A FAST MATRIX-ANALYTIC APPROXIMATION FOR THE TWO CLASS GI/G/1 NON-PREEMPTIVE PRIORITY QUEUE

Gábor Horváth
Department of Telecommunication
Budapest University of Technology and Economics
1521. Budapest Pf. 91., Hungary
E-mail: ghorvath@hit.bme.hu

KEYWORDS

Markov models, queueing systems, performance modeling, priority queue, multi class queue model

ABSTRACT

In this paper we present the approximate waiting time analysis of two class non-preemptive priority queues. The traffics of the queue are characterized by "two parameter description", which means that the mean and the squared coefficient of variation of the inter arrival times and of the service times are given. The solution is based on the separate analysis of the low and high priority queue. The resulting single class queues have a homogeneous QBD (quasi birth death) structure, therefore their analysis is numerically efficient. We check the performance of the approximation extensively, and conclude that it gives good accuracy in a wide range of traffic parameters.

1 INTRODUCTION

In several computer and telecommunication systems jobs can be grouped to job classes. Jobs belonging to different classes can have different behavior, e.g. different arrival process, or different service requirement.

The non-preemptive priority scheduling defines strict priorities: always the highest priority job is selected by the server, but there is no job-preemption if an even higher priority job arrives. Such priority based solutions often appear in computer systems. For example, the task scheduler of an operating system may distinguish between kernel and user tasks, and give higher priority to the kernel tasks, since for them it is important to finish faster. The signaling channels in telecommunication networks have often higher priority as well, not to allow the other traffics to grab the whole link capacity.

From the stochastic analysis point of view, modeling of multi class systems is more challenging than modeling of single class systems. Even if we consider the simplest case – exponentially distributed inter arrival times and job sizes – the arisen model is a multi dimensional Markov chain, which usually does not have a closed form steady state solution. Nevertheless, the engineering applications need more general arrival and job size characterization, the exponential distribution is usually not a good model for the real behavior.

There are a number of papers and books which study priority queueing systems. The initial works are [4],[1],[7] and [8]. In all of these works the arrival process is assumed to be a Poisson process. In [2] a simple formula for the mean waiting time of M/G/1 type non-preemptive priority queues is given. But the arrival process is still a Poisson process in that paper, and the mean waiting time is the only performance measure that can be computed that way. In [9] the full analysis of the MAP/G/1 non-preemptive priority queue is given, including the distribution and the moments of the waiting time. The results of that paper are rather theoretically than application ready, since there are still some unsolved questions related to the numerical solution.

Contrary to the above mentioned works, we present an approximate analysis in this paper. Our goal was to develop a fast approximation to compute performance measures of priority queues with traffic characterized by a two parameter description.

Section 2 introduces the idea of the approximation. Section 4 presents the analysis of the low, Section 5 the analysis of the high priority queue in details. The solution and performance measures of QBDs are summarized in Section 6. To verify the accuracy of the method, we compared our results to simulation results in Section 7. Finally, Section 8 concludes the paper.

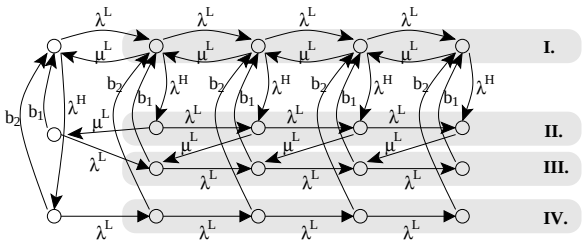
2 CONCEPT OF THE APPROXIMATION

In this paper we limit our attention to the two class case. However, the concept itself can be easily extended to arbitrary number of classes.

The concept is to approximate the two class system as the classes were separated, and construct a service process for both classes that approximately imitates the behavior of the original server.

Proceedings 12th International Conference ASMTA 2005
Khalid Al-Begain, Gunter Bolch, Miklos Telek © ECMS, 2005
ISBN 1-84233-112-4 (Set) / ISBN 1-84233-113-2 (CD)

Let consider the low priority queue first. From the point of view of the low priority customer class, the exact number of high priority customers does not play any role. When there are no high priority customers, the server is available, and when there are high priority customers, the server is not available for low priority service. Therefore, during the analysis of the low priority queue, the two dimensionally infinity state space is eliminated such, that the number of high priority customers is modeled by only 2 states: zero, and more than zero. The transitions between these states are obtained by the busy period analysis of the high priority queue. Figure 1 shows the structure of the approximating Markov chain. The detailed description and analysis can be found in Section 4.



I. No high priority customer

II. There are high pr. customers (blocked by low pr) low pr. customer in the server

III. There are high pr customers (not blocked any more) high pr. customer in the server

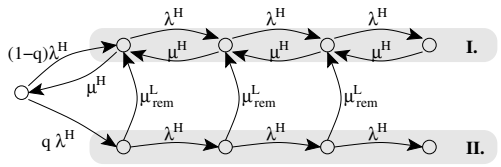
IV. There are high pr. customers high pr. customer in the server

Figure 1: Queue model of the low priority class

On the other hand, the high priority customers can be affected by the low priority customers only at one point: when the high priority queue is empty at the arrival of a high priority customer, and a low priority customer is in the server. In this case the arrived high priority customer has to await the remaining service time of the low priority customer, since the service is non-preemptive. The probability of this event (q) will be computed from the queue model of the low priority class. Figure 2 shows the structure of the corresponding Markov chain. The detailed description and analysis can be found in Section 5.

3 APH-2 CHARACTERIZATION OF THE ARRIVAL AND SERVICE TIMES

In our framework the jobs are characterized by a two parameter description. The arrival process is described by the arrival rate ($\lambda^{(H)}$ for the high, $\lambda^{(L)}$ for the low priority class), and by the squared coefficient of variation of the inter arrival times ($c_A^{2\,(H)}$ and $c_A^{2\,(L)}$). The service time is characterized by the



I. No low pr. customer at first arrival

II. There was a low pr. customer at first arrival (high pr. customer is blocked)

Figure 2: Queue model of the high priority class

mean ($\tau^{(H)}$, $\tau^{(L)}$) and by the squared coefficient of variation ($c_S^{2\,(H)}$, $c_S^{2\,(L)}$).

To exploit the quasi birth death (QBD) structure of the queue models (check Figure 1 and 2), we use matrix analytical techniques for the solution. Therefore we convert the two parameter description of the inter arrival and service times to phase type (PH) distributions. For this reason a small acyclic PH distribution having only 2 states (APH-2) will be constructed, which has the same mean and squared coefficient of variation as given. (It can be applied only if the squared coefficient of variation is ≥ 0.5).

The transient generator of the PH distribution describing the inter arrival times of high priority customers is denoted by $\boldsymbol{D_0}^{(H)}$, the rate vector to the absorbing state is denoted by $\underline{d}^{(H)}$ (column vector), and the initial probability vector is denoted by $\underline{\delta}^{(H)}$ (row vector). They are constructed the following way:

$$\boldsymbol{D_0}^{(H)} = \begin{bmatrix} -\lambda^{(H)}/c_A^{2\,(H)} & \lambda^{(H)}/c_A^{2\,(H)} \\ 0 & -2\,\lambda^{(H)} \end{bmatrix}, \underline{d}^{(H)} = \begin{bmatrix} 0 \\ 2\lambda^{(H)} \end{bmatrix},$$
$$\underline{\delta}^{(H)} = \begin{bmatrix} 1/(2\,c_A^{2\,(H)}) & 1 - 1/(2\,c_A^{2\,(H)}) \end{bmatrix}. \tag{1}$$

It is easy to check that the mean of this PH distribution is $1/\lambda^{(H)}$, and its squared coefficient of variation is $c_A^{2\,(H)}$. The PH fitting for the low priority queue is performed similarly. (The corresponding quantities of the low priority class are indicated by superscript $^{(L)}$ instead of $^{(H)}$).

The distribution of the service time is a phase type distribution with transient generator denoted by $\boldsymbol{S}^{(H)}$, rate vector to the absorbing state denoted by $\underline{s}^{(H)}$ (column vector), and initial probability vector denoted by $\underline{\sigma}^{(H)}$ (row vector). The construction of these parameters is as follows:

$$\boldsymbol{S}^{(H)} = \begin{bmatrix} -1/(\tau^{(H)} c_S^{2\,(H)}) & 1/(\tau^{(H)} c_S^{2\,(H)}) \\ 0 & -2/\tau^{(H)} \end{bmatrix},$$
$$\underline{s}^{(H)} = \begin{bmatrix} 0 \\ 2/\tau^{(H)} \end{bmatrix}, \ \underline{\sigma}^{(H)} = \begin{bmatrix} 1/(2c_S^{2\,(H)}) & 1 - 1/(2c_S^{2\,(H)}) \end{bmatrix}.$$

4 THE LOW PRIORITY QUEUE

The structure of the Markov chain that models the low priority queue is depicted on Figure 1. One circle on that figure does not correspond to only one state, it indicates a kind on "macro-state" (the phases of the PH arrival and service time are hidden in favor of readability). The levels in that Markov chain correspond to the number of low priority customers in the system. This quasi birth-death structure leads to a well-known block-tridiagonal generator matrix:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{B_0} & \boldsymbol{A_0} & & & & \\ \boldsymbol{C_1} & \boldsymbol{B} & \boldsymbol{A} & & & \\ & \boldsymbol{C} & \boldsymbol{B} & \boldsymbol{A} & & \\ & & \boldsymbol{C} & \boldsymbol{B} & \boldsymbol{A} & \\ & & & \ddots & \ddots & \ddots \end{bmatrix} \tag{2}$$

At levels > 0 we distinguish between 4 state groups, to follow the presence and the state of the high priority class.

In state group "I." there are no high priority customers in the system, the server serves low priority customers. In this state group each "macro-state" consists of the following phases:

- phase of arrival process of the low pr. class
- phase of service time of the low pr. class
- phase of arrival process of the high pr. class

When a high priority customer arrives in state group "I.", a transition occurs to state group "II.". In state group "II." there is a high priority customer in the system, but it has to wait until the current service finishes. In this state group still the low priority customer is in the server. The states in this state group consist of the following phases:

- phase of arrival process of the low pr. class
- phase of service time of the low pr. class

When the service of the low priority customer is finished in state group "II.", a transition occurs to state group "III.". In this group the server works on the high priority class. When all the high priority customers are served (the busy period is finished), the state group "I." will follow again. We compute 2 moments of the busy period of the high priority customers, and approximate the duration of the busy period by an APH-2 distribution. Therefore in state group "III." we have to keep track the following phases:

- phase of arrival process of the low pr. class
- phase of APH-2 distribution that approximates the busy period of the high priority queue. Several high pr. customers can be in the queue when the server starts to work on high pr. customers, because several arrivals can occur while being blocked in state group "II.". This kind of busy period will be called a "Type 1" busy period (random variable denoted by $\mathcal{B}$), and the parameters of its APH-2 approximate are denoted by: $\boldsymbol{B}, \underline{b}, \underline{\beta}$ (transient generator, transition rates to absorbing states and initial probability vector, respectively). (The "Type 1" busy period is indicated by label b_1 on Figure 1).

The length of the "Type 1" busy period depends a lot on the number of high pr. arrivals until the server finishes the service of the low pr. customer. The number of arrivals is directly related to the phase of service time of the low pr. customer at the instant of the arrival of the first high pr. customer. Thus, we refine our queue model by distinguishing between "Type 1" busy periods with low priority service being in phase 1 ($\boldsymbol{B_1}, \underline{b_1}, \underline{\beta_1}$) and being in phase 2 ($\boldsymbol{B_2}, \underline{b_2}, \underline{\beta_2}$) at the instant of the first high pr. arrival.

If there are no low pr. customers in the system (at level 0), and a high pr. customer arrives, the server starts its service immediately. This is a different kind of busy period, because it starts from exactly 1 high pr. customer (the one that initiated the busy period). This will be called a "Type 2" busy period (random variable denoted by $\hat{\mathcal{B}}$). The "Type 2" busy period is indicated by label b_2 on Figure 1, and the parameters of the APH-2 fitted on its first two moments are denoted by: $\hat{\boldsymbol{B}}, \hat{\underline{b}}, \hat{\underline{\beta}}$. When a "Type 2" busy period starts (with a high pr. arrival at level 0), a transition occurs to state group "IV." in the Markov chain. This state group consists of the following phases:

- phase of arrival process of the low pr. class
- phase of APH-2 distribution that approximates the duration of the "Type 2" busy period of the high pr. queue.

When the "Type 1" or "Type 2" busy period is over, the state group "I." will be active again, since all the high pr. customers left the system.

Based on the above explanation, we can construct the matrices of generator matrix Q: Matrices $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ (see eq. (3)) consist of 6 blocks: the first block corresponds to state group "I.", the second and third correspond to state group "II." (according to the refinement described above), the forth and fifth correspond to state group "III." (duplicated due to the refinement again), finally the sixth block corresponds to state group "IV.". $\boldsymbol{I}$ denotes the identity matrix, and the $\underline{\alpha}$ (row-) vectors denote the probability vector of the phase of the arrival process of the high priority customers right after the busy period is finished.

The construction of matrices in the irregular level 0 are shown by eq. (4).

Due to the lack of space, we have to omit the computation of the duration of Type "1" and Type "2"

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} \otimes \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{D_0}^{(L)} \oplus \boldsymbol{D_0}^{(H)} \oplus \boldsymbol{S}^{(L)} & \boldsymbol{I} \otimes \underline{d}^{(H)} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} & \boldsymbol{I} \otimes \underline{d}^{(H)} \otimes \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D_0}^{(L)} \oplus \boldsymbol{S}^{(L)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_0}^{(L)} \oplus \boldsymbol{S}^{(L)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I} \otimes \underline{b_1} \cdot \underline{\alpha_1} \otimes \delta^{(L)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_0}^{(L)} \oplus \boldsymbol{B_1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I} \otimes \underline{b_2} \cdot \underline{\alpha_2} \otimes \delta^{(L)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_0}^{(L)} \oplus \boldsymbol{B_2} & \boldsymbol{0} \\ \boldsymbol{I} \otimes \hat{\underline{b}} \cdot \hat{\underline{\alpha}} \otimes \underline{\delta}^{(L)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_0}^{(L)} \oplus \hat{\boldsymbol{B}} \end{bmatrix} \quad (3)$$

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I} \otimes \boldsymbol{I} \otimes \underline{s}^{(L)} \cdot \underline{\delta}^{(L)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \otimes \underline{s}^{(L)} \cdot \underline{\beta_1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \otimes \underline{s}^{(L)} \cdot \underline{\beta_2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

$$\boldsymbol{A_0} = \begin{bmatrix} \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} \otimes \underline{\delta}^{(L)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_1}^{(L)} \otimes \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{B_0} = \begin{bmatrix} \boldsymbol{D_0}^{(L)} \oplus \boldsymbol{D_0}^{(H)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \otimes \underline{d}^{(H)} \cdot \underline{\hat{\beta}} \\ \boldsymbol{I} \otimes \underline{b_1} \cdot \underline{\alpha_1} & \boldsymbol{D_0}^{(L)} \oplus \boldsymbol{B_1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I} \otimes \underline{b_2} \cdot \underline{\alpha_2} & \boldsymbol{0} & \boldsymbol{D_0}^{(L)} \oplus \boldsymbol{B_2} & \boldsymbol{0} \\ \boldsymbol{I} \otimes \hat{\underline{b}} \cdot \underline{\hat{\alpha}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D_0}^{(L)} \oplus \hat{\boldsymbol{B}} \end{bmatrix} \quad (4)$$

$$\boldsymbol{C_1} = \begin{bmatrix} \boldsymbol{I} \otimes \boldsymbol{I} \otimes \underline{s}^{(L)} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \otimes \underline{s}^{(L)} \cdot \underline{\beta_1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \otimes \underline{s}^{(L)} \cdot \underline{\beta_2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

busy periods. The idea is to compute two moments of these durations (by expressing them in Laplace domain, and taking their derivative), and the construction of the corresponding APH-2 distribution is performed as in eq. (1).

5 THE HIGH PRIORITY QUEUE

Figure 2 depicts the structure of the Markov chain that models the high priority queue.

When the high priority queue is empty, and a high priority customer arrives, there are two possibilities.

With probability q, the arrived customer finds a low priority customer in the server, whose service can not be interrupted. Therefore it has to await its remaining service time. During this time, other high priority customers can arrive. This case is covered by state group "II." on the figure.

On the other hand, with probability $1-q$ the server is empty, and the service of the arrived customer can start immediately. This case corresponds to state group "I." on the figure.

According to this explanation, the matrix blocks of the generator matrix can be constructed the following way:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{D_1}^{(H)} \otimes \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D_1}^{(H)} \otimes \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{D_0}^{(H)} \oplus \boldsymbol{S}^{(H)} & \boldsymbol{0} \\ \boldsymbol{I} \otimes \underline{s}^{(L)} \cdot \underline{\sigma}^{(H)} & \boldsymbol{D_0}^{(H)} \oplus \boldsymbol{S}^{(L)} \end{bmatrix}$$

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I} \otimes \underline{s}^{(H)} \cdot \underline{\sigma}^{(H)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

The matrices in level 0 are:

$$\boldsymbol{A_0} = \begin{bmatrix} (1-q) \cdot \boldsymbol{D_1}^{(H)} \otimes \underline{\sigma}^{(H)} & q \cdot \boldsymbol{D_1}^{(H)} \otimes \underline{\sigma}_r^{(L)} \end{bmatrix}$$

$$\boldsymbol{B_0} = \begin{bmatrix} \boldsymbol{D_0}^{(H)} \end{bmatrix}$$

$$\boldsymbol{C_1} = \begin{bmatrix} \boldsymbol{I} \otimes \underline{s}^{(H)} \\ \boldsymbol{0} \end{bmatrix}$$

Two quantities are missing to accomplish the analysis: the above mentioned q probability, and the phase distribution of the service PH distribution of the low priority class embedded at the arrival instant of a high probability customer when it arrived into an empty high priority queue, denoted by $\underline{\sigma}_r^{(L)}$. These missing quantities are derived from the low priority queue model, from its steady state distribution $\underline{v}_k$.

The probability that there are k customers in the low priority queue when a high priority customer arrives in the empty high priority queue is denoted by $\underline{p}_k$. The ith element of this vector corresponds to the ith phase of the service distribution of the low priority customer.

$$\underline{p}_k = \frac{\underline{v}_k \cdot \boldsymbol{M} \cdot (\underline{h} \otimes \underline{d}^{(H)} \otimes \boldsymbol{I})}{\sum_{i=1}^{\infty} \underline{v}_i \cdot \boldsymbol{M} \cdot (\underline{h} \otimes \underline{d}^{(H)} \otimes \underline{h}) + \underline{v}_0 \cdot \boldsymbol{M'} \cdot (\underline{h} \otimes \underline{d}^{(H)})},$$

$$p_0 = \frac{\underline{v}_0 \cdot \boldsymbol{M'} \cdot (\underline{h} \otimes \underline{d}^{(H)})}{\sum_{i=1}^{\infty} \underline{v}_i \cdot \boldsymbol{M} \cdot (\underline{h} \otimes \underline{d}^{(H)} \otimes \underline{h}) + \underline{v}_0 \cdot \boldsymbol{M'} \cdot (\underline{h} \otimes \underline{d}^{(H)})}.$$

Here the purpose of $\boldsymbol{M}$ and $\boldsymbol{M'}$ are to select the states of state group "I." of the low priority queue, which corresponds to the empty high priority queue. They are defined by:

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{I}_{8\times 8} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \qquad \boldsymbol{M'} = \begin{bmatrix} \boldsymbol{I}_{4\times 4} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}.$$

Using the introduced embedded $\underline{p}_k$ distribution, the q probability equals $1 - p_0$:

$$q = 1 - p_0.$$

To obtain the $\underline{\sigma}_r^{(L)}$ phase distribution, we have to sum up $\underline{p}_k$ from 1 to ∞:

$$\underline{\sigma}_r^{(L)} = \sum_{i=1}^{\infty} \frac{\underline{p}_i}{1 - p_0} = \frac{\underline{v}_1 \cdot (\boldsymbol{I} - \boldsymbol{R})^{-1} \cdot \boldsymbol{M} \cdot (\underline{h} \otimes \underline{d}^{(H)} \otimes \boldsymbol{I})}{\underline{v}_1 \cdot (\boldsymbol{I} - \boldsymbol{R})^{-1} \cdot \boldsymbol{M} \cdot (\underline{h} \otimes \underline{d}^{(H)} \otimes \underline{h})}.$$

Now all the blocks of the generator matrix are defined. The next section gives a short summary on the steady state analysis and waiting time moments of QBDs.

6 PERFORMANCE MEASURES

In this section we summarize how to compute the steady state distribution and waiting time moments of infinite homogenous QBDs (for more details see [3]).

6.1 Steady State Distribution

The steady state probabilities of block tri-diagonal structured Markov chains (see eq. (2)) are obtained from the well known formula $\underline{0} = \underline{\pi}\boldsymbol{Q}$, where $\underline{\pi}$ consists of vectors $\underline{v}_k$: $\underline{\pi} = [\underline{v}_0, \underline{v}_1, \dots]$.

In [5] it is proven that the solution of the Markov chain has the following matrix geometric form:

$$\begin{aligned} \underline{v}_0 &= \underline{v}_1 \boldsymbol{C_1} (-\boldsymbol{B_0})^{-1} \\ \underline{v}_k &= \underline{v}_1 \boldsymbol{R}^{k-1} \qquad\qquad k = 2, 3, \dots, \end{aligned}$$

where $\boldsymbol{R}$ satisfies the following matrix-quadratic equation:

$$\boldsymbol{0} = \boldsymbol{A} + \boldsymbol{R}\boldsymbol{B} + \boldsymbol{R}^2\boldsymbol{C}.$$

There are many numerical methods to compute matrix $\boldsymbol{R}$, [3] provides a good collection of them.

For several performance measures (e.g. for the waiting time distribution) the state probabilities embedded at arrival epochs (denoted by $\underline{v}_k^A$) are required instead of the general time probabilities. They are obtained via multiplying the general time steady state probability vectors by a weight vector of the probability that the arrival occurs in the given phase:

$$\underline{v}_0^A = \underline{v}_0 \frac{\boldsymbol{A_0}}{\underline{v}_0 \boldsymbol{A_0} \underline{h} + \sum_{k=1}^{\infty} \underline{v}_k \boldsymbol{A} \underline{h}} = \frac{1}{\lambda} \underline{v}_0 \boldsymbol{A_0}$$

$$\underline{v}_k^A = \underline{v}_k \frac{\boldsymbol{A}}{\underline{v}_0 \boldsymbol{A_0} \underline{h} + \sum_{j=1}^{\infty} \underline{v}_j \boldsymbol{A} \underline{h}} = \frac{1}{\lambda} \underline{v}_1 \boldsymbol{R}^{k-1} \boldsymbol{A} \quad k > 0.$$

(Here $\lambda = \underline{\gamma} \boldsymbol{A} \underline{h}$, where $\underline{\gamma}$ is the steady state phase distribution (it satisfies $\underline{0} = \underline{\gamma}(\boldsymbol{A} + \boldsymbol{B} + \boldsymbol{C})$)).

6.2 Moments of the Waiting Time

To compute the moments of the waiting time, we first express the distribution of the waiting time in Laplace domain.

When a customer arrives into the queue, the queue length distribution it experiences is $\underline{v}_k^A$, thus, with probability $\underline{v}_k^A$ it has to wait $k+1$ services (including its own) to leave the system. The Laplace transform of one service in a QBD equals $(s\boldsymbol{I} - \boldsymbol{B} - \boldsymbol{A})^{-1}\boldsymbol{C}$, so the distribution of the waiting time can be expressed as:

$$\begin{aligned} T(s) &= \sum_{k=0}^{\infty} \underline{v}_k^A \left[(s\boldsymbol{I} - \boldsymbol{B} - \boldsymbol{A})^{-1} \boldsymbol{C} \right]^{k+1} \cdot \underline{h} = \\ &= \underline{v}_0^A (s\boldsymbol{I} - \boldsymbol{B} - \boldsymbol{A})^{-1} \boldsymbol{C} \underline{h} + \\ &\quad + \underline{v}_1^A \sum_{k=1}^{\infty} \boldsymbol{R}^{k-1} \left[(s\boldsymbol{I} - \boldsymbol{B} - \boldsymbol{A})^{-1} \boldsymbol{C} \right]^{k+1} \cdot \underline{h}. \end{aligned}$$

The moments of the waiting time are computed by taking the derivative of $T(s)$ at $s = 0$. The infinite sum with respect to k can be eliminated by replacing $\boldsymbol{R}^k$ by its spectral decomposition.

7 NUMERICAL RESULTS

The presented computation method has been implemented in MATLAB. Since QBDs can be solved efficiently with such a small number of phases we have,

this method is very fast. Generating the results of one plot (15 executions) took about 3 seconds on a modern PC (Intel Pentium IV 2.4 GHz). A C implementation would be even faster.

In case of all approximation methods, the verification has a special importance. We checked the influence of all the parameters on the accuracy of the results (compared to the simulation results). The following parameters have been used: $\lambda^{(H)} = 0.4, c_A^{2\,(H)} = 1, \tau^{(H)} = 1, c_S^{2\,(H)} = 1, \lambda^{(L)} = 0.2, c_A^{2\,(L)} = 1, \tau^{(L)} = 1, c_S^{2\,(L)} = 1$. Some of the plots are depicted on Figure 3.

Looking at the figures we can say that the mean waiting time is approximated very accurately in most cases. The results of the high priority queue are more accurate, in case of the low priority queue the error grows up to 6-7 percent when $c_A^{2\,(H)}$ or $c_S^{2\,(L)}$ is large.

The presented method gives a tight approximation for the squared coefficient of variation of the waiting time, too. The results of the high priority queue are more accurate again. The worst matching in the results is at $c_S^{2\,(L)} = 5$, where the error is 12 percent, but in most cases the difference is less than 5 percent.

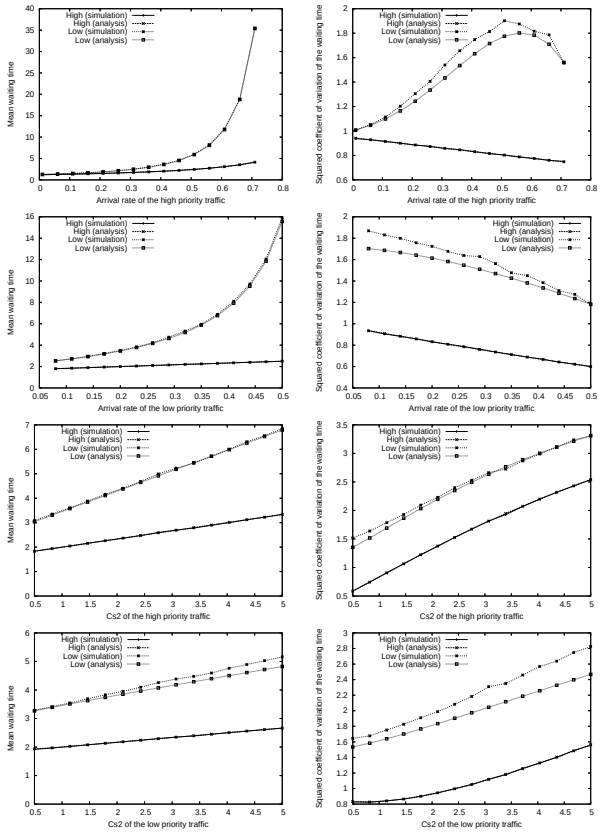


Figure 3: Results

8 CONCLUSION

We presented an approximate analysis method for two class non-preemptive priority systems with traffic characterized by two parameter description. Using Matrix analytical techniques, we described the construction and solution of the QBDs that model the behavior of the low and high priority queues.

To check the accuracy of the approximation, the results of our implementation has been compared to simulation results. We found that the presented computation method provides tight approximation for the waiting time moments in less than a second. This makes it possible to use it for tasks where a queueing analysis is performed many times iteratively, such as in case of a dimensioning problem, or in case of a traffic based queueing network analysis.

References

[1] N.K. Jaiswal. *Priority Queues.* Academic Press, New York, 1968.

[2] L. Kleinrock. *Queueing Systems Volume II: Computer Applications.* John Wiley & Sons, Inc., 1976.

[3] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* American Statistical Association and the Society for Industrial and Applied Mathematics, 1999.

[4] R.G. Miller. Priority Queues. *Ann. Math. Statist.*, 31:86–103.

[5] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models.* The Johns Hopkins University Press, 1981.

[6] Rózsa Pál. *Lineáris algebra és alkalmazásai.* Tankönyvkiadó, Budapest, 1991.

[7] L. Takács. Priority Queues. *Operation Research*, 12:63–74, 1964.

[8] H Takagi. *Queueing Analysis - A Foundation of Performance Evaluation: Volume 1. Vacation and Priority Systems, Part 1.* North-Holland, Amsterdam, 1991.

[9] Tetsuya Takine. The Nonpreemptive Priority MAP/G/1 Queue. *Operation Research*, 47(6):917–927, 1999.

Gábor Horváth received the M. Sc. degree in Computer Science in 2001 from the Budapest University of Technology and Economics. He was a PhD student from 2001 to 2004 supervised by Miklós Telek at the same university at the Department of Telecommunications, where he is an assistant professor now. His research interest includes performance modeling of telecommunication systems.