pi_{0,0}^{\mathrm{UWC}} = \underline{\pi}_1^{\mathrm{UWC}}\underline{B}_0', \tag{3.15}$$

$$\lambda\underline{\pi}_i^{\mathrm{UWC}}\underline{e}' = \underline{\pi}_{i+1}^{\mathrm{UWC}}\underline{B}_0',\ i = 1, 2, \ldots, C-2, \tag{3.16}$$

$$\lambda\underline{\pi}_{C-1}^{\mathrm{UWC}}\underline{e}' = \underline{\pi}_C^{\mathrm{UWC}}(1 - p_C^*)\underline{B}_0'.$$

Since Equations (3.15) and (3.16) have the same form as Equations (3.6) and (3.7), we similarly find that

$$\underline{\pi}_i^{\mathrm{UWC}} = \pi_{0,0}^{\mathrm{UWC}}\underline{\beta}\lambda^i(\lambda I - \lambda\underline{e}'\underline{\beta} - B)^{-i},\ i = 1, 2, \ldots, C-1. \tag{3.17}$$

However, rearranging Equation (3.14) for $\underline{\pi}_C^{\mathrm{UWC}}$ and substituting Equation (3.17) for $i = C-1$, we have

$$\begin{aligned} \underline{\pi}_C^{\mathrm{UWC}} &= \underline{\pi}_{C-1}^{\mathrm{UWC}}\lambda(-(B + p_C^*\underline{B}_0'\underline{\beta})^{-1}) \\ &= \pi_{0,0}^{\mathrm{UWC}}\underline{\beta}\lambda^C(\lambda I - \lambda\underline{e}'\underline{\beta} - B)^{-(C-1)}(-(B + p_C^*\underline{B}_0'\underline{\beta})^{-1}). \end{aligned}$$

The normalization condition for the UWC model becomes

$$\begin{aligned} 1 &= \pi_{0,0}^{\mathrm{UWC}} + \sum_{i=1}^{C} \underline{\pi}_i^{\mathrm{UWC}}\underline{e}' \\ &= \pi_{0,0}^{\mathrm{UWC}}\left(1 + \sum_{i=1}^{C-1} \underline{\beta}R^i\underline{e}' + \underline{\beta}R^{C-1}\lambda(-(B + p_C^*\underline{B}_0'\underline{\beta})^{-1}\underline{e}')\right). \end{aligned} \tag{3.18}$$

Note that we can alternately express Equation (3.10) as

$$1 = \pi_{0,0}\left(1 + \sum_{i=1}^{C-1} \underline{\beta} R^i \underline{e}' + \sum_{i=C}^{\infty} \underline{\beta} R^i \underline{e}'\right) = \pi_{0,0}\left(1 + \sum_{i=1}^{C-1} \underline{\beta} R^i \underline{e}' + \underline{\beta} R^C (I-R)^{-1} \underline{e}'\right),$$

so if p_C^* satisfies

$$\underline{\beta} R^{C-1} \lambda(-(B + p_C^* \underline{B}_0' \underline{\beta})^{-1})\underline{e}' = \underline{\beta} R^C (I-R)^{-1} \underline{e}', \tag{3.19}$$

then $\pi_{0,0}^{\text{UWC}} = \pi_{0,0}$, and by Equation (3.17), $\underline{\pi}_i^{\text{UWC}} = \underline{\pi}_i$, $i = 1, 2, \ldots, C-1$.

We must now select a value of p_C^*. As before, we aim to equate the expected number of observed customer departures during level-C busy periods in the UWC and IB models. Note, however, that unlike the exponential service case, we must now consider the service phase that is underway at the beginning of a level-C busy period, BP_C. We define $q_{x,y}$ as the steady-state probability of the IB model being in state (x, y) immediately prior to a customer arrival that initiates a level-C busy period (i.e., an arrival that increases $X(t)$ from $C-1$ to C). It follows that at steady-state (e.g., Lakatos, Szeidl, and Telek [22], Chapter 9),

$$\begin{aligned}
q_{C-1,y} &= \lim_{h\to 0} P((X(t), Y(t)) = (C-1, y) | X(t+h) = C) \\
&= \lim_{h\to 0} \frac{P(X(t+h) = C | (X(t), Y(t)) = (C-1, y)) P((X(t), Y(t)) = (C-1, y))}{\sum_{m,n} P(X(t+h) = C | (X(t), Y(t)) = (m, n)) P((X(t), Y(t)) = (m, n))} \\
&= \lim_{h\to 0} \frac{(\lambda h + o(h))\pi_{C-1,y}}{\sum_n (\lambda h + o(h))\pi_{C-1,n}} \\
&= \lim_{h\to 0} \frac{\lambda \pi_{C-1,y} + o(h)/h}{\sum_n \lambda \pi_{C-1,n} + o(h)/h} \\
&= \frac{\pi_{C-1,y}}{\underline{\pi}_{C-1}\underline{e}'}.
\end{aligned} \tag{3.20}$$

Applying Equations (3.12) and (3.20), we define the modified initial probability row vector

$$\underline{\beta}_C^* = (q_{C-1,1}, q_{C-1,2}, \ldots, q_{C-1,b}) = \frac{\underline{\pi}_{C-1}}{\underline{\pi}_{C-1}\underline{e}'} = \frac{\underline{\beta}(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-(C-1)}}{\underline{\beta}(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-(C-1)}\underline{e}'}. \tag{3.21}$$

It now follows that BP_C will be identical in distribution to a busy period of a modified IB $M/PH/1$ queue where the first customer of a busy period has a service time with distribution $Ser_C^* \sim \text{PH}_b(\underline{\beta}_C^*, B)$, but all future service times within the same busy period will be iid following the original $\text{PH}_b(\underline{\beta}, B)$ distribution. We can calculate $\text{E}[BP_C]$ by setting $Q_{0,1} = \lambda \underline{\beta}_C^*$ in Equation (3.2) and solving for the modified steady-state distribution which we will define as $\underline{\pi}^{*C} = (\pi_{0,0}^{*C}, \underline{\pi}_1^{*C}, \underline{\pi}_2^{*C}, \ldots)$. By Equation (2.6), it readily follows that

$$\text{E}[BP_C] = \frac{1 - \pi_{0,0}^{*C}}{\lambda \pi_{0,0}^{*C}}. \tag{3.22}$$

Following similar steps to the original analysis for the IB model, we can show that

$$\underline{\pi}_i^{*C} = \pi_{0,0}^{*C}\underline{\beta}_C^*\lambda^i(\lambda I - \lambda\underline{e}'\underline{\beta} - B)^{-i},\ i \in \mathbb{Z}^+.$$

Using Equation (3.11), we can now solve for $\pi_{0,0}^{*C}$ through the normalization condition,

$$\begin{aligned}
1 &= \pi_{0,0}^{*C} + \underline{\pi}_1^{*C}(I-R)^{-1}\underline{e}' \\
&= \pi_{0,0}^{*C}\left(1 + \underline{\beta}_C^* R(I-R)^{-1}\underline{e}'\right) \\
&= \pi_{0,0}^{*C}\left(1 + \frac{\lambda(-\underline{\beta}_C^* B^{-1}\underline{e}')}{1-\lambda\mathrm{E}[Ser]}\right) \\
&= \pi_{0,0}^{*C}\left(1 + \frac{\lambda\mathrm{E}[Ser_C^*]}{1-\lambda\mathrm{E}[Ser]}\right),
\end{aligned}$$

resulting in

$$\pi_{0,0}^{*C} = \frac{1-\lambda\mathrm{E}[Ser]}{1+\lambda\mathrm{E}[Ser_C^*]-\lambda\mathrm{E}[Ser]},$$

and by substituting into Equation (3.22), it is straightforward to show that

$$\mathrm{E}[BP_C^*] = \frac{\mathrm{E}[Ser_C^*]}{1-\lambda\mathrm{E}[Ser]}. \tag{3.23}$$

We now recall the left-hand side of Equation (3.19), which we can rewrite via Equation (3.21) as

$$\begin{aligned}
&\underline{\beta}R^{C-1}\lambda(-(B+p_C^*\underline{B}_0'\underline{\beta})^{-1})\underline{e}' \\
&= \lambda^C(\underline{\beta}(\lambda I - \lambda\underline{e}'\underline{\beta} - B)^{-(C-1)}\underline{e}')(-\underline{\beta}_C^*(B+p_C^*\underline{B}_0'\underline{\beta})^{-1}\underline{e}').
\end{aligned} \tag{3.24}$$

Note that the term $-\underline{\beta}_C^*(B+p_C^*\underline{B}_0'\underline{\beta})^{-1}\underline{e}'$ is simply the expected value of a $\mathrm{PH}_b(\underline{\beta}_C^*, B + p_C^*\underline{B}_0'\underline{\beta})$ random variable. This corresponds to a phase-type distribution with initial probability row vector $\underline{\beta}_C^*$ and rate matrix B that restarts with initial probability row vector $\underline{\beta}$ every time it would reach absorption with probability p_C^*. We can express

$$BP_C = Ser_C^* + \sum_{j=1}^{N_C^*} Ser_j, \tag{3.25}$$

where $\{Ser_j\}_{j=1}^{\infty}$ are iid service times having distribution $Ser_j \sim \mathrm{PH}_b(\underline{\beta}, B)$ within a busy period after the first service $Ser_C^* \sim \mathrm{PH}_b(\underline{\beta}_C^*, B)$, and N_C^* is some discrete random variable depending on λ, C, and the random service times. If we approximate N_C^* by an independent geometric distribution having probability mass function (PMF) $P(N=n) = (p_C^*)^n(1-p_C^*)$, $n \in \mathbb{N}$, then this would be distributionally equivalent to $\mathrm{PH}_b(\underline{\beta}_C^*, B + p_C^*\underline{B}_0'\underline{\beta})$ (using the convention $\sum_{j=1}^{0} Ser_j = 0$).

Taking the expectation of Equation (3.25) under this approximation, we have

$$\mathrm{E}[BP_C] = \mathrm{E}[Ser_C^*] + \frac{p_C^*}{1-p_C^*}\mathrm{E}[Ser]. \tag{3.26}$$

Equating Equations (3.23) and (3.26) and solving for p_C^*, we set

$$p_C^* = \frac{\lambda \mathrm{E}[Ser_C^*]}{1+\lambda \mathrm{E}[Ser_C^*] - \lambda \mathrm{E}[Ser]}. \tag{3.27}$$

Therefore, if we use this choice of p_C^*, Equation (3.24) becomes

$$\begin{aligned}
\underline{\beta} R^{C-1}\lambda(-(B+p_C^*\underline{B}_0'\underline{\beta})^{-1})\underline{e}' &= \lambda^C(\underline{\beta}(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-(C-1)}\underline{e}')\mathrm{E}[BP_C] \\
&= \frac{\lambda^C}{1-\lambda \mathrm{E}[Ser]}(\underline{\beta}(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-(C-1)}\underline{e}')\mathrm{E}[Ser_C^*] \\
&= \frac{\lambda^C}{1-\lambda \mathrm{E}[Ser]}(\underline{\beta}(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-(C-1)}\underline{e}')(-\underline{\beta}_C^* B^{-1}\underline{e}') \\
&= \frac{\lambda^C}{1-\lambda \mathrm{E}[Ser]}\underline{\beta}(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-(C-1)}(-B^{-1}\underline{e}').
\end{aligned}$$

Substituting Equation (3.11) into the right-hand side of Equation (3.19) yields

$$\underline{\beta} R^C (I-R)^{-1}\underline{e}' = \frac{\lambda^C}{1-\lambda \mathrm{E}[Ser]}\underline{\beta}(\lambda I - \lambda \underline{e}'\underline{\beta}\underline{e}' - B)^{-(C-1)}(-B^{-1}\underline{e}'). \tag{3.28}$$

Thus, we have shown that the choice of p_C^* in Equation (3.27) satisfies Equation (3.11), and it will hold that

$$\pi_{0,0}^{\mathrm{UWC}} = 1 - \lambda \mathrm{E}[Ser] = \pi_{0,0},$$

$$\underline{\pi}_i^{\mathrm{UWC}} = (1-\lambda \mathrm{E}[Ser])\underline{\beta}\lambda^i(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-i} = \underline{\pi}_i,\ i = 1, 2, \ldots, C-1,$$

and

$$\underline{\pi}_C^{\mathrm{UWC}} = (1-\lambda \mathrm{E}[Ser])\underline{\beta}\lambda^C(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-(C-1)}(-(B+p_C^*\underline{B}_0'\underline{\beta})^{-1}), \tag{3.29}$$

which must satisfy $\underline{\pi}_C^{\mathrm{UWC}}\underline{e}' = \sum_{i=C}^{\infty}\underline{\pi}_i\underline{e}'$.

Remark 1 We can obtain the corresponding FB model results for the $M/PH/1$ queue by setting $p_C^* = 0$ in the above analysis. From Equations (3.19) and (3.28), since

$$\begin{aligned}
\underline{\beta} R^{C-1}\lambda(-B^{-1}\underline{e}') &= \lambda^C\underline{\beta}(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-(C-1)}(-B^{-1}\underline{e}') \\
&= (1-\lambda \mathrm{E}[Ser])\underline{\beta} R^C (I-R)^{-1}\underline{e}' \\
&< \underline{\beta} R^C (I-R)^{-1}\underline{e}',
\end{aligned}$$

it follows that $\pi_{0,0}^{\mathrm{FB}} > \pi_{0,0}$, and hence by Equation (3.17),

$$\underline{\pi}_i^{\mathrm{FB}} = \pi_{0,0}^{\mathrm{FB}}\underline{\beta}\lambda^i(\lambda I - \lambda \underline{e}'\underline{\beta} - B)^{-i} = \frac{\pi_{0,0}^{\mathrm{FB}}}{\pi_{0,0}} \cdot \underline{\pi}_i > \underline{\pi}_i,\ i = 1, 2, \ldots, C-1. \tag{3.30}$$

Interestingly, since Equation (3.30) confirms that $\underline{\pi}^{\text{FB}}_{C-1}$ is proportional to $\underline{\pi}_{C-1}$, this implies that we can express Equation (3.21) in terms of the FB steady-state probabilities to obtain

$$\underline{\beta}^*_C = \frac{\underline{\pi}_{C-1}}{\underline{\pi}_{C-1}\underline{e}'} = \frac{\underline{\pi}^{\text{FB}}_{C-1}}{\underline{\pi}^{\text{FB}}_{C-1}\underline{e}'}.$$

This observation will inspire an approximation that we will examine in Section 3.2.2.

3.1.1 $M/M/1$ Queue

Let us confirm that we recover the results of Section 2.2 if we let $\underline{\beta} = 1$ and $B = -\mu$ (i.e., if $Ser \sim \text{Exp}(\mu)$). First of all, Equation (3.21) clearly simplifies to $\underline{\beta}^*_C = 1 = \underline{\beta}$. Therefore, Ser^*_C and Ser have identical distributions, implying that $\text{E}[Ser^*_C] = \text{E}[Ser]$ and Equation (3.27) simplifies to $p^*_C = \lambda\text{E}[Ser] = \lambda/\mu = \rho$. Thus, Equation (3.29) reduces to

$$\begin{aligned}\pi^{\text{UWC}}_C &= (1-\rho)\lambda^C(\lambda - \lambda - (-\mu))^{-(C-1)}(-(-\mu + \rho(\mu))^{-1}) \\ &= (1-\rho)\rho^{C-1}\frac{\lambda}{\mu - \lambda} = \rho^C,\end{aligned}$$

as required. Therefore, our two different UWC methods derived from the $M/M/1 + M$ and $M/PH/1$ queues both simplify to give the same desired results for the $M/M/1$ queue.

3.2 $M/PH/1 + M$ Queue

Suppose now that individual customers not currently receiving service in the $M/PH/1$ queue from Section 3.1 are at risk of reneging according to iid $\text{Exp}(\alpha)$ impatience timers (as in Section 2.1). This too may be modelled by a CTMC $\{(X(t), Y(t)), t \geq 0\}$ with the same interpretations as previously described. This CTMC is still a QBD whose infinitesimal generator matrix takes the form of Equation (3.1), with non-zero matrix blocks:

$$\begin{array}{rclcrclcrcl} & & & & Q_{0,0} &=& -\lambda, & & Q_{0,1} &=& \lambda\underline{\beta}, \\ Q_{1,0} &=& \underline{B}'_0, & & Q_{1,1} &=& B - \lambda I, & & Q_{1,2} &=& \lambda\overline{I}, \\ Q_{i,i-1} &=& \underline{B}'_0\underline{\beta} + (i-1)\alpha I, & & Q_{i,i} &=& B - (\lambda + (i-1)\alpha)I, & & Q_{i,i+1} &=& \lambda I,\ i = 2, 3, \ldots. \end{array}$$

As $Q_{i,j}$ now depends on i for $j = i-1, i$, this is a level-dependent QBD. Information concerning the analytical approach required to calculate the steady-state distribution of a level-dependent QBD may be found in, for example, Bright and Taylor [9]. In our notation, this algorithm assumes that

$$\underline{\pi}_i = \underline{\pi}_0 \prod_{j=1}^{i} R_j,\ i \in \mathbb{Z}^+, \tag{3.31}$$

which in combination with the QBD form of Equation (3.1) results in the recursive solution

$$R_i = -Q_{i-1,i}\left(Q_{i,i} + R_{i+1}Q_{i+1,i}\right)^{-1},\ i \in \mathbb{Z}^+. \tag{3.32}$$

As each R_i references the value of R_{i+1}, in order to actually calculate the steady-state probabilities, we must implement a state truncation at some level C by setting $R_i = \mathbf{0}$ for all $j > C$. This provides us with the boundary equation

$$R_C = -Q_{C-1,C}(Q_{C,C})^{-1}. \tag{3.33}$$

After calculating R_C, we iteratively obtain R_i, $i = C-1, C-2, \ldots, 1$. Defining

$$R_0 = Q_{0,0} + R_1 Q_{1,0}, \tag{3.34}$$

it must hold that

$$\underline{\pi}_0 R_0 = \underline{0}. \tag{3.35}$$

The normalization equation becomes

$$1 = \sum_{i=0}^{C} \underline{\pi}_i \underline{e}' = \underline{\pi}_0 \left(I + \sum_{i=1}^{C} \prod_{j=1}^{i} R_j \right) \underline{e}' = \underline{\pi}_0 \underline{u}', \tag{3.36}$$

which in combination with Equation (3.35) provides us with the means to calculate $\underline{\pi}_0$,

$$\underline{\pi}_0 \left[\begin{array}{cc} R_0 & \underline{u}' \end{array} \right] = \left[\begin{array}{cc} \underline{0} & 1 \end{array} \right].$$

The remaining $\underline{\pi}_i$'s may then be obtained through Equation (3.31). As this numerical algorithm requires a truncation of the state space at a given level C, it calculates the steady-state distribution of the FB model approximation by default (when no adjustments are applied), which will converge to that of the IB model as $C \to \infty$.

For the UWC model (with truncation at level C), we must modify our approach due to the presence of both reneging and phase-type service. In the $M/PH/1$ queue, the only way a customer could depart the system was following the completion of their service, after which the time until the next observed departure would have an iid distribution (i.e., a new service phase is always selected according to the probability vector $\underline{\beta}$). In the $M/PH/1+M$ queue, while we re-initialize the service phase after service completions, the current service phase is unchanged if we observe a departure due to impatience. Therefore, the random time intervals between observed departures are no longer iid and we cannot make a similar breakdown of a level-C busy period as in Equation (3.25). That is, we are unable to obtain the expected number of observed departures from the expected duration of a level-C busy period. In the analysis of the $M/M/1+M$ queue, this was not a concern due to the existence of only a single service phase. We will now propose two versions of the UWC approximation to tackle this more difficult problem.

3.2.1 $M/PH/1+M$ Queue: UWC Version 1

For UWC version 1, we obtain an analytic approximation that is comparable in computational complexity to the FB model approximation. To illustrate, we consider the UWC

model of a $M/PH/1+M$ queue having two service phases. We visualize the state transition diagram of this model for states near level C in Figure 1, where we denote the absorption rate out of the j^{th} phase of *Ser* by $B_{j,0} = (\underline{B}'_0)_j$. While in state $(C,1)$, an observed departure will decrease the queue length with probability $1-p^*_{C,1}$, and the CTMC will remain in state $(C,1)$ with probability

$$\frac{(C-1)\alpha + B_{1,0}\beta_1}{(C-1)\alpha + B_{1,0}} \cdot p^*_{C,1}, \tag{3.37}$$

or the CTMC will transition to state $(C,2)$ with probability

$$\frac{B_{1,0}\beta_2}{(C-1)\alpha + B_{1,0}} \cdot p^*_{C,1}. \tag{3.38}$$

That is, the UWC approximation will respond to departures in the same way in this model as in a $M/M/1+M$ model with service rate $B_{1,0}$, with the exception of service completions that re-initialize the service phase in a *different* phase.

As $C \to \infty$, Equations (3.37) and (3.38) converge to $p^*_{C,1}$ and 0, respectively, as observed customer departures while at level C become dominated by reneging, reducing the probability of an observed departure that would change the service phase but not decrement the queue length. Thus, in terms of the UWC behaviour, we approximate being in state (C,j) as if in state C of a $M/M/1+M$ model with service rate μ equal to $B_{j,0}$. In the $M/M/1+M$ queue, the probability p^*_C in Equation (2.8) was optimal for estimating the probability of requiring at least one more observed departure to lower the level of the CTMC to $C-1$. Therefore, upon observing a departure while in state (C,j), we elect to let

$$p^*_{C,j} = 1 - \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}(B_{j,0} + (C-1)\alpha)}{\prod_{n=0}^{k-1}(B_{j,0} + (C-1+n)\alpha)} \right)^{-1}, \; j = 1,2,\ldots,b, \tag{3.39}$$

be the (approximate) probability of having one or more unobserved customers present in the system.

Returning to our two-phase illustration, in addition to service completions that result in a change of service phase but not a reduction in the number of observed customers, there is an independent competing $\text{Exp}(B_{1,2})$ timer whose completion would result in a transition to state $(C,2)$. When transitioning into state $(C,2)$, the CTMC retains no memory of the previous state being $(C,1)$ or $(C-1,2)$, and in either case uses UWC probability $p^*_{C,2}$ while in this state. Thus, no information concerning how long the system has remained in level C is preserved, and we may interpret this event as the removal of any unobserved waiting customers present in the system at that time instant followed by the initialization of a new level-C busy period in an $M/M/1+M$ queue with service rate $B_{2,0}$. This intuition generalizes logically to any number of service phases, b. Therefore, while $p^*_{C,1}, p^*_{C,2}, \ldots, p^*_{C,b}$ will shift some steady-state probability mass to the truncation level C, they will underestimate the true expected number of required customer departures to transition to level $C-1$ due to the removal of unseen waiting customers when the service phase (but not the number of observed

customers) changes. While the gain in accuracy is not as high as in the simpler models, it will still outperform the standard FB model since only removing unobserved waiting customers upon changes in service phase results in fewer lost customers than rejecting everyone who arrives while the queue length is at C (i.e., when enforcing a finite buffer).

Unlike in Section 3.1, the equations for $\underline{\pi}_i^{\text{UWC}}$ will recursively depend on the form of $Q_{C,C}^{\text{UWC}}$ through Equations (3.32) and (3.33), whereas previously they only depended on the value of $\underline{\pi}_C^{\text{UWC}}\underline{e}'$ through the normalization condition in Equation (3.18) used to obtain $\pi_{0,0}^{\text{UWC}}$. Therefore, as we will see in Tables 2 and 3, the less precise UWC approximation used in $Q_{C,C}^{\text{UWC}}$ and $Q_{C,C-1}^{\text{UWC}}$ may cause slight irregularities in levels near the buffer, where the dependency on R_C is larger. However, these irregularities within a given level vanish as we increase C causing the distance between that level and the truncation level to grow.

Note that if $B_{0,1} = B_{0,2} = \cdots = B_{0,b}$, then $p_{C,1}^* = p_{C,2}^* = \cdots = p_{C,b}^*$ and no accuracy in the UWC approximation is lost. Also, these choices of $p_{C,j}^*$ will reduce to the p_C^* of the $M/M/1+M$ UWC model if we assume exponentially distributed service times, as required. Moreover, as the proportion of departures due to reneging increases with C, fewer instances of "lost" unobserved waiting customers will occur and the accuracy of the UWC approximation itself will improve with larger C. Thus, like the other UWC models, the UWC steady-state distribution will converge to that of the IB model as $C \to \infty$.

3.2.2 $M/PH/1+M$ Queue: UWC Version 2

We now consider a second version of UWC for this particular model. Rather than making use of earlier results derived for a simpler model, we apply phase-type theory to approximate the PMF of N_C^* directly, from which we can obtain its expected value and use it to set a single UWC probability p_C^*. As in the analysis of Section 3.1, the initial service phase of the level-C busy period matters, and like Equation (3.21), we would find that

$$\underline{\beta}_C^* = \frac{\underline{\pi}_{C-1}}{\underline{\pi}_{C-1}\underline{e}'}.$$

Unfortunately, we do not have a precise analytic solution of $\underline{\pi}_{C-1}$ from the IB model. However, as seen in Remark 1, these IB steady-state probabilities may be replaced by those from the FB model with no loss of accuracy for the $M/PH/1$ model. While this is not necessarily the case for the $M/PH/1+M$ model, we still propose to use $\underline{\pi}_{C-1}^{\text{FB}}$ in place of $\underline{\pi}_{C-1}$, and we denote this approximated initial probability row vector by $\underline{\hat{\beta}}_C^*$. As we will see in Tables 2 and 3, while we do not end up obtaining exact steady-state probabilities for levels below C, this approximation works very well. Note that this does imply that we must calculate the steady-state probabilities of the FB model prior to those of this version of the UWC model, effectively doubling our computational requirement. This is the main limitation of UWC version 2, which is otherwise typically more accurate than UWC version 1.

Letting $D \in \mathbb{Z}^+$, we may model the number of customers beyond level C (up until the

next observed departure) by an absorbing CTMC having infinitesimal generator matrix

$$Q = \begin{bmatrix} Q_{TT} & Q_{TA} \\ \mathbf{0} & I \end{bmatrix},$$

where we let $\Delta = B - (\lambda + (C-1)\alpha)I_b$, $\Delta_A = \underline{B}_0'\underline{\beta} + (C-1)\alpha I_b$,

$$Q_{TT} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ D-1 \\ D \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \cdots & D-1 & D \end{array} \\ \begin{bmatrix} \Delta & \lambda I_b & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \alpha I_b & \Delta - \alpha I_b & \lambda I_b & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\alpha I_b & \Delta - 2\alpha I_b & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Delta - (D-1)\alpha I_b & \lambda I_b \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & D\alpha I_b & \Delta - (D\alpha - \lambda)I_b \end{bmatrix} \end{array},$$

and

$$Q_{TA} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ D-1 \\ D \end{array} \begin{array}{c} \begin{array}{cccccc} 0^* & 1^* & \cdots & (D-2)^* & (D-1)^* & -1^* \end{array} \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \Delta_A\underline{e}' \\ \Delta_A & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \underline{0}_b' \\ \mathbf{0} & \Delta_A & \ddots & \mathbf{0} & \mathbf{0} & \underline{0}_b' \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Delta_A & \mathbf{0} & \underline{0}_b' \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \Delta_A & \underline{0}_b' \end{bmatrix} \end{array}.$$

Here, I_b denotes a $b \times b$ identity matrix and $\underline{0}_b'$ is a column vector of b zeroes. This CTMC applies a FB approximation to the unobserved portion of the queue (considering an effective total queue length of $C + D$). If it is absorbed into state (i^*, j), $i^* \in \{0^*, 1^*, \ldots, (D-1)^*\}$, $j \in \{1, 2, \ldots, b\}$, then the queue length does not decrease after the next observed departure. An unobserved customer immediately joins the observed portion of the queue, there are i unobserved customers in the system, and the next service time begins in phase j. If it is absorbed into state -1^*, then there were no unobserved customers and the observed queue length will decrement.

Given the initial probability row vector $\underline{\hat{\beta}}_C^*$, if we let D^* be a set of dummy absorption states (which cannot actually be observed) and define

$$Q_{TA}^* = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ D-1 \\ D \end{array} \begin{array}{c} \begin{array}{cccccc} 0^* & 1^* & \cdots & (D-2)^* & (D-1)^* & D^* \end{array} \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \underline{0}_b'\underline{0}_b \\ \Delta_A & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Delta_A & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Delta_A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \Delta_A & \mathbf{0} \end{bmatrix} \end{array},$$

while we let the right-most column of Q_{TA} be denoted by

$$\underline{Q}'_{-1^*} = \begin{bmatrix} \Delta_A \underline{e}' \\ \underline{0}' \end{bmatrix},$$

then the probability that the queue length will decrement after the first observed departure is simply

$$P(N_C^* = 0) = \begin{bmatrix} \hat{\underline{\beta}}_C^* & \underline{0} \end{bmatrix} (-Q_{TT}^{-1}) \underline{Q}'_{-1^*}.$$

Since we can use the knowledge of the absorption state to re-initialize the process to run until the next observed departure without losing track of the number of unobserved customers, we can actually express the (approximated) PMF of N_C^* as

$$P(N_C^* = n) = \begin{bmatrix} \hat{\underline{\beta}}_C^* & \underline{0} \end{bmatrix} \left[(-Q_{TT}^{-1}) Q_{TA}^*\right]^n (-Q_{TT}^{-1}) \underline{Q}'_{-1^*}, \ n \in \mathbb{N}.$$

From here, we evaluate $\mathrm{E}[N_C^*]$ and select a UWC probability that equates

$$\mathrm{E}[N_C^*] = \frac{p_C^*}{1 - p_C^*},$$

or equivalently,

$$p_{C,j}^* = p_C^* = \frac{\mathrm{E}[N_C^*]}{1 + \mathrm{E}[N_C^*]}, \ j = 1, 2, \ldots, b.$$

Note that this is an approximation for a given choice of $D \in \mathbb{Z}^+$. As such, a large enough D should be selected such that this FB approximation approaches that of the true IB model. For the calculations in Section 3.2.3, we use $D = 40$.

3.2.3 $M/PH/1 + M$ Queue: Comparing UWC Versions

Defining $\underline{p}_C^* = (p_{C,1}^*, p_{C,2}^*, \ldots, p_{C,b}^*)$, we let the non-zero QBD blocks of the UWC model be given by $Q_{i,j}^{\mathrm{UWC}} = Q_{i,j}$, $i = 0, 1, \ldots, C-1$, $j = 0, 1, \ldots, C$,

$$\begin{aligned} Q_{C,C-1}^{\mathrm{UWC}} &= (I - \mathrm{diag}(\underline{p}_C^*)) Q_{C,C-1} \\ &= (I - \mathrm{diag}(\underline{p}_C^*))(\underline{B}_0' \underline{\beta} + (C-1)\alpha I), \end{aligned} \tag{3.40}$$

and

$$\begin{aligned} Q_{C,C}^{\mathrm{UWC}} &= Q_{C,C} + \lambda I + \mathrm{diag}(\underline{p}_C^*) Q_{C,C-1} \\ &= B + \mathrm{diag}(\underline{p}_C^*) \underline{B}_0' \underline{\beta} - (I - \mathrm{diag}(\underline{p}_C^*))(C-1)\alpha I. \end{aligned} \tag{3.41}$$

In Tables 2 and 3, we illustrate the relative efficiency gains of both versions of the UWC model over the FB model. We apply the level-dependent QBD algorithm to approximate the IB model steady-state distribution $\underline{\pi}$ using a truncation level of 1000, as well as to calculate $\underline{\pi}^{\mathrm{UWC}}$ and $\underline{\pi}^{\mathrm{FB}}$ for $C = 3, 7$. We let $\lambda = 0.9$, $\alpha = 0.1$, and consider the following service time distributions:

- (E_2) Erlang-2 with $\mathrm{E}[Ser] = 1$ and $\mathrm{Var}(Ser) = 0.5$:

$$Ser \sim \mathrm{PH}_2 \left(\underline{\beta} = (1, 0), B = \begin{bmatrix} -2 & 2 \\ 0 & -2 \end{bmatrix} \right).$$

- (E_2^f) Erlang-2 with feedback with $\mathrm{E}[Ser] = 1$ and $\mathrm{Var}(Ser) = 0.75$:

$$Ser \sim \mathrm{PH}_2 \left(\underline{\beta} = (1, 0), B = \begin{bmatrix} -4 & 4 \\ 2 & -4 \end{bmatrix} \right).$$

- (C_2^f) Coxian-2 with feedback with $\mathrm{E}[Ser] = 1$, $\mathrm{Var}(Ser) = 1.5$:

$$Ser \sim \mathrm{PH}_2 \left(\underline{\beta} = (0.5, 0.5), B = \begin{bmatrix} -\left(\frac{8+4\sqrt{3}}{7+4\sqrt{3}}\right) & \frac{4+2\sqrt{3}}{7+4\sqrt{3}} \\ 4 + 2\sqrt{3} & -(8 + 4\sqrt{3}) \end{bmatrix} \right).$$

- (H_2) Hyperexponential-2 with $\mathrm{E}[Ser] = 1$ and $\mathrm{Var}(Ser) = 2$:

$$Ser \sim \mathrm{PH}_2 \left(\underline{\beta} = (0.5, 0.5), B = \begin{bmatrix} -(2 + \sqrt{2}) & 0 \\ 0 & -(2 - \sqrt{2}) \end{bmatrix} \right).$$

As the goal of UWC is to improve the accuracy of approximated steady-state probabilities at levels below the buffer, we consider the absolute differences between approximated and IB model probabilities for queue lengths less than C. Specifically, we are interested in the following measures, which we define as the *global maximum measure* (M_g) and the *marginal maximum measure* (M_m):

$$M_g = \max_{(n,y):n<C} \left| \pi_{n,y} - \pi^A_{n,y} \right|, \tag{3.42}$$

$$M_m = \max_{n:n<C} \left| \pi_{n,\bullet} - \pi^A_{n,\bullet} \right|, \tag{3.43}$$

where n is a queue length, y is a service phase (which we let take on a value of 0 when the server is idle at $n = 0$), A is a placeholder for which model approximation we are using (UWC version 1, UWC version 2, or FB), and

$$\pi_{n,\bullet} = \sum_y \pi_{n,y} \text{ and } \pi^A_{n,\bullet} = \sum_y \pi^A_{n,y}$$

are marginal queue length probabilities. Essentially, M_g acts as a measure of the state-wise convergence to the true IB values while M_m considers the accuracy of the total probability of observing the queue at a given length. Due to the small state space of the considered example, the usefulness of M_m over M_g is limited, but it will be more valuable in later sections when we consider substantially larger state spaces having smaller probabilities at individual states (and hence, very small values of M_g).

Examining Tables 2 and 3, while it is clear that we do not recover $\underline{\pi}_i^{\text{UWC}} = \underline{\pi}_i$, $i = 0, 1, \ldots, C-1$, by comparing the values of M_g and M_m, UWC version 2 results in approximations that are very close to the true steady-state distribution and UWC version 1 still gives a better overall approximation than the FB model at these levels (for a given C). Letting $\text{E}[X]$ denote the expected queue length at steady state for a given model, UWC version 2 provides the best estimates followed by version 1 and then the FB model. The UWC probability vectors $\underline{p}_C^*$ are also provided. Note that these probabilities are identical for version 1 under E_2 and E_2^f service time distributions since they have equal column vectors of absorption rates.

Comparing the four service time distributions and using our global and marginal measures as a benchmark, UWC version 1 has its best performance under H_2 service while UWC version 2 has its worst. This is intuitive, as H_2 does not permit service phase transitions without service completions (and hence out of the four considered distributions, it acts most similar to an exponential distribution). In contrast, the E_2 distribution will always observe one such transition, while the C_2^f and E_2^f distributions will have an expected value of two and four phase transitions between service completions, respectively.

The values of our measures for a given UWC version under the other three distributions at $C = 7$ are comparable. For UWC version 1 at $C = 3$, C_2^f has the smallest measures, despite the fact that it observes twice as many phase transitions on average relative to E_2. Therefore, the presence of absorbing rates equalling zero (e.g., $B_{1,0} = 0$) also appears to have a slight negative impact on the efficacy of UWC. Also, since $\text{E}[X^{\text{UWC}}]$ at $C = 3$ for E_2 is larger than that for E_2^f, despite E_2^f having the larger $\text{E}[X^{\text{IB}}]$, we can conclude that E_2 has a slight edge in UWC version 1 performance over E_2^f due to the latter's larger number of expected phase transitions. This observation is in agreement with the sizes of M_g and M_m.

We must also point out that it is possible to observe $\pi_{i,j}^{\text{UWC}} < \pi_{i,j}$. For example, at $i = 6$, $j = 2$, $C = 7$, and H_2 service, we have $\pi_{6,2}^{\text{UWC}} = 0.0408$ for UWC version 1 (while $\pi_{6,2} = 0.0419$). Being only one level below the truncation level, this is an illustration of the possible irregularities mentioned previously. However, if we further increase C, the distance between level 6 and the truncation increases and we observe $\pi_{6,2}^{\text{UWC}} = 0.0418$ for $C = 8$ and $\pi_{6,2}^{\text{UWC}} = 0.0419$ for $C = 9$. In either of these cases, UWC version 1 provides a closer estimate than FB, so it is still preferable to use over FB for any of these service time distributions at a given C. Observe that at $C = 7$, M_g and M_m are eight to ten times larger for the FB model than for the UWC version 1 model.

For UWC version 2, the performance for either distribution is notably better than that of version 1. This holds even for the H_2 service time distribution where UWC version 1 performed its best and UWC version 2 performed its worst. It is also possible to observe underestimation of the true steady-state probabilities (e.g., $i = 0$, $C = 3$, H_2 service). If one can afford the extra computation time, it seems that it would be preferable to use version 2 in these cases. However, note that its gains relative to version 1 are much smaller for $C = 7$, where a larger proportion of observed departures during a level-C busy period are caused by reneging. Note also that in the case of exponential service, both versions of UWC will in fact result in the same optimal p_C^* (since we must have $\underline{\beta}_C^* = 1$, we have no error from

approximating it by $\underline{\hat{\beta}}_C^*$), and so version 2 is not required.

Remark 2 In our numerical investigations, we have observed that it is not true that UWC version 2 is always superior to version 1 in terms of accuracy. We found a counterexample demonstrating poor performance by UWC version 2 within a two-class polling model employing a k_i-limited service policy with H_2 service time distributions having extreme mixing weights (0.001 and 0.999, with very slow or very fast service times, respectively), as well as impatient customers. For an example of the analysis of this type of queueing system without the application of UWC, we refer the reader to Granville and Drekic [16]. While this extreme H_2 service time distribution caused issues for UWC version 2 at moderate values of C within that particular queueing system, the negative impact did vanish as C was increased. Using this service time distribution within the one-queue $M/PH/1+M$ model considered in this subsection, we found that UWC version 2 outperformed version 1 (which was itself better than FB). Thus, it would appear that version 2's poor performance was due to the combination of that service time distribution and the polling model framework (i.e., having multiple queues with k_i-limited service policies). We therefore recommend to carefully consider the structure of the underlying phase-type service distribution(s) and the features of the queueing system when deciding which version of UWC to use.

Table 2: Steady-state probabilities and expected queue lengths for UWC versions 1 and 2, FB, and IB $M/PH/1+M$ queueing models, under $C = 3, 7$, $\lambda = 0.9$, $\alpha = 0.1$, and E_2 or E_2^f service time distributions

E_2	$C = 3$			$C = 7$			
	UWC ver. 1	UWC ver. 2	FB	UWC ver. 1	UWC ver. 2	FB	IB
$\underline{\pi}_0$	0.2541	0.2263	0.2886	0.2269	0.2263	0.2293	0.2262
$\underline{\pi}_1$	(0.1602, 0.1144)	(0.1426, 0.1018)	(0.1819, 0.1299)	(0.1430, 0.1021)	(0.1426, 0.1018)	(0.1446, 0.1032)	(0.1426, 0.1018)
$\underline{\pi}_2$	(0.1083, 0.1125)	(0.0969, 0.1002)	(0.1234, 0.1277)	(0.0972, 0.1004)	(0.0969, 0.1001)	(0.0982, 0.1015)	(0.0969, 0.1001)
$\underline{\pi}_3$	(0.1045, 0.1461)	(0.1473, 0.1849)	(0.0505, 0.0981)	(0.0638, 0.0751)	(0.0636, 0.0748)	(0.0644, 0.0759)	(0.0636, 0.0748)
$\underline{\pi}_4$	-	-	-	(0.0391, 0.0492)	(0.0390, 0.0491)	(0.0395, 0.0497)	(0.0390, 0.0491)
$\underline{\pi}_5$	-	-	-	(0.0223, 0.0294)	(0.0222, 0.0293)	(0.0225, 0.0297)	(0.0222, 0.0293)
$\underline{\pi}_6$	-	-	-	(0.0116, 0.0163)	(0.0118, 0.0162)	(0.0119, 0.0164)	(0.0118, 0.0162)
$\underline{\pi}_7$	-	-	-	(0.0095, 0.0140)	(0.0107, 0.0155)	(0.0041, 0.0089)	(0.0059, 0.0083)
$E[X]$	1.4677	1.6352	1.2597	2.0010	2.0154	1.9473	2.0362
(M_g, M_m)	(0.0279, 0.0301)	(<0.0001, <0.0001)	(0.0623, 0.0673)	(0.0007, 0.0007)	(<0.0001, <0.0001)	(0.0031, 0.0033)	-
$\underline{p}_C^*$	(0.9950, 0.3820)	(0.5934, 0.5934)	(0, 0)	(0.8970, 0.3278)	(0.4604, 0.4604)	(0, 0)	-
E_2^f	$C = 3$			$C = 7$			
	UWC ver. 1	UWC ver. 2	FB	UWC ver. 1	UWC ver. 2	FB	IB
$\underline{\pi}_0$	0.2655	0.2342	0.3009	0.2351	0.2342	0.2382	0.2342
$\underline{\pi}_1$	(0.1437, 0.1195)	(0.1267, 0.1054)	(0.1628, 0.1354)	(0.1272, 0.1058)	(0.1267, 0.1054)	(0.1289, 0.1072)	(0.1267, 0.1054)
$\underline{\pi}_2$	(0.1036, 0.1079)	(0.0916, 0.0951)	(0.1177, 0.1222)	(0.0920, 0.0955)	(0.0916, 0.0951)	(0.0932, 0.0968)	(0.0916, 0.0951)
$\underline{\pi}_3$	(0.1199, 0.1399)	(0.1645, 0.1824)	(0.0690, 0.0919)	(0.0652, 0.0708)	(0.0649, 0.0705)	(0.0661, 0.0717)	(0.0649, 0.0705)
$\underline{\pi}_4$	-	-	-	(0.0430, 0.0476)	(0.0428, 0.0474)	(0.0436, 0.0482)	(0.0428, 0.0474)
$\underline{\pi}_5$	-	-	-	(0.0263, 0.0296)	(0.0262, 0.0295)	(0.0267, 0.0300)	(0.0262, 0.0295)
$\underline{\pi}_6$	-	-	-	(0.0149, 0.0172)	(0.0150, 0.0171)	(0.0152, 0.0174)	(0.0150, 0.0171)
$\underline{\pi}_7$	-	-	-	(0.0138, 0.0160)	(0.0155, 0.0180)	(0.0072, 0.0096)	(0.0080, 0.0092)
$E[X]$	1.4656	1.6465	1.2607	2.0587	2.0784	1.9931	2.1077
(M_g, M_m)	(0.0313, 0.0313)	(<0.0001, <0.0001)	(0.0668, 0.0668)	(0.0009, 0.0009)	(<0.0001, <0.0001)	(0.0041, 0.0041)	-
$\underline{p}_C^*$	(0.9950, 0.3820)	(0.6129, 0.6129)	(0, 0)	(0.8970, 0.3278)	(0.4857, 0.4857)	(0, 0)	-

Table 3: Steady-state probabilities and expected queue lengths for UWC versions 1 and 2, FB, and IB $M/PH/1+M$ queueing models, under $C = 3, 7$, $\lambda = 0.9$, $\alpha = 0.1$, and C_2^f or H_2 service time distributions

C_2^f	$C=3$			$C=7$			
	UWC ver. 1	UWC ver. 2	FB	UWC ver. 1	UWC ver. 2	FB	IB
$\underline{\pi}_0$	0.2772	0.2559	0.3355	0.2568	0.2560	0.2630	0.2560
$\underline{\pi}_1$	(0.1899, 0.0198)	(0.1754, 0.0183)	(0.2300, 0.0239)	(0.1760, 0.0183)	(0.1755, 0.0183)	(0.1803, 0.0188)	(0.1755, 0.0183)
$\underline{\pi}_2$	(0.1627, 0.0113)	(0.1539, 0.0101)	(0.2029, 0.0132)	(0.1543, 0.0101)	(0.1538, 0.0101)	(0.1580, 0.0104)	(0.1538, 0.0101)
$\underline{\pi}_3$	(0.3218, 0.0174)	(0.3649, 0.0215)	(0.1871, 0.0074)	(0.1227, 0.0075)	(0.1223, 0.0075)	(0.1257, 0.0077)	(0.1223, 0.0075)
$\underline{\pi}_4$	-	-	-	(0.0903, 0.0054)	(0.0901, 0.0054)	(0.0925, 0.0055)	(0.0901, 0.0054)
$\underline{\pi}_5$	-	-	-	(0.0618, 0.0036)	(0.0617, 0.0036)	(0.0635, 0.0037)	(0.0617, 0.0036)
$\underline{\pi}_6$	-	-	-	(0.0389, 0.0023)	(0.0396, 0.0022)	(0.0409, 0.0023)	(0.0396, 0.0022)
$\underline{\pi}_7$	-	-	-	(0.0494, 0.0025)	(0.0512, 0.0028)	(0.0267, 0.0011)	(0.0238, 0.0013)
E[X]	1.5751	1.6808	1.2695	2.2342	2.2479	2.1173	2.3042
(M_g, M_m)	(0.0212, 0.0212)	(0.0001, 0.0001)	(0.0795, 0.0795)	(0.0008, 0.0008)	(<0.0001, <0.0001)	(0.0070, 0.0070)	-
$\underline{p}_C^*$	(0.8355, 0.1157)	(0.6592, 0.6592)	(0, 0)	(0.6555, 0.1101)	(0.5331, 0.5331)	(0, 0)	-
H_2	$C=3$			$C=7$			
	UWC ver. 1	UWC ver. 2	FB	UWC ver. 1	UWC ver. 2	FB	IB
$\underline{\pi}_0$	0.2791	0.2649	0.3472	0.2663	0.2657	0.2746	0.2658
$\underline{\pi}_1$	(0.0483, 0.1473)	(0.0458, 0.1401)	(0.0600, 0.1839)	(0.0460, 0.1408)	(0.0459, 0.1405)	(0.0475, 0.1452)	(0.0460, 0.1406)
$\underline{\pi}_2$	(0.0247, 0.1300)	(0.0224, 0.1289)	(0.0285, 0.1740)	(0.0227, 0.1289)	(0.0227, 0.1286)	(0.0234, 0.1328)	(0.0227, 0.1286)
$\underline{\pi}_3$	(0.0325, 0.3380)	(0.0394, 0.3584)	(0.0071, 0.1993)	(0.0142, 0.1081)	(0.0142, 0.1079)	(0.0147, 0.1115)	(0.0142, 0.1079)
$\underline{\pi}_4$	-	-	-	(0.0095, 0.0843)	(0.0095, 0.0841)	(0.0098, 0.0870)	(0.0095, 0.0841)
$\underline{\pi}_5$	-	-	-	(0.0063, 0.0612)	(0.0062, 0.0613)	(0.0064, 0.0637)	(0.0063, 0.0613)
$\underline{\pi}_6$	-	-	-	(0.0042, 0.0408)	(0.0039, 0.0420)	(0.0037, 0.0449)	(0.0039, 0.0419)
$\underline{\pi}_7$	-	-	-	(0.0044, 0.0621)	(0.0051, 0.0625)	(0.0008, 0.0341)	(0.0023, 0.0269)
E[X]	1.6166	1.6821	1.2680	2.3054	2.3153	2.1570	2.3922
(M_g, M_m)	(0.0133, 0.0133)	(0.0009, 0.0009)	(0.0814, 0.0814)	(0.0010, 0.0008)	(0.0001, 0.0001)	(0.0088, 0.0088)	-
$\underline{p}_C^*$	(0.2404, 0.8119)	(0.6787, 0.6787)	(0, 0)	(0.2173, 0.6360)	(0.5630, 0.5630)	(0, 0)	-

4 N-Queue $M/PH/1+M$ Exhaustive Polling System

4.1 Model Assumptions

We consider a polling system of N queues, $Q_1, Q_2, \ldots, Q_N$, which are visited in a cyclic order by a lone server. The server follows an exhaustive service discipline such that after visiting a queue, they do not leave until it has emptied. If the server arrives to a queue and finds it to be empty, they immediately move on to the next queue. Let a class-i switch-in time denote the amount of time that it takes the server to switch from Q_{i-1} to Q_i (where Q_0 represents Q_N). It is assumed that switch-in times are independent, and class-i switch-in times follow a $\text{PH}_{s_i}(\underline{\gamma}_i, S_i)$ distribution with column vector of absorption rates $\underline{S}'_{0,i} = -S_i\underline{e}'$, $i = 1, 2, \ldots, N$. Furthermore, we assume that switch-in times are strictly positive in duration (i.e., $\underline{\gamma}_i\underline{e}' = 1$).

Each Q_i has its own class of customers who arrive according to an independent Poisson process with parameter λ_i, $i = 1, 2, \ldots, N$. Class-i customers are served on a first-come-first-served basis within their queue, having iid service time requirements $Ser_i \sim \text{PH}_{b_i}(\underline{\beta}_i, B_i)$ with column vector of absorption rates $\underline{B}'_{0,i} = -B_i\underline{e}'$. Additionally, class-$i$ customers are assumed to have independent $\text{Exp}(\alpha_i)$ impatience timers and are at risk of reneging until they reach the server. By setting $\alpha_i = 0$, we would have the case where class-i customers are patient and are not at risk of reneging, but for the purposes of this section we assume that $\alpha_i > 0$, $i = 1, 2, \ldots, N$.

We truncate the length of Q_i at $C_i < \infty$. Define $p^*_{i,j}$ as the UWC probability applied to class i when Q_i has C_i observed customers and the server is in phase j of a class-i customer's service, $j = 1, 2, \ldots, b_i$. Here, for ease of notation, we suppress the dependency of $p^*_{i,j}$ on C_i. When the server is not currently visiting Q_i and there are C_i observed class-i customers, we use UWC probability $p^*_{i,0}$. If the server is not at Q_i, then Q_i acts as an $M/M/\infty$ queue with $Ser \sim \text{Exp}(\alpha_i)$. It is therefore a logical choice to apply Equation (2.10) and set

$$p^*_{i,0} = 1 - \frac{(\lambda_i/\alpha_i)^{C_i}}{C_i!}\left(e^{\lambda_i/\alpha_i} - \sum_{k=0}^{C_i-1}\frac{(\lambda_i/\alpha_i)^k}{k!}\right)^{-1}.$$

For $p^*_{i,j}$, $j = 1, 2, \ldots, b_i$, we elect to use analogues of UWC versions 1 and 2 from Section 3.2. For UWC version 1, we use Equation (3.39) and set

$$p^*_{i,j} = 1 - \left(\sum_{k=1}^{\infty}\frac{\lambda^{k-1}(B_{0,i,j} + (C_i-1)\alpha_i)}{\prod_{n=0}^{k-1}(B_{0,i,j} + (C_i - 1 + n)\alpha_i)}\right)^{-1}, \ j = 1, 2, \ldots, b_i, \tag{4.1}$$

where $B_{0,i,j} = (\underline{B}'_{0,i})_j$. For UWC version 2, we require an initial probability vector for the first service in the level-C busy period for every class $i = 1, 2, \ldots, N$. Following a similar logic to what was used to derive Equation (3.20), the steady-state probability of the IB model initializing a level-C_i busy period in service phase y, $y = 1, 2, \ldots, b_i$, is

$$\frac{\sum_{n_j, j\neq i}(\lambda_i\pi_{n_1,\ldots,n_{i-1},C_i-1,n_{i+1},\ldots,n_N,2i,y} + \underline{\pi}_{n_1,\ldots,n_{i-1},C_i,n_{i+1},\ldots,n_N,2i-1}\underline{S}'_{0,i}\beta_{i,y})}{\sum_{n_j, j\neq i}(\lambda_i\underline{\pi}_{n_1,\ldots,n_{i-1},C_i-1,n_{i+1},\ldots,n_N,2i}\underline{e}' + \underline{\pi}_{n_1,\ldots,n_{i-1},C_i,n_{i+1},\ldots,n_N,2i-1}\underline{S}'_{0,i})},$$

where $\beta_{i,y} = (\underline{\beta}_i)_y$ and the steady-state probabilities are defined in the following subsection. We now define the corresponding modified phase-type initial probability row vector

$$\underline{\beta}_i^* = \frac{\sum\limits_{n_j, j \neq i} (\lambda_i \underline{\pi}_{n_1,\ldots,n_{i-1},C_i-1,n_{i+1},\ldots,n_N,2i} + \underline{\pi}_{n_1,\ldots,n_{i-1},C_i,n_{i+1},\ldots,n_N,2i-1} \underline{S}'_{0,i} \underline{\beta}_i)}{\sum\limits_{n_j, j \neq i} (\lambda_i \underline{\pi}_{n_1,\ldots,n_{i-1},C_i-1,n_{i+1},\ldots,n_N,2i} \underline{e}' + \underline{\pi}_{n_1,\ldots,n_{i-1},C_i,n_{i+1},\ldots,n_N,2i-1} \underline{S}'_{0,i})}. \tag{4.2}$$

In general, we do not know the true IB model steady-state probabilities, so we again approximate this by using FB model steady-state probabilities and refer to the approximated vector as $\hat{\underline{\beta}}_i^*$. Given these probability vectors, we repeat the numerical procedure for UWC version 2 for every class to obtain UWC probabilities p_i^*, and then let each $p_{i,j}^* = p_i^*$, $j = 1, 2, \ldots, b_i$, $i = 1, 2, \ldots, N$.

4.2 State Space and Steady-State Probabilities

This N-queue system may be modelled by the CTMC

$$\{(X_1(t), X_2(t), \ldots, X_N(t), L(t), Y(t)), t \geq 0\},$$

where $X_i(t) \in \{0, 1, \ldots, C_i\}$ is the number of class-i customers in the system, $i = 1, 2, \ldots, N$, $L(t) \in \{1, 2, \ldots, 2N-1, 2N\}$ denotes the location of the server, where $L(t) = 2i - 1$ if the server is conducting a class-i switch-in or $L(t) = 2i$ if they are serving class i, such that

$$L(t) \in \Omega_L(X_1(t), X_2(t), \ldots, X_N(t)) = \bigcup_{i=1}^{N} \Omega_L(X_i(t)),$$

where we define for $i = 1, 2, \ldots, N$,

$$\Omega_L(X_i(t)) = \begin{cases} \{2i-1\} & , \text{ if } X_i(t) = 0, \\ \{2i-1, 2i\} & , \text{ if } X_i(t) > 0, \end{cases}$$

and $Y(t)$ tracks the current service or switch-in phase, taking possible values depending on $L(t)$ as follows:

$$Y(t) \in \Omega_Y(L(t)) = \begin{cases} \{1, 2, \ldots, s_i\} & , \text{ if } L(t) = 2i - 1, \\ \{1, 2, \ldots, b_i\} & , \text{ if } L(t) = 2i. \end{cases}$$

Letting $s = \sum_{i=1}^{N} s_i$, the total number of states of this CTMC are

$$s \prod_{i=1}^{N} (C_i + 1) + \sum_{j=1}^{N} b_j \prod_{i=1}^{N} (C_i + 1 - \delta_{i,j}). \tag{4.3}$$

Let $\pi_{n_1,n_2,\ldots,n_N,l,y}$ denote the steady-state probability of observing the CTMC in state $(n_1, n_2, \ldots, n_N, l, y)$. As we are truncating Q_i at C_i, $i = 1, 2, \ldots, N$, these are not IB model

probabilities, but rather UWC model probabilities by default, or FB model probabilities if we let every $p^*_{i,j} = 0$. If we are to treat them as being approximately equal to the true IB model probabilities, sufficiently large values of C_i must be chosen.

For $i = 1, 2, \ldots, N$, we organize the steady-state probabilities into ordered row vectors as follows:

$$\underline{\pi}_{n_1,n_2,\ldots,n_N,l} = \begin{cases} (\pi_{n_1,n_2,\ldots,n_N,l,1}, \pi_{n_1,n_2,\ldots,n_N,l,2}, \ldots, \pi_{n_1,n_2,\ldots,n_N,l,s_i}) & , \text{ if } l = 2i-1, \\ (\pi_{n_1,n_2,\ldots,n_N,l,1}, \pi_{n_1,n_2,\ldots,n_N,l,2}, \ldots, \pi_{n_1,n_2,\ldots,n_N,l,b_i}) & , \text{ if } l = 2i. \end{cases}$$

Next, these vectors are further sorted into

$$\underline{\pi}_{n_1,n_2,\ldots,n_N} = (\underline{\pi}^{[1]}_{n_1,n_2,\ldots,n_N}, \underline{\pi}^{[2]}_{n_1,n_2,\ldots,n_N}, \ldots, \underline{\pi}^{[N]}_{n_1,n_2,\ldots,n_N}),$$

where

$$\underline{\pi}^{[i]}_{n_1,n_2,\ldots,n_N} = \begin{cases} \underline{\pi}_{n_1,n_2,\ldots,n_N,2i-1} & , \text{ if } n_i = 0, \\ (\underline{\pi}_{n_1,n_2,\ldots,n_N,2i-1}, \underline{\pi}_{n_1,n_2,\ldots,n_N,2i}) & , \text{ if } n_i > 0. \end{cases}$$

We finally group these vectors into probability row vectors as follows:

$$\underline{\pi}_{n_1} = (\underline{\pi}_{n_1,0}, \underline{\pi}_{n_1,1}, \ldots, \underline{\pi}_{n_1,C_2}),$$

$$\underline{\pi}_{n_1,n_2} = (\underline{\pi}_{n_1,n_2,0}, \underline{\pi}_{n_1,n_2,1}, \ldots, \underline{\pi}_{n_1,n_2,C_3}),$$

and in general for $i = 1, 2, \ldots, N-1$,

$$\underline{\pi}_{n_1,n_2,\ldots,n_i} = (\underline{\pi}_{n_1,n_2,\ldots,n_i,0}, \underline{\pi}_{n_1,n_2,\ldots,n_i,1}, \ldots, \underline{\pi}_{n_1,n_2,\ldots,n_i,C_{i+1}}),$$

such that $\underline{\pi} = (\underline{\pi}_0, \underline{\pi}_1, \ldots, \underline{\pi}_{C_1})$ is the combined probability row vector having $C_1 + 1$ levels. We can solve for these probabilities by applying the level-dependent QBD algorithm to the QBD specified in Section 4.3.

We extend the measures defined in Equations (3.42) and (3.43) to correspond to the state space of this queueing model. As before, the global maximum measure is the maximum absolute difference between steady-state probabilities at any given state (omitting states with $n_i = C_i$), whereas we now must consider marginal maximum measures for each queue. Allowing A to once again represent which model approximation we are considering, these measures are defined as:

$$M_g = \max_{(n_1,n_2,\ldots,n_N,l,y):n_i<C_i,i=1,2,\ldots,N} \left| \pi_{n_1,n_2,\ldots,n_N,l,y} - \pi^A_{n_1,n_2,\ldots,n_N,l,y} \right|, \tag{4.4}$$

$$M_{m,i} = \max_{n_i:n_i<C_i} \left| \pi^m_{i,n_i} - \pi^{m,A}_{i,n_i} \right|, \ i = 1, 2, \ldots, N. \tag{4.5}$$

where

$$\pi^m_{i,n_i} = \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{n_j=1}^{C_j} \sum_{l} \sum_{y} \pi_{n_1,\ldots,n_{i-1},n_i,n_{i+1},\ldots,n_N,l,y}$$

and

$$\pi^{m,A}_{i,n_i} = \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{n_j=1}^{C_j} \sum_{l} \sum_{y} \pi^{A}_{n_1,\ldots,n_{i-1},n_i,n_{i+1},\ldots,n_N,l,y}$$

are marginal queue length probabilities for queue i. Note that we do allow the consideration of states with $n_j = C_j$ when calculating $M_{m,i}$, $i \neq j$.

4.3 Infinitesimal Generator Matrix

Letting the value of $X_1(t)$ denote the level of the process, we now construct the generator blocks, $Q_{i,j}$, which contain all transition probabilities that result in the level changing from i to j. To begin, we let

$$\underline{n}_i = (n_1, \ldots, n_i),\ i = 2, 3, \ldots, N,$$

denote a row vector of particular queue lengths of the first i queues, so that $(\underline{n}_i, m)$ is equal to $\underline{n}_{i+1}$ with $n_{i+1} = m$. Next, we define

$$\lambda_{\underline{n}_N} = \sum_{i=1}^{N} \lambda_i (1 - \delta_{n_i,C_i}),$$

$$a^{[m,n]}_{\underline{n}_N} = \sum_{i=m}^{n} (s_i + (1 - \delta_{n_i,0}) b_i),\ 1 \leq m \leq n \leq N,$$

$$\underline{p}^*_i = (p^*_{i,1}, p^*_{i,2}, \ldots, p^*_{i,b_i}),$$

$$p^*_{i,n_i,l,y} = \begin{cases} p^*_{i,0} & ,\ \text{if } n_i = C_i,\ l \neq 2i, \\ p^*_{i,y} & ,\ \text{if } n_i = C_i,\ l = 2i, \\ 0 & ,\ \text{otherwise,} \end{cases}$$

$$\alpha_{\underline{n}_N,l,y} = \sum_{i=1}^{N} \alpha_i (n_i - \delta_{l,2i})(1 - p^*_{i,n_i,l,y}),$$

$$\underline{\alpha}_{\underline{n}_N,l} = \begin{cases} (\alpha_{\underline{n}_N,l,1}, \alpha_{\underline{n}_N,l,2}, \ldots, \alpha_{\underline{n}_N,l,s_i}) & ,\ \text{if } l = 2i-1, \\ (\alpha_{\underline{n}_N,l,1}, \alpha_{\underline{n}_N,l,2}, \ldots, \alpha_{\underline{n}_N,l,b_i}) & ,\ \text{if } l = 2i, \end{cases}$$

and

$$B^*_{i,n_i} = B_i + \delta_{n_i,C_i} \text{diag}(\underline{p}^*_i) \underline{B}'_{0,i} \underline{\beta}_i,\ i = 1, 2, \ldots, N.$$

We first construct blocks to track movements in $X_N(t)$, after which we will recursively build outwards to track all queue lengths. We achieve this by modelling changes of $X_j(t)$ for given values of $X_1(t), X_2(t), \ldots, X_{j-1}(t)$ using a QBD structure. These will be nested

within each other, with the innermost QBD describing $X_N(t)$. For $n_i = 0, 1, \ldots, C_i$, $i = 1, 2, \ldots, N-1$, we define

$$Q^{[N]}_{\underline{n}_{N-1}} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ C_N-1 \\ C_N \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \cdots & C_N-1 & C_N \end{array} \\ \begin{bmatrix} \Delta_{(\underline{n}_{N-1},0)} & (UD)^{[N]}_{(\underline{n}_{N-1},0)} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ (LD)^{[N]}_{(\underline{n}_{N-1},1)} & \Delta_{(\underline{n}_{N-1},1)} & (UD)^{[N]}_{(\underline{n}_{N-1},1)} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (LD)^{[N]}_{(\underline{n}_{N-1},2)} & \Delta_{(\underline{n}_{N-1},2)} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \Delta_{(\underline{n}_{N-1},C_N-1)} & (UD)^{[N]}_{(\underline{n}_{N-1},C_N-1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & (LD)^{[N]}_{(\underline{n}_{N-1},C_N)} & \Delta_{(\underline{n}_{N-1},C_N)} \end{bmatrix} \end{array}.$$

Letting

$$\zeta_{\underline{n}_N,l} = \begin{cases} S_j - \lambda_{\underline{n}_N} I_{s_j} - \mathrm{diag}(\underline{\alpha}_{\underline{n}_N,2j-1}) & , \text{ if } l = 2j-1, \\ B^*_{j,n_j} - \lambda_{\underline{n}_N} I_{b_j} - \mathrm{diag}(\underline{\alpha}_{\underline{n}_N,2j}) & , \text{ if } l = 2j, \end{cases}$$

the main diagonal blocks of $Q^{[N]}_{n_1,\ldots,n_{N-1}}$ are

$$\Delta_{\underline{n}_N} = \begin{bmatrix} \Delta^{[1]}_{\underline{n}_N} \\ \Delta^{[2]}_{\underline{n}_N} \\ \vdots \\ \Delta^{[N]}_{\underline{n}_N} \end{bmatrix},$$

where

$$\Delta^{[1]}_{\underline{n}_N} = \begin{cases} \begin{bmatrix} \zeta_{\underline{n}_N,1} & \underline{S}'_{0,1}\underline{\gamma}_2 & \underline{0}'_{s_1}\underline{0}_{a^{[2,N]}_{\underline{n}_N}-s_2} \end{bmatrix} & , \text{ if } n_1 = 0, \\ \begin{bmatrix} \zeta_{\underline{n}_N,1} & \underline{S}'_{0,1}\underline{\beta}_1 & \underline{0}'_{s_1}\underline{0}_{a^{[2,N]}_{\underline{n}_N}} \\ \underline{0}'_{b_1}\underline{0}_{s_1} & \zeta_{\underline{n}_N,2} & \underline{0}'_{b_1}\underline{0}_{a^{[2,N]}_{\underline{n}_N}} \end{bmatrix} & , \text{ if } n_1 = 1, 2, \ldots, C_1, \end{cases}$$

while for $j = 2, 3 \ldots, N-1$,

$$\Delta^{[j]}_{\underline{n}_N} = \begin{cases} \begin{bmatrix} \underline{0}'_{s_j}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}} & \zeta_{\underline{n}_N,2j-1} & \underline{S}'_{0,j}\underline{\gamma}_{j+1} & \underline{0}'_{s_j}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}-s_{j+1}} \end{bmatrix} & , \text{ if } n_j = 0, \\ \begin{bmatrix} \underline{0}'_{s_j}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}} & \zeta_{\underline{n}_N,2j-1} & \underline{S}'_{0,j}\underline{\beta}_j & \underline{0}'_{s_j}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}} \\ \underline{0}'_{b_j}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}} & \underline{0}'_{b_j}\underline{0}_{s_j} & \zeta_{\underline{n}_N,2j} & \underline{0}'_{b_j}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}} \end{bmatrix} & , \text{ if } n_j = 1, 2, \ldots, C_j, \end{cases}$$

and

$$\Delta^{[N]}_{\underline{n}_N} = \begin{cases} \begin{bmatrix} \underline{S}'_{0,N}\underline{\gamma}_1 & \underline{0}'_{s_N}\underline{0}_{a^{[1,N-1]}_{\underline{n}_N}-s_1} & \zeta_{n_1,\ldots,n_N,2N-1} \end{bmatrix} & , \text{ if } n_N = 0, \\ \begin{bmatrix} \underline{0}'_{s_N}\underline{0}_{a^{[1,N-1]}_{\underline{n}_N}} & \zeta_{\underline{n}_N,2N-1} & \underline{S}'_{0,N}\beta_N \\ \underline{0}'_{b_N}\underline{0}_{a^{[1,N-1]}_{\underline{n}_N}} & \underline{0}'_{b_N}\underline{0}_{s_N} & \zeta_{\underline{n}_N,2N} \end{bmatrix} & , \text{ if } n_N = 1, 2, \ldots, C_N. \end{cases}$$

The upper diagonal blocks of $Q^{[N]}_{n_1,\ldots,n_{N-1}}$ are

$$(UD)^{[N]}_{\underline{n}_N} = \left[\begin{array}{cc} \lambda_N I_{a^{[1,N-1]}_{\underline{n}_N}+s_N} & \underline{0}'_{a^{[1,N-1]}_{\underline{n}_N}+s_N}\underline{0}_{b_N} \end{array} \right]$$

for $n_N = 0$ and

$$(UD)^{[N]}_{\underline{n}_N} = \lambda_N I_{a^{[1,N]}_{\underline{n}_N}}$$

for $n_N = 1, 2, \ldots, C_N - 1$, and the lower diagonal blocks are

$$(LD)^{[N]}_{\underline{n}_N} = \left[\begin{array}{cc} \alpha_N(1-\delta_{n_N,C_N}p^*_{N,0})I_{s_1} & \underline{0}'_{s_1}\underline{0}_{a^{[1,N-1]}_{\underline{n}_N}+s_N-s_1} \\ \underline{0}'_{a^{[1,N-1]}_{\underline{n}_N}+s_N-s_1}\underline{0}_{s_1} & \alpha_N(1-\delta_{n_N,C_N}p^*_{N,0})I_{a^{[1,N-1]}_{\underline{n}_N}+s_N-s_1} \\ (I_{b_N}-\delta_{n_N,C_N}\mathrm{diag}(\underline{p}^*_N))\underline{B}'_{0,N}\underline{\gamma}_1 & \underline{0}'_{b_N}\underline{0}_{a^{[1,N-1]}_{\underline{n}_N}+s_N-s_1} \end{array} \right]$$

for $n_N = 1$ and

$$(LD)^{[N]}_{\underline{n}_N} = \left[\begin{array}{cc} n_N\alpha_N(1-\delta_{n_N,C_N}p^*_{N,0})I_{a^{[1,N-1]}_{\underline{n}_N}+s_N} & \underline{0}'_{a^{[1,N-1]}_{\underline{n}_N}+s_N}\underline{0}_{b_N} \\ \underline{0}'_{b_N}\underline{0}_{a^{[1,N-1]}_{\underline{n}_N}+s_N} & (I_{b_N}-\delta_{n_N,c_N}\mathrm{diag}(\underline{p}^*_N))((n_N-1)\alpha_N I_{b_N}+\underline{B}'_{0,N}\underline{\beta}_N) \end{array} \right]$$

for $n_N = 2, 3, \ldots, C_N$.

We now build the QBD structured blocks that are needed to track changes in $X_j(t)$, $j = 2, 3, \ldots, N-1$. For $n_i = 0, 1, \ldots, C_i$, $i = 1, 2, \ldots, j-1$, we define

$$Q^{[j]}_{\underline{n}_{j-1}} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ C_j - 1 \\ C_j \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \cdots & C_j-1 & C_j \end{array} \\ \left[\begin{array}{ccccccc} Q^{[j+1]}_{(\underline{n}_{j-1},0)} & (UD)^{[j]}_{(\underline{n}_{j-1},0)} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ (LD)^{[j]}_{(\underline{n}_{j-1},1)} & Q^{[j+1]}_{(\underline{n}_{j-1},1)} & (UD)^{[j]}_{(\underline{n}_{j-1},1)} & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (LD)^{[j]}_{(\underline{n}_{j-1},2)} & Q^{[j+1]}_{(\underline{n}_{j-1},2)} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & Q^{[j+1]}_{(\underline{n}_{j-1},C_j-1)} & (UD)^{[j]}_{(\underline{n}_{j-1},C_j-1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & (LD)^{[j]}_{(\underline{n}_{j-1},C_j)} & Q^{[j+1]}_{(\underline{n}_{j-1},C_j)} \end{array} \right] \end{array}.$$

Note how the main diagonal blocks of $Q^{[j]}_{\underline{n}_{j-1}}$ take the form of $Q^{[j+1]}_{\underline{n}_j}$, implying that these must be constructed recursively, starting with our original $Q^{[N]}_{\underline{n}_{N-1}}$ blocks. The upper and lower diagonal blocks make use of a similar recursion in their definitions. The upper diagonal blocks are $(UD)^{[j]}_{\underline{n}_j}$, where

$$(UD)^{[j]}_{\underline{n}_{j+k}} = \left[\begin{array}{cccc} (UD)^{[j]}_{(\underline{n}_{j+k},0)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & (UD)^{[j]}_{(\underline{n}_{j+k},1)} & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & (UD)^{[j]}_{(\underline{n}_{j+k},C_{j+k+1})} \end{array} \right] \tag{4.6}$$

for $k = 0, 1, \ldots, N-j-1$, with

$$(UD)^{[j]}_{\underline{n}_N} = \begin{bmatrix} \lambda_j I_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & \underline{0}'_{a^{[1,j-1]}_{\underline{n}_N}+s_j}\underline{0}_{b_j} & \underline{0}'_{a^{[1,j-1]}_{\underline{n}_N}+s_j}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}} \\ \underline{0}'_{a^{[j+1,N]}_{\underline{n}_N}}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & \underline{0}'_{a^{[j+1,N]}_{\underline{n}_N}}\underline{0}_{b_j} & \lambda_j I_{a^{[j+1,N]}_{\underline{n}_N}} \end{bmatrix}$$

for $n_j = 0$ and

$$(UD)^{[j]}_{\underline{n}_N} = \lambda_j I_{a^{[1,N]}_{\underline{n}_N}}$$

for $n_j = 1, 2, \ldots, C_j - 1$. Similarly, the lower diagonal blocks are $(LD)^{[j]}_{\underline{n}_j}$, where

$$(LD)^{[j]}_{\underline{n}_{j+k}} = \begin{bmatrix} (LD)^{[j]}_{(\underline{n}_{j+k},0)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & (LD)^{[j]}_{(\underline{n}_{j+k},1)} & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & (LD)^{[j]}_{(\underline{n}_{j+k},C_{j+k+1})} \end{bmatrix} \tag{4.7}$$

for $k = 0, 1, \ldots, N-j-1$, with

$$(LD)^{[j]}_{\underline{n}_N} = \begin{bmatrix} \alpha_j(1-\delta_{n_j,C_j}p^*_{j,0})I_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & \underline{0}'_{a^{[1,j-1]}_{\underline{n}_N}+s_j}\underline{0}_{s_{j+1}} & \underline{0}'_{a^{[1,j-1]}_{\underline{n}_N}+s_j}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}-s_{j+1}} \\ \underline{0}'_{b_j}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & (I_{b_j}-\delta_{n_j,C_j}\mathrm{diag}(\underline{p}^*_j))\underline{B}'_{0,j}\underline{\gamma}_{j+1} & \underline{0}'_{b_j}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}-s_{j+1}} \\ \underline{0}'_{s_{j+1}}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & \alpha_j(1-\delta_{n_j,C_j}p^*_{j,0})I_{s_{j+1}} & \underline{0}'_{s_{j+1}}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}-s_{j+1}} \\ \underline{0}'_{a^{[j+1,N]}_{\underline{n}_N}-s_{j+1}}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & \underline{0}'_{a^{[j+1,N]}_{\underline{n}_N}-s_{j+1}}\underline{0}_{s_{j+1}} & \alpha_j(1-\delta_{n_j,C_j}p^*_{j,0})I_{a^{[j+1,N]}_{\underline{n}_N}-s_{j+1}} \end{bmatrix}$$

for $n_j = 1$ and

$$(LD)^{[j]}_{\underline{n}_N} = \begin{bmatrix} n_j\alpha_j(1-\delta_{n_j,C_j}p^*_{j,0})I_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & \underline{0}'_{a^{[1,j-1]}_{\underline{n}_N}+s_j}\underline{0}_{b_j} & \underline{0}'_{a^{[1,j-1]}_{\underline{n}_N}+s_j}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}} \\ \underline{0}'_{b_j}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & (I_{b_j}-\delta_{n_j,C_j}\mathrm{diag}(\underline{p}^*_j))((n_j-1)\alpha_j I_{b_j}+\underline{B}'_{0,j}\underline{\beta}_j) & \underline{0}'_{b_j}\underline{0}_{a^{[j+1,N]}_{\underline{n}_N}} \\ \underline{0}'_{a^{[j+1,N]}_{\underline{n}_N}}\underline{0}_{a^{[1,j-1]}_{\underline{n}_N}+s_j} & \underline{0}'_{a^{[j+1,N]}_{\underline{n}_N}}\underline{0}_{b_j} & n_j\alpha_j(1-\delta_{n_j,C_j}p^*_{j,0})I_{a^{[j+1,N]}_{\underline{n}_N}} \end{bmatrix}$$

for $n_j = 2, 3, \ldots, C_j$.

Finally, the complete infinitesimal generator is simply the QBD modelling changes in $X_1(t)$, namely

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ C_1-1 \\ C_1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & C_1-1 & C_1 \end{array} \\ \begin{bmatrix} Q^{[2]}_0 & (UD)^{[1]}_0 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ (LD)^{[1]}_1 & Q^{[2]}_1 & (UD)^{[1]}_1 & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (LD)^{[1]}_2 & Q^{[2]}_2 & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & Q^{[2]}_{C_1-1} & (UD)^{[1]}_{C_1-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & (LD)^{[1]}_{C_1} & Q^{[2]}_{C_1} \end{bmatrix} \end{array},$$

where we again use Equations (4.6) and (4.7), with

$$(UD)^{[1]}_{\underline{n}_N} = \begin{bmatrix} \lambda_1 I_{s_1} & \underline{0}'_{s_1}\underline{0}_{b_1} & \underline{0}'_{s_1}\underline{0}_{a^{[2,N]}_{\underline{n}_N}} \\ \underline{0}'_{a^{[2,N]}_{\underline{n}_N}}\underline{0}_{s_1} & \underline{0}'_{a^{[2,N]}_{\underline{n}_N}}\underline{0}_{b_1} & \lambda_1 I_{a^{[2,N]}_{\underline{n}_N}} \end{bmatrix}$$

for $n_1 = 0$ and

$$(UD)^{[1]}_{\underline{n}_N} = \lambda_1 I_{a^{[1,N]}_{\underline{n}_N}}$$

for $n_1 = 1, 2, \ldots, C_1 - 1$, and

$$(LD)^{[1]}_{\underline{n}_N} = \begin{bmatrix} \alpha_1(1-\delta_{n_1,C_1}p^*_{1,0})I_{s_1} & \underline{0}'_{s_1}\underline{0}_{s_2} & \underline{0}'_{s_1}\underline{0}_{a^{[2,N]}_{\underline{n}_N}-s_2} \\ \underline{0}'_{b_1}\underline{0}_{s_1} & (I_{b_1}-\delta_{n_1,C_1}\mathrm{diag}(\underline{p}^*_1))\underline{B}'_{0,1}\underline{\gamma}_2 & \underline{0}'_{b_1}\underline{0}_{a^{[2,N]}_{\underline{n}_N}-s_2} \\ \underline{0}'_{s_2}\underline{0}_{s_1} & \alpha_1(1-\delta_{n_1,C_1}p^*_{1,0})I_{s_2} & \underline{0}'_{s_2}\underline{0}_{a^{[2,N]}_{\underline{n}_N}-s_2} \\ \underline{0}'_{a^{[2,N]}_{\underline{n}_N}-s_2}\underline{0}_{s_1} & \underline{0}'_{a^{[2,N]}_{\underline{n}_N}-s_2}\underline{0}_{s_2} & \alpha_1(1-\delta_{n_1,C_1}p^*_{1,0})I_{a^{[2,N]}_{\underline{n}_N}-s_2} \end{bmatrix}$$

for $n_1 = 1$ and

$$(LD)^{[1]}_{\underline{n}_N} = \begin{bmatrix} n_1\alpha_1(1-\delta_{n_1,C_1}p^*_{1,0})I_{s_1} & \underline{0}'_{s_1}\underline{0}_{b_1} & \underline{0}'_{s_1}\underline{0}_{a^{[2,N]}_{\underline{n}_N}} \\ \underline{0}'_{b_1}\underline{0}_{s_1} & (I_{b_1}-\delta_{n_1,C_1}\mathrm{diag}(\underline{p}^*_1))((n_1-1)\alpha_1 I_{b_1}+\underline{B}'_{0,1}\underline{\beta}_1) & \underline{0}'_{b_1}\underline{0}_{a^{[2,N]}_{\underline{n}_N}} \\ \underline{0}'_{a^{[2,N]}_{\underline{n}_N}}\underline{0}_{s_1} & \underline{0}'_{a^{[2,N]}_{\underline{n}_N}}\underline{0}_{b_1} & n_1\alpha_1(1-\delta_{n_1,C_1}p^*_{1,0})I_{a^{[2,N]}_{\underline{n}_N}} \end{bmatrix}$$

for $n_1 = 2, 3, \ldots, C_1$.

4.4 Numerical Examples

4.4.1 The Impact of Service Phase Transitions

Continuing from our earlier discussion of the potential impact on the effectiveness of UWC by the switching of service phases without observing customer departures, we compare mean queue lengths between our two UWC models as well as the FB model in a 2-queue system with arrival rates $\lambda_1 = \lambda_2 = 8/15$, reneging rates $\alpha_1 = \alpha_2 = 0.05$, and iid Exp(1) switch-in times. To easily scale the number of service phases while controlling for expected value, we select Erlang-k (E_k) service time distributions with means of 1 and 2 for classes 1 and 2, respectively. Note that when there is only one service phase (i.e., $k = 1$), these are simply Exp(1) and Exp(1/2) distributions, respectively.

In Figure 2, we plot $\mathrm{E}[X_i]$ for both classes for the corresponding UWC and FB models. As an impatient customer example, we treat this as a level-dependent QBD and let $C_1 = C_2 = C$, $C = 2, 3, \ldots, 20$. The number of service phases are similarly kept constant between the classes, and we present the cases for $k = 1, 2, 5, 10$. Within these plots, the light grey horizontal lines are approximated $\mathrm{E}[X_i^{\mathrm{IB}}]$ values, obtained via calculating the mean queue lengths for the FB model with $C = 40$. As expected, we would rank UWC version 2 above version 1, with both outperforming the FB model in all cases. When $k = 1$, both UWC

versions are identical due to the presence of only a single service phase. Performance is comparable for UWC version 2 across all cases, while version 1 has the widest margins between itself and the FB model in the $k = 1$ case, which we know enables the best performance by this version of UWC. We also observe in all cases that the difference in effectiveness between the UWC versions decrease as C is increased, reflecting the impact of reneging causing a larger proportion of customer departures during a level-C busy period.

This combination of parameters resulted in values of $\mathrm{E}[X_i^{\mathrm{IB}}]$ between 6.0212 ($k = 1$) and 6.3087 ($k = 10$) for class 1, and 4.3682 ($k = 1$) and 4.1088 ($k = 10$) for class 2. Due to the narrow range of limiting values, this allows us to more accurately compare rates of convergence. In Table 4, we present $\mathrm{E}[X_i^{\mathrm{UWC}}]/\mathrm{E}[X_i^{\mathrm{IB}}]$ for $C = 2, 3, \ldots, 10$ for both versions of UWC. As previously observed, the convergence percentages for the $k = 1$ cases of UWC version 1 are noticeably higher than those for $k \geq 2$. In fact, the difference in percentage for a given C between $k = 1$ and $k = 2$ is larger than that between $k = 2$ and $k = 10$! In contrast, the percentages for UWC version 2 are not very sensitive to changes in k, and even increase for class 2 (due to the fact that $\mathrm{E}[X_2^{\mathrm{IB}}]$ is decreasing in k).

The values of the global and marginal maximum measures defined in Equations (4.4) and (4.5) are plotted for these cases in Figure 3. Note that the scales for these measures are fairly different, as $M_{m,i}$ considers the total probability mass spread across multiple states whereas M_g considers individual states. Additionally, observe that the slopes of these plots are (in general) decreasing in C as the models converge towards the true IB model and their measures trend to zero. A tipping point appears to exist near $C = 5$, after which the rate of reduction of M_g and $M_{m,1}$ largely levels out (i.e., we observe a slope nearer to zero in absolute value), while the convergence of $M_{m,2}$ to zero is smoother. This may be due to class 1 having a larger expected queue length (as presented in Figure 2), resulting in larger gains in accuracy per increase in C for small C (due to more probability mass at the truncation level).

We observe that M_g decreases in k for both UWC version 2 and FB, as the larger number of service phases results in a larger state space. However, it actually appears to be increasing in k for UWC version 1, demonstrating version 1's decrease in effectiveness. Of course, this measure still remains smaller for version 1 than FB at any given k and C. Unlike M_g, $M_{m,i}$ does control for the difference in number of states. With respect to the marginal maximum measures, the performance of UWC version 2 is relatively insensitive to the increase in k while FB's accuracy gradually worsens and UWC version 1's effectiveness at low C experiences a huge reduction. Therefore, in the absence of exponential service time distributions or moderate to large C (or alternatively, large reneging rates), UWC version 2 should be used for service time distributions involving multiple phase transitions.

4.4.2 Examining Marginal Queue Length Probabilities

In Section 3.2, we considered a numerical example which compared the steady-state probabilities of the UWC version 1 and 2 models of a $M/PH/1+M$ queue against those of the FB model. We now consider a comparison of steady-state probabilities in our polling system.

Table 4: UWC version 1 and UWC version 2 model steady-state expected marginal queue length convergence percentages ($\mathrm{E}[X_i^{\mathrm{UWC}}]/\mathrm{E}[X_i^{\mathrm{IB}}]$, $i = 1, 2$) at various buffers $C_1 = C_2 = C$ under E_k service, $k = 1, 2, 5, 10$

Class 1, UWC ver. 1					C				
k	2	3	4	5	6	7	8	9	10
1	0.2788	0.4036	0.5172	0.6193	0.7090	0.7852	0.8473	0.8957	0.9315
2	0.2464	0.3693	0.4864	0.5944	0.6905	0.7725	0.8391	0.8906	0.9285
5	0.2294	0.3532	0.4728	0.5835	0.6822	0.7664	0.8348	0.8877	0.9266
10	0.2245	0.3490	0.4694	0.5808	0.6800	0.7647	0.8335	0.8868	0.9260
Class 1, UWC ver. 2					C				
k	2	3	4	5	6	7	8	9	10
1	0.2788	0.4036	0.5172	0.6193	0.7090	0.7852	0.8473	0.8957	0.9315
2	0.2762	0.3998	0.5129	0.6150	0.7050	0.7819	0.8447	0.8937	0.9302
5	0.2740	0.3967	0.5094	0.6115	0.7018	0.7790	0.8424	0.8920	0.9289
10	0.2730	0.3955	0.5080	0.6101	0.7004	0.7778	0.8414	0.8912	0.9283
Class 2, UWC ver. 1					C				
k	2	3	4	5	6	7	8	9	10
1	0.3810	0.5351	0.6615	0.7616	0.8380	0.8940	0.9333	0.9597	0.9766
2	0.3478	0.4964	0.6249	0.7317	0.8163	0.8800	0.9251	0.9553	0.9744
5	0.3283	0.4758	0.6069	0.7176	0.8063	0.8733	0.9211	0.9530	0.9731
10	0.3220	0.4697	0.6020	0.7140	0.8038	0.8717	0.9200	0.9523	0.9727
Class 2, UWC ver. 2					C				
k	2	3	4	5	6	7	8	9	10
1	0.3810	0.5351	0.6615	0.7616	0.8380	0.8940	0.9333	0.9597	0.9766
2	0.3949	0.5506	0.6762	0.7738	0.8471	0.9002	0.9373	0.9620	0.9779
5	0.4056	0.5624	0.6869	0.7823	0.8531	0.9041	0.9395	0.9632	0.9784
10	0.4097	0.5670	0.6910	0.7854	0.8552	0.9053	0.9401	0.9635	0.9785

For the benefit of simplified (and condensed) presentation of data, we consider steady-state probabilities for marginal queue lengths (rather than for individual states), and we limit ourselves to a 2-queue system. As marginal queue length probabilities are highly related to $M_{m,i}$, we elect to forgo investigating either measure within this example. We allow the service time distributions for classes 1 and 2 to be E_2^f with the following parameters:

- Class 1: $\mathrm{E}[Ser] = 1$ and $\mathrm{Var}(Ser) = 0.75$:

$$Ser \sim \mathrm{PH}_2 \left(\underline{\beta} = (1,0), B = \begin{bmatrix} -4 & 4 \\ 2 & -4 \end{bmatrix} \right).$$

- Class 2: $\mathrm{E}[Ser] = 2$ and $\mathrm{Var}(Ser) = 3$:

$$Ser \sim \mathrm{PH}_2 \left(\underline{\beta} = (1,0), B = \begin{bmatrix} -2 & 2 \\ 1 & -2 \end{bmatrix} \right).$$

Lastly, as in Section 4.4.1, we let $\lambda_1 = \lambda_2 = 8/15$ and $\alpha_1 = \alpha_2 = 0.05$, while switch-in times are assumed to be iid Exp(1) random variables.

In Figures 4 and 5, we present barplots of the marginal queue length probabilities for classes 1 and 2, respectively, for both versions of UWC as well as FB models at even buffer sizes $C_1 = C_2 = C$, $C = 2, 4, \ldots, 16$. Plotted along with these values are those from the corresponding IB model, approximated via a FB model with $C = 40$, which are unchanged between plots within a figure.

Unlike the simple single-queue case, version 2 does not immediately result in near exact steady-state probabilities for levels below C. While there is still some error present as a result of using the FB probabilities, we more importantly do not separate the cases (in terms of UWC probability) where the server begins a level-C busy period due to an arrival versus after a switch-in time. If the busy period begins after a switch, then as in our earlier discussion of the original $M/PH/1 + M$ queue UWC version 1 approximation, any cases where possible unobserved customers could be in the system at this instant are treated as if there are exactly zero unobserved customers. Unsurprisingly, this results in a failure to capture all excess probability mass at level C. However, for both classes, we are in fact observing the intended effect of probability mass being shifted to the truncation level C, resulting in better approximations at lower levels. The relative difference in gains by the two versions over the FB model are larger for small C and decrease as C is increased, consistent with what we have seen previously.

For moderate values of C, we again observe instances of underestimating the IB model steady-state probabilities at queue lengths near the buffer. While more common for UWC, it is also observed for FB (e.g., case $C = 10$ in Figure 5). Fortunately, even if the steady-state probabilities are slightly underestimated by UWC, they are still generally closer to the target probabilities than those of the FB model at the same value of C, and the underestimation vanishes as C is increased. These results indicate that the either version of the UWC model would indeed be preferable to the FB model at any given C. This experiment was also

replicated in a more optimal setting using exponentially distributed service times (the results of which we omit), which led to the same conclusions while observing higher relative accuracy gains by UWC (with the largest relative gains at small C). Overall, we maintain the same conclusion that version 2 is preferable although version 1 is comparable at moderate to high values of C.

4.4.3 Accuracy and Run Time Comparisons Between UWC and a Level-Independent Approximation

We have compared accuracy gains between both versions of UWC as well as the FB method. We now consider comparisons of both accuracy and run time between these and a *level-independent approximation* (LIA) inspired by the work of Shin and Choo [31] who made the simplifying approximation that the total effective rate of reneging does not change beyond a certain level, allowing for the use of level-independent QBD solution techniques. We consider a LIA approach rather than ETAQA since ETAQA, despite being comparable to UWC due to its aggregation nature, is restricted to only calculating exact probabilities for states in levels 0 and 1. To explain LIA, we use an infinitesimal generator in the style of Section 4.3, but remove the truncation at C_1 so that we never reference class-1 UWC probabilities and we no longer "turn off" class-1 arrivals at level C_1:

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ C_1-1 \\ C_1 \\ C_1+1 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccccc} 0 & 1 & 2 & \cdots & C_1-1 & C_1 & C_1+1 & C_1+2 & \cdots \end{array} \\ \begin{bmatrix} Q_0^{[2]} & (UD)_0^{[1]} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ (LD)_1^{[1]} & Q_1^{[2]} & (UD)_1^{[1]} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & (LD)_2^{[2]} & Q_2^{[2]} & \ddots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & Q_{C_1-1}^{[2]} & (UD)_{C_1-1}^{[1]} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & (LD)_{C_1}^{[1]} & Q_{C_1}^{[2]} & (UD)_{C_1}^{[1]} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & (LD)_{C_1+1}^{[1]} & Q_{C_1+1}^{[2]} & (UD)_{C_1+1}^{[1]} & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \end{array}.$$

In practice, we construct QBD blocks for level C_1 as if the truncation point was instead set at some arbitrary higher level. As is, this would still be a level-dependent QBD, so we make the simplifying assumptions that for $i \in \mathbb{N}$,

$$\begin{aligned} (LD)_{C_1+i}^{[1]} &= (LD)_{C_1}^{[1]}, \\ Q_{C_1+i}^{[2]} &= Q_{C_1}^{[2]}, \\ (UD)_{C_1+i}^{[1]} &= (UD)_{C_1}^{[1]}, \end{aligned}$$

and

$$\underline{\pi}_i = \begin{cases} \underline{\pi}_0 \prod_{j=1}^{i} R_j & , \text{ if } i = 1, 2, \ldots, C_1, \\ \underline{\pi}_{C_1} R^{i-C_1} & , \text{ if } i = C_1+1, C_1+2, \ldots. \end{cases}$$

Therefore, the QBD is level dependent up to level C_1, after which it becomes level independent. The implication of this approximation is that the calculated total rate of reneging for class 1 only takes into account the impatience of class-1 customers waiting in the first C_1 positions of their queue. Any additional waiting class-1 customers are assumed to be patient. This is reasonable for a large value of C_1 when the times between departures due to reneging are short and the marginal increase to the total reneging rate from an additional customer is minimial in comparison to the previous total rate. However, it can prove to be very unrealistic for small values of C_1.

It is straightforward to obtain the equation

$$\begin{aligned}\underline{0} &= \underline{\pi}_{C_1+i-1}(UD)^{[1]}_{C_1} + \underline{\pi}_{C_1+i}Q^{[2]}_{C_1+1} + \underline{\pi}_{C_1+i+2}(LD)^{[1]}_{C_1+2} \\ &= \underline{\pi}_{C_1}R^{i-1}\left((UD)^{[1]}_{C_1} + RQ^{[2]}_{C_1} + R^2(LD)^{[1]}_{C_1}\right),\end{aligned}$$

which provides us with the matrix quadratic equation

$$\mathbf{0} = (UD)^{[1]}_{C_1} + RQ^{[2]}_{C_1} + R^2(LD)^{[1]}_{C_1},$$

having matrix R as its minimal non-negative solution (e.g., Neuts [25], Theorem 1.7.1). R may be calculated by letting $R(0) = \mathbf{0}$ and iteratively applying the algorithm

$$R(i) = -\left((UD)^{[1]}_{C_1} + R(i-1)^2(LD)^{[1]}_{C_1}\right)\left(Q^{[2]}_{C_1}\right)^{-1},\ i \in \mathbb{Z}^+,$$

where $\lim_{i\to\infty} R(i) = R$ converges monotonically, element-wise. Once R is obtained, we may calculate R_i, $i = 1, 2, \ldots, C_1$, using the following formulas which are obtained in a similar fashion as Equations (3.32) and (3.33):

$$R_{C_1} = -(UD)^{[1]}_{C_1-1}\left(Q^{[2]}_{C_1} + R(LD)^{[1]}_{C_1}\right)^{-1}$$

and

$$R_i = -(UD)^{[1]}_{i-1}\left(Q^{[2]}_{i} + R_{i+1}(LD)^{[1]}_{i+1}\right)^{-1},\ i = 1, 2, \ldots, C_1 - 1,$$

which may be calculated iteratively in reverse order. Finally, in order to obtain all of our steady-state probabilities, we simply need to solve for $\underline{\pi}_0$ through the adjusted normalization condition

$$1 = \underline{\pi}\,\underline{e}' = \sum_{i=0}^{\infty}\underline{\pi}_i\underline{e}' = \underline{\pi}_0\left[I + \sum_{i=1}^{C_1-1}\left(\prod_{j=1}^{i}R_j\right) + \left(\prod_{j=1}^{C_1}R_j\right)(I-R)^{-1}\right]\underline{e}'.$$

To numerically compare these approximation methods, we consider a symmetric 3-queue system so that the choice of which class to assign as class 1 does not matter. For $i = 1, 2, 3$, we set $\lambda_i = 4/9$, $\alpha_i = 0.05$, switch-in times to be iid Exp(1), and service times to be iid E_2^f such that

$$Ser_i \sim \mathrm{PH}_2\left(\underline{\beta} = (1,0),\ B = \begin{bmatrix} -4 & 4 \\ 2 & -4 \end{bmatrix}\right).$$

We consider values of C_i from 2 to 12, $i = 1, 2, 3$, and calculate the steady-state distribution under FB as well as UWC versions 1 and 2, with and without the LIA approximation. The times to calculate the UWC probabilities as well as the steady-state distributions are recorded. Since UWC version 2 requires results from the FB model, the FB run times are included in the UWC version 2 run times. All calculations were done using the statistical software package R [27] on a computer with 16 GB of RAM and a 4.0 GHz Intel Core i7-6700K processor.

For the combination of UWC version 2 and LIA, the class-2 and class-3 UWC probabilities were computed only using steady-state probabilities from levels $0, 1, \ldots, C_1$, despite the fact that we can calculate the stationary distribution at higher levels using LIA, to be more directly comparable to UWC version 2 without LIA (i.e., what method is best at approximating the steady-state probabilities for levels $0, 1, \ldots, C_1$). It would be slightly more accurate to use steady-state probabilities from higher levels in these calculations as well, but as we normalize our probability vector $\underline{\beta}_i^*$ in Equation (4.2), the discrepancies should be small.

The IB model steady-state probabilities are approximated through the use of the FB method with $C_1 = C_2 = C_3 = 25$. Both measures from Equations (4.4) and (4.5) are calculated for each case. Note that the global maximum measure for the LIA cases do not consider states above level C_1. The class-1 marginal maximum measures for the LIA cases also do not look at states above level C_1, but levels up to 25 are used in the calculation of the class-2 and class-3 marginal maximum measures. Otherwise, the probabilities for those particular marginal distributions would be systematically underestimated.

In Figure 6, we plot the global maximum measures (denoted by M_g) and run times for each combination of approximation methods for $C_1 = 3, 6, 9, 12$ and $C_2 = C_3 = C$, $C = 2, 3, \ldots, 12$. We see that in Figure 6 (a), for $C_1 = 3$ which is less than the mean queue length that is approximately equal to 3.395, the performance of UWC version 2 (followed by version 1) is the best until $C = 5$, after which FB with LIA edges out a small advantage. It appears that for small values of C_1 for which LIA is inappropriate, there is nothing to be gained by combining LIA with UWC. As C_1 is increased, LIA simultaneously becomes more accurate and less impactful, so that the LIA and non-LIA M_g values converge for a given choice of FB or UWC version. In Figure 6 (g), we observe the relative ordering that we would expect, where models using UWC version 2 are slightly more accurate than those using UWC version 1 which is more accurate than FB.

In Figure 6 (b), (d), (f), and (h), we observe a consistent relative ordering of run times. Since the calculation of the UWC probabilities for version 1 is very simple, UWC version 1 and FB have near identical computational complexity and hence run times. In contrast, UWC version 2, as expected, takes significantly longer to run. For the LIA methods, the necessity of calculating R iteratively incurs a significant computational cost. With that in mind, we do see the same comparable run times for LIA with UWC version 1 and LIA with FB, while LIA with UWC version 2 takes the longest to calculate. In these plots, we incorporate a log-10 scale for the run times, so we can see that they increase exponentially in C for the non-UWC version 2 cases (which begin with a flatter relationship with C at

lower values, but observe a similar exponential relationship at higher C).

In Figure 7, we plot the marginal maximum measures for each class (denoted by $M_{m,i}$) at the same combinations of C_i. For $C_1 = 3$, the LIA methods prove the most effective at small values of C, but they get worse as C increases beyond 3 and are surpassed by UWC version 2 (and version 1) at $C = 7$. UWC version 2 and version 1 also eventually outperform the LIA methods for the other classes' $M_{m,i}$ values, but the FB and LIA method performs better until a larger value of C. Just as in Figure 6, the respective LIA and non-LIA methods converge as C_1 increases, while we once again are able to make the same conclusions with respect to FB and the UWC versions.

Overall, it seems that LIA on its own is more accurate than FB (other than when all the C_i are small). For the marginal maximum measures, LIA in combination with UWC may result in some small improvements, but it can come at a notable increase in global maximum measure values if C_1 is small. This is intuitive, since its approximation allows us to model more levels of the queue and hence the amount of probability mass at lower levels is smaller relative to when we must normalize over fewer levels. In addition, the approximation is less appropriate at small values of C_1 and we observe this through discrepancies in the exact probabilities assigned to individual states within these levels. However, as seen in Figure 6, increasing C_1 has a much smaller impact on run times than increasing C, so there should not be a problem in selecting a large enough value of C_1 so that LIA is reliable. Unfortunately, the use of LIA does incur a large computational cost when calculating the matrix R. Given that there is minimal to no benefit to its use when C_1 is large enough, UWC is a more time efficient option. Assuming that the values of C_i are not small for $i \neq 1$, it is advisable to use UWC version 1 as it is nearly equivalent to the FB method in computational cost, while UWC version 2 may strategically be used if one or more C_i's are small.

4.5 Sources of Approximation Error

Due to the necessity of maintaining the Markov property within our CTMCs, we are restricted to using a geometric approximation for the number of observed customer departures necessary to decrement the observed queue length. That is, we must assume that the events of observed departures reducing the observed queue length below its truncation level are iid, having the same success probability p^*_C. Since we have this single parameter to manipulate, we can use such an approximation to equate the expected time spent in that state (or level) in the UWC model of an elementary queue to the expected duration of a level-C busy period of the true IB queue. However, the geometric distribution is a simplification and not the true distribution in general. If we relaxed our restriction on not being able to introduce new states, it may be possible to model the departures using a phase-type distribution that can also match higher order moments, presumably improving approximation accuracy.

UWC version 1's approximation works by considering each service phase as its own $M/M/1+M$ queue. As discussed in Section 3.2.1, we can consider transitions from one service phase to another having a different absorption rate as removing all currently unobserved waiting customers and beginning a new busy period in a different $M/M/1+M$ queue. While

we lose fewer arriving customers relative to a FB model which completely blocks any arrivals while the queue is at full capacity, the corresponding UWC model would still typically reject some number of customers that otherwise would have received service in the IB model. This type of error is non-existent if service times are exponentially distributed, or minimal when a service time distribution involves few phase transitions. Within the exhaustive service policy, the queue length must be zero when the server switches away, and while the server is away, the queue length is treated as in a $M/M/\infty$ model, which (as we have seen in Section 2.3) can be modelled accurately using UWC. However, if the queue length happens to be at its truncation length when the server arrives, this departure of unobserved waiting customers would occur assuming that the initial service phase's absorption rate differs from that of the lead customer's reneging rate. Fortunately, this additional chance for approximation error may only occur once per cycle.

While UWC version 2 does not suffer as much from service phase transitions, its major restriction is having a single UWC probability for each service phase. When applied to a polling model, no distinctions are made between a server arriving and finding the queue at its truncation length versus the queue reaching this length during the server's visit (i.e., while they are actively serving that class of customers). If a server switches in to find a "full" queue, then the level-C busy period would likely initiate with multiple unobserved customers, in contrast to reaching that level during a visit which would begin with no unobserved customers. Therefore, not only would the distribution of the initial service phase for that level-C busy period be different, but the expected number of observed customer departures should be higher (since we begin with a longer queue). If we removed the restriction on increasing the number of states, then we could add duplicate states and use a separate UWC probability for each case, thereby eliminating this inaccuracy. Another inaccuracy unique to UWC version 2 is due to the approximation of IB model probabilities with FB model probabilities. While this approximation is known to be accurate in the $M/PH/1$ queue, it is likely less correct in more complex queues. If using service time distributions which perform well under UWC version 1 (i.e., in which UWC version 1 experiences minimal approximation errors), it may be more accurate to replace the unknown IB model probabilities in the calculation of $\underline{\hat{\beta}}^*_C$ with UWC version 1 model probabilities, rather than FB model probabilities. Alternatively, one could iteratively apply UWC version 2 and use its own approximations of the steady-state distribution in the next iteration's calculation of $\underline{\hat{\beta}}^*_C$, although that would further increase the difference in compuational cost between the UWC versions.

In Remark 2, we discussed a numerical example in which we observed issues with UWC version 2. In that example, we considered H_2 service times with extreme mixing weights. In such a distribution, it would be more accurate to use UWC version 1, since it would either spend a long time completing the current service in the rare slow service phase (and hence go a long time before switching to the common service phase), or experience multiple service completions in the common service phase with only a minimal risk of switching to the rare service phase after each completion. Therefore, we advise the use of UWC version 1 over version 2 when a service phase distribution is similar to an exponential distribution, since not only does UWC version 1 have a lower computational cost, it should never experience

unexpected approximation errors as UWC version 2 is prone to do.

5 Concluding Remarks

Within this paper, we have introduced a new method which can improve truncated queueing models' ability to approximate true IB systems. This is particularly useful for complex queues that require a very large number of states to model accurately, such as a polling system, in which it may not be computationally feasible to simply use sufficiently large truncation levels. Through the inclusion of a probability to not decrease an observed queue length from its finite buffer following an observed departure, we are able to emulate the presence of unobserved waiting customers who are ready to immediately enter the queue. Optimal choices for this probability have been derived for the $M/M/1$, $M/M/1+M$, $M/M/\infty$, and $M/PH/1$ queues, which result in exact steady-state probabilities for states below the buffer and the aggregation of excess steady-state probability mass from truncated states to the state(s) at the highest observed queue length.

Two versions of UWC have been presented for use within the $M/PH/1+M$ queue as well as a N-queue polling system with phase-type service and customers who may be impatient. While these versions make use of simplifying approximations and do not result in exact steady-state probabilities for non-buffer states, they do consistently outperform their equivalent finite buffer models in the absence of UWC. As UWC version 2 does require additional computations, version 1 may be more suitable given a sufficiently high reneging rate or truncation level, or if service time distributions are exponential (or similar). Several numerical examples have been presented to contrast their performance (both in terms of accuracy and computer run time) against each other as well as the FB model and the LIA method inspired by Shin and Choo [31]. A thorough discussion of potential sources of approximation error is included, and in what situations our two versions of UWC work best in providing accurate steady-state probability calculations.

For future work, we intend to investigate the generalization of UWC theory to queues with level-dependent reneging rates. Additionally, we will extend server behaviour within a polling model framework beyond that of exhaustive and consider UWC alongside the non-branching k-limited and Bernoulli service disciplines. It is also of interest to observe the impact of UWC on densities of customer waiting times and to investigate a version of UWC which relaxes the constraint of not increasing the number of states relative to its corresponding FB model (as discussed in Section 4.5).

References

[1] Altman, E., Yechiali, U.: Analysis of customers' impatience in queues with server vacations. Queueing Systems, 52(4), 261-279 (2006)

[2] Asmussen, S., Nerman, O., Olsson, M.: Fitting phase-type distributions via the EM algorithm. Scandinavian Journal of Statistics, 23(4), 419-441 (1996)

[3] Avrachenkov, K., Perel, E., Yechiali, U.: Finite-buffer polling systems with threshold-based switching policy. TOP, 24(3), 541-571 (2016)

[4] Baumann, H., Sandmann, W.: Numerical solution of level dependent quasi-birth-and-death processes. Procedia Computer Science, 1(1), 1561-1569 (2010)

[5] Baumann, H., Sandmann, W.: Computing stationary expectations in level-dependent QBD processes. Journal of Applied Probability, 50(1), 151-165 (2013)

[6] Boon, M.A.A.: Polling Models: From Theory to Traffic Intersections. Doctoral dissertation, Eindhoven: Technische Universiteit Eindhoven, 190 pages (2011)

[7] Boon, M.A.A., van der Mei, R.D., Winands, E.M.M.: Applications of polling systems. Surveys in Operations Research and Management Science, 16(2), 67-82 (2011)

[8] Boxma, O.J., de Waal, P.R.: Multiserver Queues with Impatient Customers. Teletraffic Science and Engineering, 1, 743-756 (1994)

[9] Bright, L., Taylor, P.G.: Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. Stochastic Models, 11(3), 497-525 (1995)

[10] Chakravarthy, S.R.: Maintenance of a deteriorating single server system with Markovian arrivals and random shocks. European Journal of Operational Research, 222(3), 508-522 (2012)

[11] Ciardo, G., Smirni, E.: ETAQA: an efficient technique for the analysis of QBD-processes by aggregation. Performance Evaluation, 36, 71-93 (1999)

[12] Diamond, J.E., Alfa, A.S.: Matrix analytic methods for a multi-server retrial queue with buffer. TOP, 7(2), 249-266 (1999)

[13] Drekic, S., Stanford, D.A., Woolford, D.G., McAlister, V.C.: A model for deceased-donor transplant queue waiting times. Queueing Systems, 79(1), 87-115 (2015)

[14] Gaver, D.P., Jacobs, P.A., Latouche, G.: Finite birth-and-death models in randomly changing environments. Advances in Applied Probability, 16(4), 715-731 (1984)

[15] Gertsbakh, I.: The shorter queue problem: A numerical study using the matrix-geometric solution. European Journal of Operational Research, 15(3), 374-381 (1984)

[16] Granville, K., Drekic, S.: On a 2-class polling model with reneging and k_i-limited service. Annals of Operations Research, 274(1), 267-290. (2018)

[17] Granville, K., Drekic, S.: A 2-class maintenance model with a finite population and competing exponential failure rates. Queueing Models and Service Management, 1(1), 141-176 (2018)

[18] Granville, K., Drekic, S.: A 2-class maintenance model with dynamic server behavior. TOP, 28, 34-96 (2020)

[19] He, Q.M.: Fundamentals of Matrix-Analytic Methods. Springer, New York (2014)

[20] Kim, J., Kim, B.: Waiting time distribution in an M/PH/1 retrial queue. Performance Evaluation, 70(4), 286-299 (2013)

[21] Krishnamoorthy, A., Babu, S., Narayanan, V.C.: The MAP/(PH/PH)/1 queue with self-generation of priorities and non-preemptive service. European Journal of Operational Research, 195(1), 174-185 (2009)

[22] Lakatos, L., Szeidl, L., Telek, M.: Introduction to Queueing Systems with Telecommunication Applications. Springer Science & Business Media, Berlin, Germany (2012)

[23] Levy, H., Sidi, M.: Polling systems: applications, modeling, and optimization. IEEE Transactions on Communications, 38(10), 1750-1760 (1990)

[24] Neuts, M.F.: Computational uses of the method of phases in the theory of queues. Computers & Mathematics with Applications, 1(2), 151-166 (1975)

[25] Neuts, M.F.: Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach. Dover Publications, Inc., New York (1981)

[26] Perel, E., Yechiali, U.: Two-queue polling systems with switching policy based on the queue that is not being served. Stochastic Models, 33(3), 1-21 (2017)

[27] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

[28] Riska, A., Smirni, E.: Exact aggregate solutions for M/G/1-type Markov processes. In ACM SIGMETRICS Performance Evaluation Review, 30(1), 86-96 (2002)

[29] Ross, S.M.: Introduction to Probability Models. Academic Press, San Diego (2014)

[30] Sakuma, Y., Takine, T.: Multi-class M/PH/1 queues with deterministic impatience times. Stochastic Models, 33(1), 1-29 (2017)

[31] Shin, Y.W., Choo, T.S.: M/M/s queue with impatient customers and retrials. Applied Mathematical Modelling, 33(6), 2596-2606 (2009)

[32] Takagi, H.: Queuing analysis of polling models. ACM Computing Surveys, 20(1), 5-28 (1988)

[33] Vishnevskii, V.M., Semenova, O.V.: Mathematical methods to study the polling systems. Automation and Remote Control, 67(2), 173-220 (2006)

[34] Yechiali, U.: Queues with system disasters and impatient customers when system is down. Queueing Systems, 56(3), 195-202 (2007)

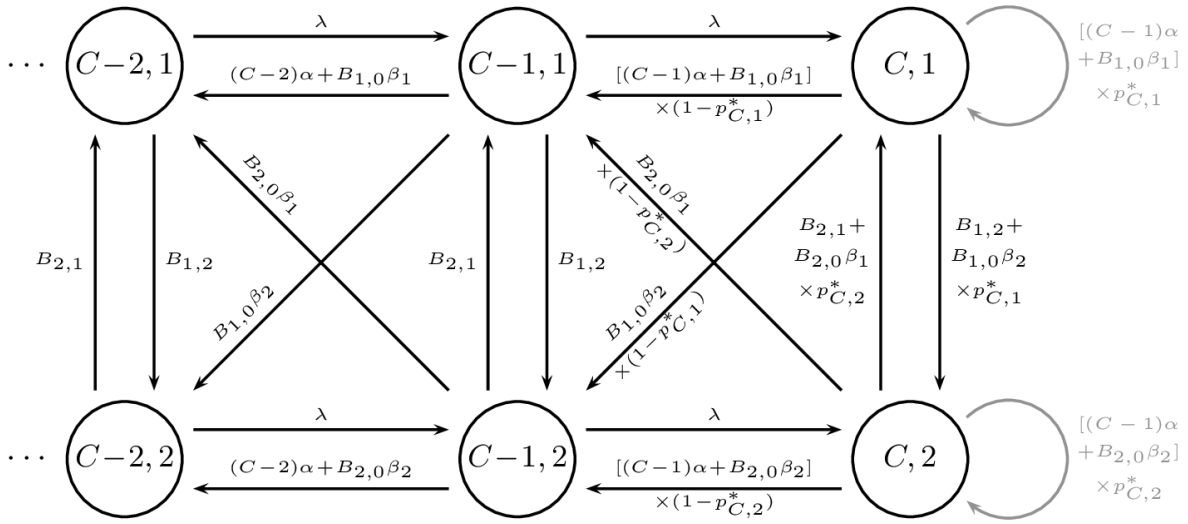

Figure 1: State transition diagram near truncation level C for a UWC model of a $M/PH/1+M$ queue with two service phases

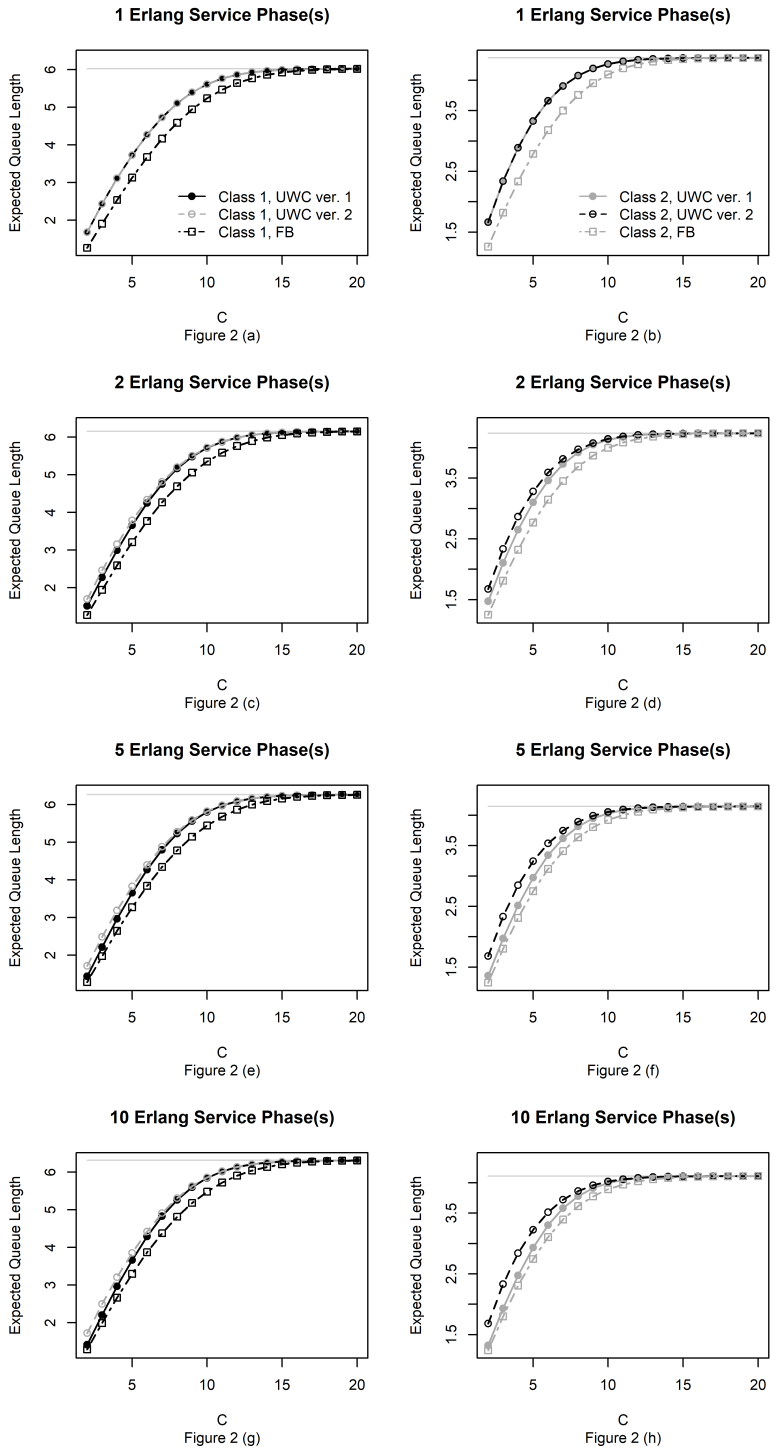

Figure 2: Plots of expected marginal queue lengths at steady state in a 2-queue system versus buffers $C_1 = C_2 = C$ for UWC version 1, UWC version 2, and FB models, under E_k service, $k = 1, 2, 5, 10$

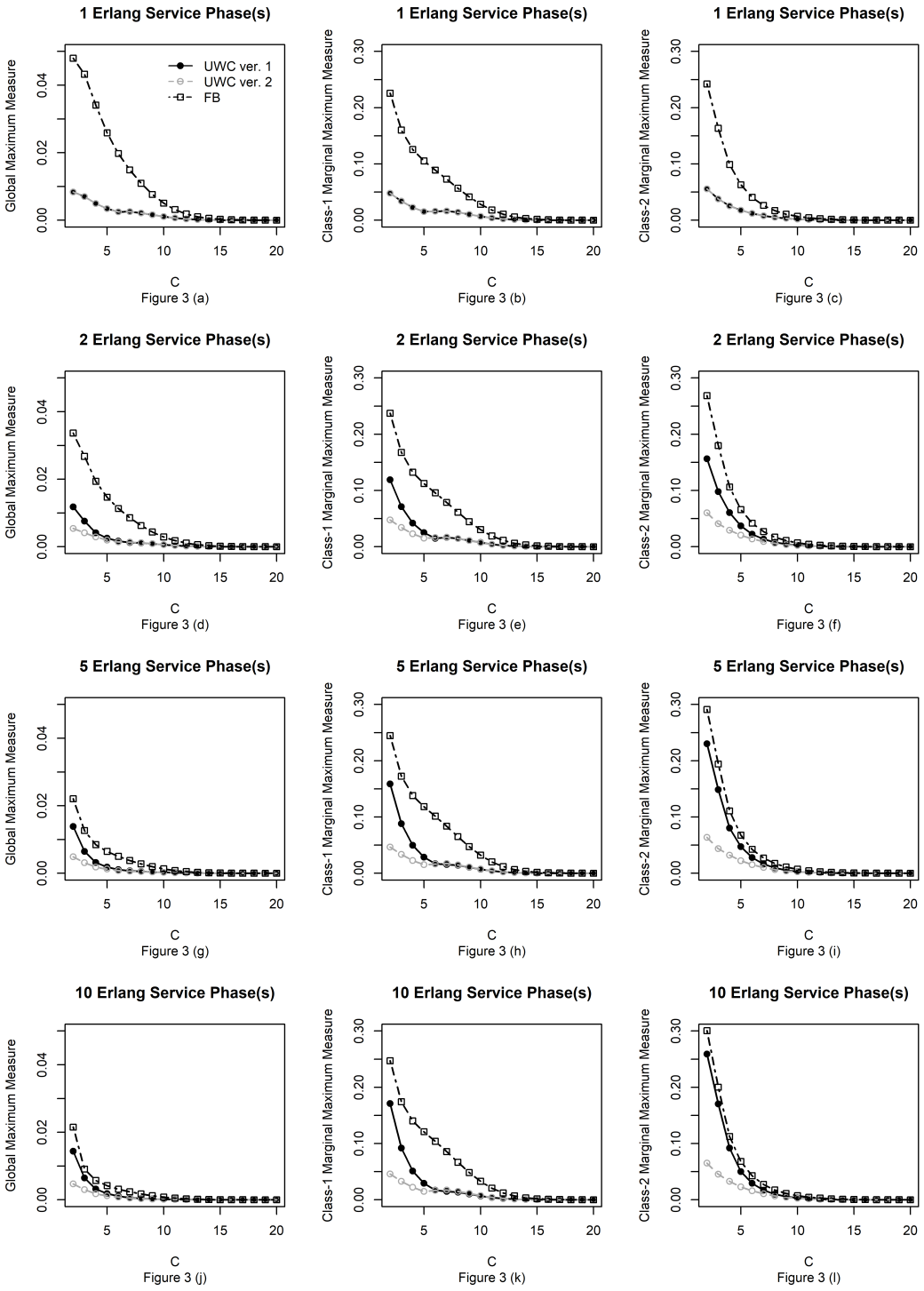

Figure 3: Plots of global and marginal maximum measures in a 2-queue system versus buffers $C_1 = C_2 = C$ for UWC version 1, UWC version 2, and FB models, under E_k service, $k = 1, 2, 5, 10$

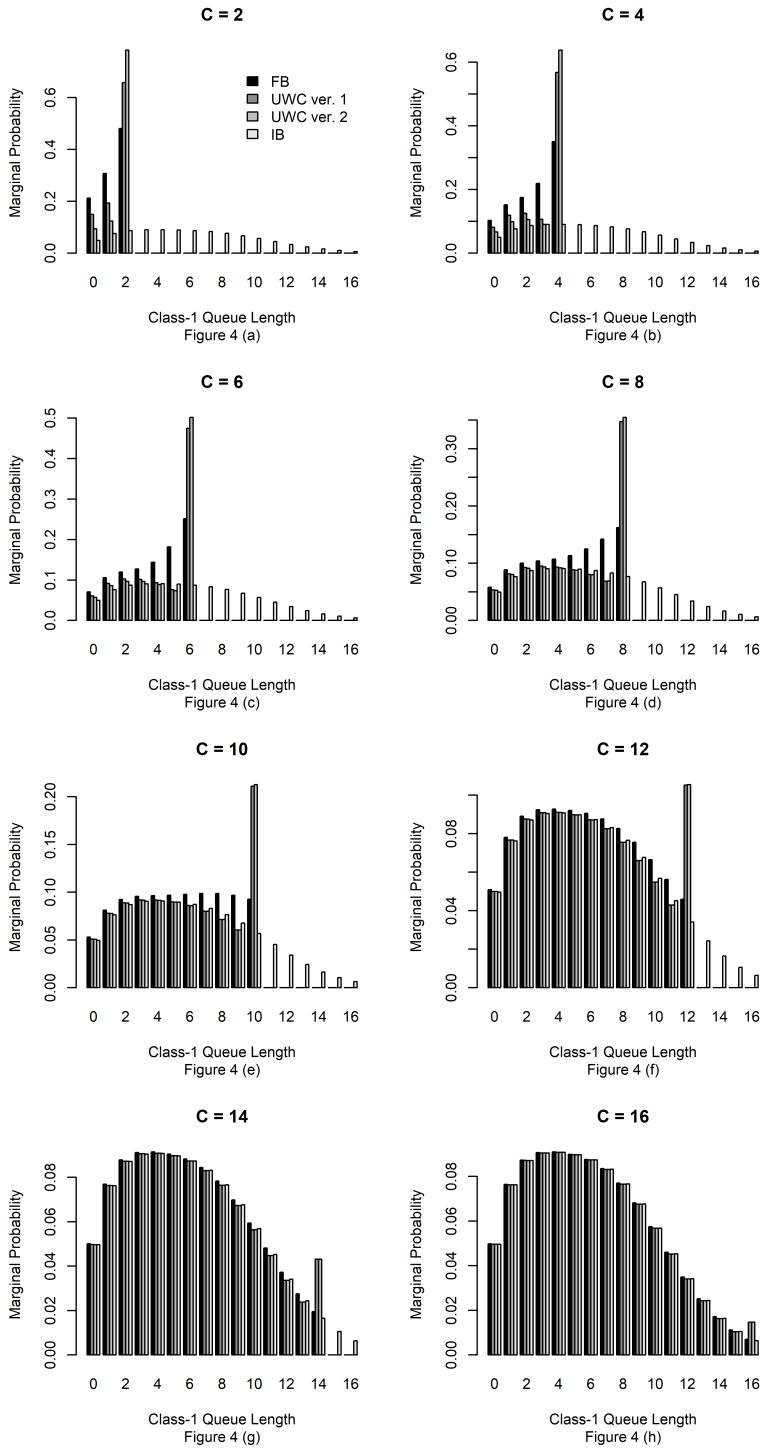

Figure 4: Barplots of class-1 marginal queue length probabilities at steady state in a 2-queue system versus buffers $C_1 = C_2 = C$ for UWC version 1, UWC version 2, FB, and IB models, under E_2^f service

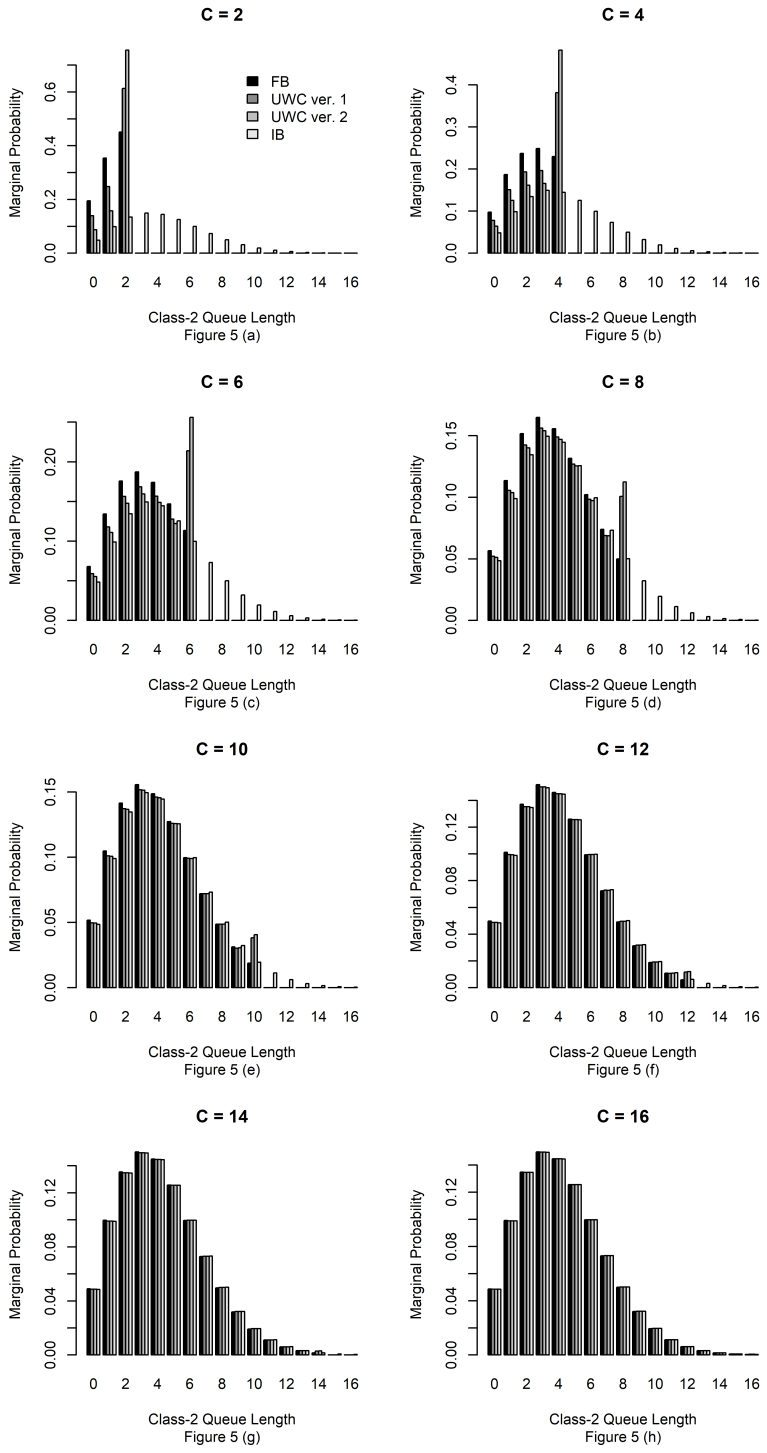

Figure 5: Barplots of class-2 marginal queue length probabilities at steady state in a 2-queue system versus buffers $C_1 = C_2 = C$ for UWC version 1, UWC version 2, FB, and IB models, under E_2^f service

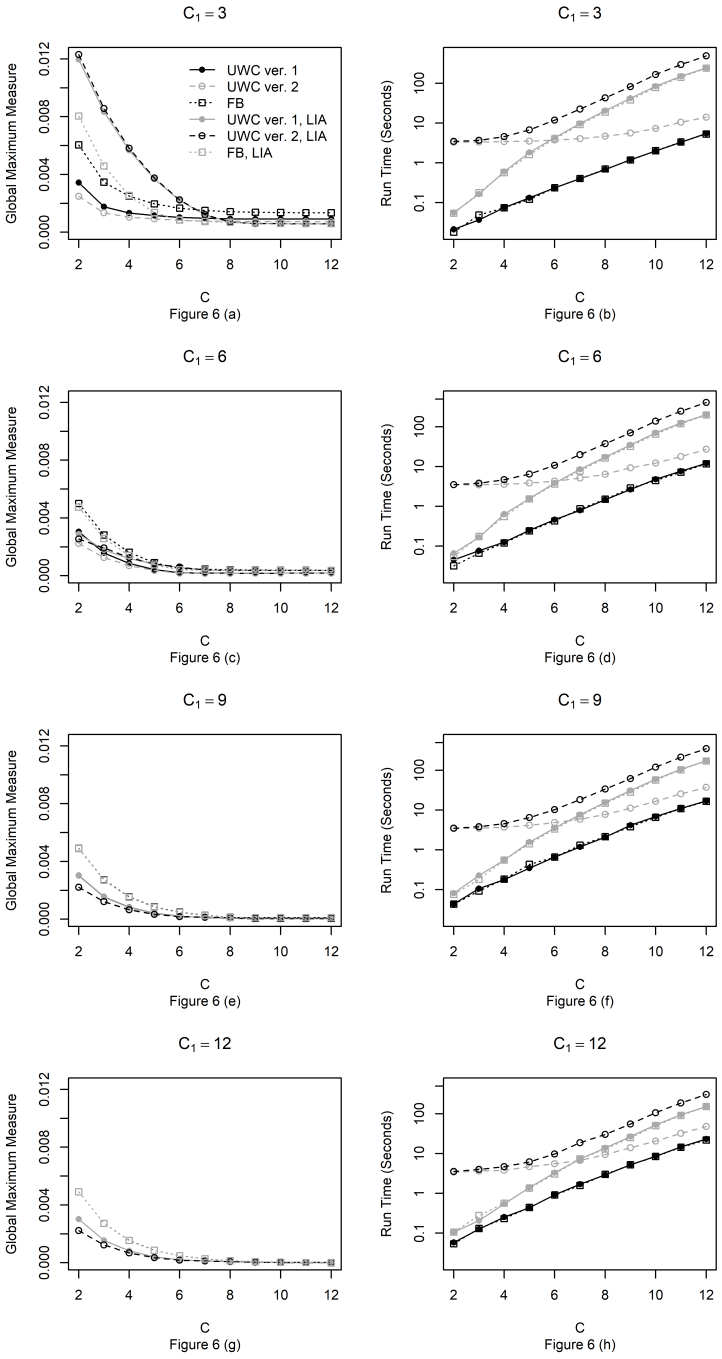

Figure 6: Plots of global maximum measure and run times to calculate UWC probabilities and steady-state distributions for a 3-queue system versus $C_1 = 3, 6, 9, 12$ and $C_2 = C_3 = C$ for UWC version 1, UWC version 2, and FB models with and without LIA.

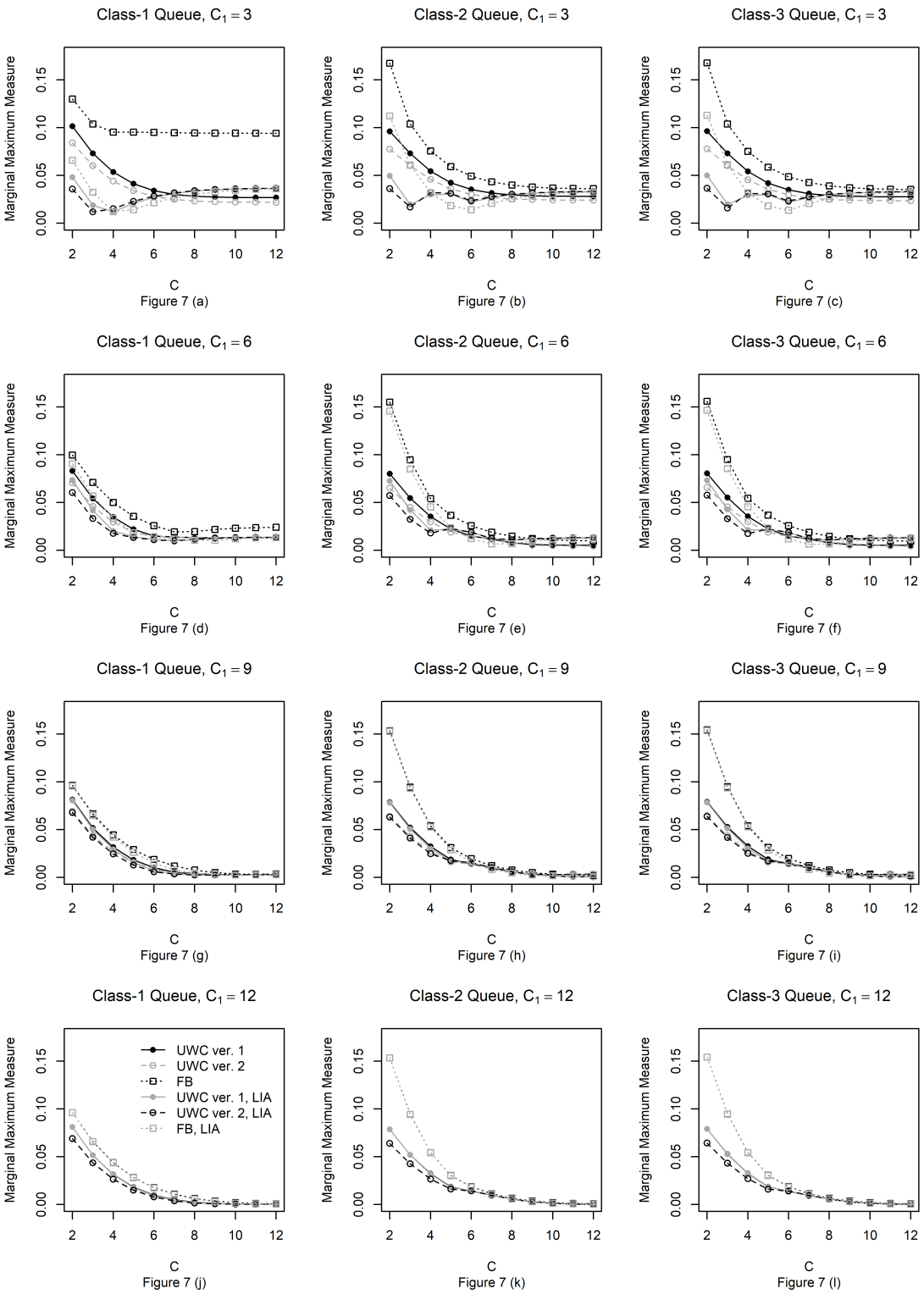

Figure 7: Plots of marginal maximum measures for a 3-queue system versus $C_1 = 3, 6, 9, 12$ and $C_2 = C_3 = C$ for UWC version 1, UWC version 2, and FB models with and without LIA.