Blocking Probabilities in Multi-Service Systems with Preemptive Scheduling

Shuna Yang, Norvald Stol
Department of Telematics
Norwegian University of Science and Technology, NTNU
Trondheim, Norway
email: {shuna, norvald.stol}@item.ntnu.no

***Abstract*—This paper investigates blocking probabilities in multi-service communication systems, in which the preemptive scheduling is adopted to implement service differentiation. A novel approximation model is proposed. In contrast with existing multi-dimensional Markov model, which focuses on analyzing the small system with only two service classes and results in non-closed form expressions of blocking probabilities, our model has three major advantages: 1) it is applicable to analyzing a general multi-service scenario, independent of the number of service classes and resources; 2) the closed form expressions of blocking probabilities can be derived directly; 3) this model shows excellent extensibility for analyzing larger system which supports more service classes or common resource units. The analytical values are compared with simulation results for two- and three-service systems. Results show that the proposed model provides a high degree of accuracy in the blocking probabilities under different scenarios.**

Keywords-Markov chain; Preemptive Scheduling; blocking probability; service differentiation .

I. INTRODUCTION

Driven by increasing communication needs worldwide, a wide variety of services and applications will be brought into the future communication networks. Some of them have comparable demands to today's services, while some demands much more strict requirements in terms of bandwidth and time delay [1]. In order to meet the diverse service demands, the scheduler (in routers or switches) has to deploy efficient handling schemes to serve the different applications in different ways. In the past the programmers resorted to a rigid, pre-determined order for execution of different services, so that the corresponding service times could be predicted in advance [2]. Unfortunately these cyclic executive methods result in programs that are hard to understand and maintain because the code for logically independent tasks is interleaved. In order to guarantee the service of the safety-sensitive applications as well as simplify the task processing on large schedulers, preemptive scheduling approaches attract notable research efforts [3, 4].

In this paper we consider multi-service communication systems which integrate different kinds of applications together (some of them are safety-sensitive applications while some are safety-nonsensitive services). Central to these systems is a service facility with multiple common shared resource units (which may be interpreted according to the application under consideration as communication channels [5], computer memory sectors [6], time slots in a TDM bus [7], wavelength channels in an OPS/OBS (optical packet/burst switched) network [8, 9], etc.) and a service discipline of preemptive scheduling. That is, each type of service class is given a fixed priority and an interrupt mechanism is executed. Each class is served according to its assigned priority and the being served user can be preempted/interrupted by the higher priority arriving users in case of no available resource units. Otherwise it occupies the required resource unit for the duration of its service time.

The performance of multi-service systems with preemptive scheduling can be evaluated by the existing multi-dimensional Markov model, which is built based on a variant of the multi-dimensional Erlang blocking model. References [8]-[10] give a detailed discussion about the blocking probabilities in two-service systems (in this case applied to an OPS/OBS network) using this model. However, the existing research focuses on two-dimensional Markov models, which can only be used for analyzing small systems supporting only two kinds of service classes. For the larger system which supports more service classes or common resource units, the model will become very complex and computationally much harder to solve. The reason is that it will introduce an excessive number of states/parameters (i.e. $O(N^R)$ states and parameters, where N is the number of shared resource units and R the number of supported service classes) [11]. Another major limitation is that no closed form expressions of blocking probabilities can be derived. Hence, the existing multi-dimensional Markov models have limited applicability in modeling multi-service systems which have practical value.

In this paper we propose a novel approximation model to analyze the performance of multi-service systems with preemptive scheduling. By using conditional partitioning method, this model builds multiple levels of one-dimensional Markov chains. Each level presents all possible service states of the corresponding service class in the system. The blocking probabilities can be calculated level by level and their closed form expressions are derived directly. Compared with the existing multi-dimensional Markov model, the proposed model has several significant advantages: 1) it is applicable to analyzing the general multi-service system, independent of the number of the supported service classes and resource units; 2) the closed form expression of the blocking probability for any service class can be derived directly and separately; 3) by using the one-dimensional Markov chains to calculate blocking probabilities, the computational complexity is decreased

 ISBN: 978-1-61208-322-3

dramatically; 4) this model shows excellent extensibility for analyzing the larger system which supports more service classes and resources. Besides the model of the general multi-service case, we also give the concrete models of the two- and three-service system cases in this paper.

The rest of this paper is organized as follows. Section II presents the operation of preemptive scheduling for the studied system. Section III proposes the approximation model and derives the closed form expressions of blocking probabilities; it also gives two concrete models of two- and three-service systems. Both analytical and simulation results are given in Section IV. Section V concludes the paper.

II. THE PREEMPTIVE SCHEDULING

As shown in Fig. 1, we consider the system with a capacity of N common resource units. The system services R (R is an integer) mutually independent classes of users: class 1 has the highest priority and class R has the lowest priority. For $1 \le i \le R$, class i users are assumed to arrive according to a Poisson process with arrival rate φ_i. Meanwhile, a class i user has a request size of one resource unit and an exponentially distributed holding time with mean value of μ_i^{-1}. Thus, the average traffic offered to the system by a class i arrival process is equal to: $A_i = \varphi_i/\mu_i$.

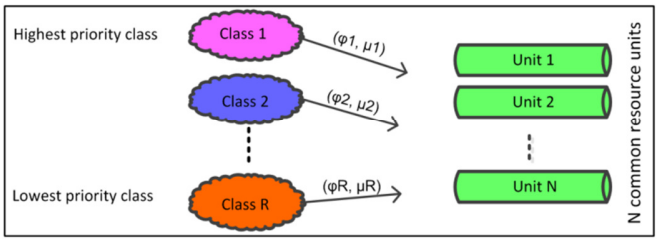

Figure 1. The traffic model of the studied system.

Fig. 2 presents the detailed operation of the preemptive scheduling when a new user arrives. All available resource units are shared among R different classes. As long as there exist available resources, the new arriving user is served directly independent of its priority. However, if all resources are occupied, this new arriving user should check its priority with that of the being served users. As illustrated in Fig. 2, we assume the lowest priority of the being served users in the system is i ($1 \le i \le R$) and the priority of this new user arrival is j. If $j \ge i$, this new user arrival will be blocked directly. If $j < i$, it will preempt/interrupt the service of class i user and takes over the respective resource unit for its own use. When i is equal to 1, all the resources are occupied by the highest priority class 1 users. Then all the new arrivals will be blocked and no preemption/interruption will happen.

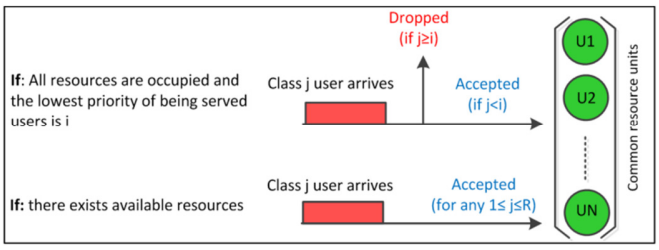

Figure 2. The operation of the preemptive scheduling.

III. ANALYTICAL MODEL

In this section, we present the approximation model to study the blocking in a multi-service communication system with preemptive scheduling. The traffic model is shown in Section II. In this part we first build the analytical model of a general R-service system and present the detailed derivation of the closed form expressions of blocking probabilities. Then we give the concrete models of the two- and three-service case systems to clarify its construction and calculation.

A. The model of the general R-service system

We model the number of the resource units occupied by each class as a continuous time Markov chain. For the R-service communication system, according to the priority of each class, the model is built from the $1st$ /top to the Rth/bottom level as shown in Fig. 3. Each level presents all possible service states of the corresponding class. The $1st$/top level gives all states of the class 1 users while the Rth/bottom level presents all states of the class R. In Fig. 3, state $i_k (0 \le i \le N, 1 \le k \le R)$ denotes that i resource units are currently serving class k users. Note that class 1 has the highest priority and class R has the lowest priority, the number of common resource units is equal to N.

For class 1 with the highest priority, blocking only happens when all resources are currently occupied by other class 1 users, hence the system can be modeled as a $M/M/N/N$ loss system as shown in the $1st$/top level.

For any state i_1 ($0 \le i_1 < N$) of class 1 in the $1st$/top level, it indicates that i_1 resources are currently serving class 1 users. Due to the higher priority of class 2 compared with classes from 3 to R, the remaining $(N - i_1)$ resources can be used for serving class 2 users. Accordingly, level 2 has a respective conditional one-dimensional Markov chain whose maximum state is $(N - i_1)$ to denote the service states of class 2. However, when i_1 is equal to N, i.e., all resources are held by class 1 users, no resource can be accessed by class 2. Hence level 2 has N conditional one-dimensional Markov chains corresponding to the different states $(i_1, 0 \le i_1 < N)$ of class 1.

For any state i_1, i_2in the $1st$/top and $2nd$ level, $i_1 + i_2 < N$, there exists a conditional one-dimensional Markov chain whose maximum state is $(N - i_1 - i_2)$ in the third level, i.e., in current i_1, i_2 resource units are busy serving class 1 and 2 users respectively. Also due to the higher priority of class 3 compared with classes from 4 to R, $(N - i_1 - i_2)$ resources can be used for serving class 3 users. Considering all possible combinations of i_1, i_2 and $i_1 + i_2 < N$, level 3 has $N * (N + 1)/2$ conditional one-dimensional Markov chains.

Using the iterative method, we build all conditional one-dimensional Markov chains in each level. Note that Fig. 3 shows only one generic one-dimensional Markov chain in each level. However, when $N > 1$, except for level 1, the other levels have more than one one-dimensional Markov chain, i.e., one for each possible combination of the states in higher levels. When calculating the blocking probabilities, conditional probability principles are used to weigh and sum contributions from each one-dimensional Markov chain in each level. Note that this model only needs to increase R or N when modeling the larger system with more classes or resources, thus the

 ISBN: 978-1-61208-322-3

proposed model shows excellent extensibility compared with multi-dimensional Markov models.

In Fig. 3, for each conditional one-dimensional Markov chain except that of level 1, the outgoing transition probability of the last state must be adjusted to take into account arrivals of higher priority users. For instance, for any one-dimensional Markov chain on level k, the last state $(N - Z_k)$ denotes that $(N - Z_k)$ common resources are currently serving class k users while Z_k resources are held by the higher priority users. Since all N resources are currently occupied, the being served class k users can be interrupted/preempted if higher priority users arrive during their holding time. Because of the lower priority of class k compared with classes from 1 to $(k-1)$, $\wedge_k$ and Z_k of the kth level in Fig. 3 are defined as:

$$\wedge_k = \sum_{j=1}^{k-1} \varphi_j , \qquad Z_k = \sum_{j=1}^{k-1} i_j . \tag{1}$$

Hence, $\wedge_R$ and Z_R in the $R\,th$ level can be written as $\wedge_R = \sum_{j=1}^{R-1} \varphi_j$, $Z_R = \sum_{j=1}^{R-1} i_j$.

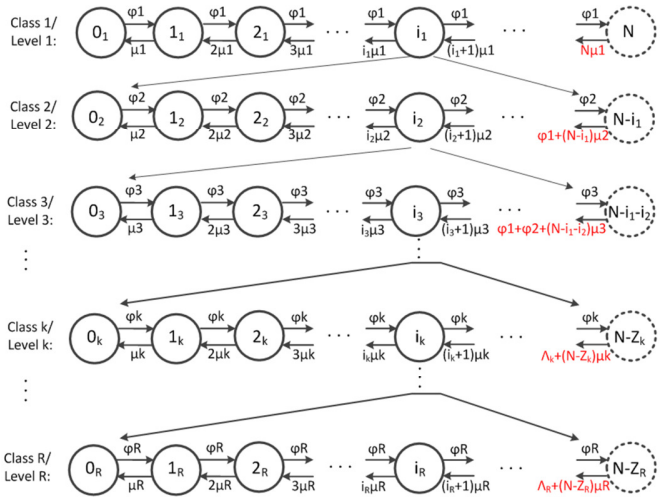

Figure 3. The model of the R-service system with capacity of N common resource units.

According to the model in Fig. 3, the blocking probability of each class can be calculated level by level. Due to the preemptive scheduling, for any class k $(1 \le k \le R)$, its blocking probability $b(k)$ only depends on the traffic pattern of classes with equal or higher priority, while not influenced by the performance of classes with lower priority. We can derive the blocking probability of each class from the highest to the lowest priority/level. In the following analysis, we use $Q_k(i_k)$ to denote the probability of state i for class k , where i resources are currently busy serving class k users.

As illustrated in Fig. 3, the blocking probability $(b(1))$ of class 1 is given directly by the $M/M/N/N$ loss formula [12]:

$$b(1) = Q_1(N) = \frac{A_1^N/N!}{\sum_{v=0}^{N} A_1^v/v!} \tag{2}$$

For any service class k $(1 < k \le R)$, $b(k)$ consists of two parts: one is the $b(k)$ introduced by the new class k user arrivals which are blocked directly; the other is the $b(k)$ given by the being served class k users which are preempted/interrupted by higher priority users. The former happens in all states where all resources are occupied by users whose priority is equal to or higher than k. According to preemptive scheduling, the new class k arrival will be blocked directly. We call these states block states. The latter happens in states where all resources are occupied by users of which the lowest priority is equal to k. In these states these being served class k users will be preempted if higher priority users arrive during their service time. We call these states preemption states. Note that the block states include all preemption states. As discussed above, in preemption states the arrival intensity of all higher priority classes is equal to $\sum_{j=1}^{k-1} \varphi_j$, and the arrival intensity of class k users is φ_k in all states of the studied system. In the following analysis, we use $Q_{k,block}$, $Q_{k,preempt}$ to denote the probabilities of block states and preemption states respectively. We also introduce Q_{all} to denote the probability of all possible service states of the studied system, which is equal to 1. Hence the blocking probability of the service class k is

$$b(k) = \frac{\varphi_k * [Q_{k,block}]}{\varphi_k * Q_{all}} + \frac{\sum_{j=1}^{k-1} \varphi_j * [Q_{k,preempt}]}{\varphi_k * Q_{all}}$$

$$= Q_{k,block} + \frac{\sum_{j=1}^{k-1} \varphi_j}{\varphi_k} * [Q_{k,preempt}] . \tag{3}$$

In order to calculate $Q_{k,block}$ and $Q_{k,preempt}$, we have to find all possible block and preemption states and their corresponding probabilities for service class k.

Block states consist of k different cases, we use $Q_{k,block,v}$ $(1 \le v \le k)$ to denote the probability of each case for class k.

$1st$ case: all resources are currently held by only class 1 when a new class k user arrives. The probability is

$$Q_{k,block,1} = Q_1(N). \tag{4}$$

$2nd$ case: all resources are currently occupied by the users whose lowest priority is 2 when a new class k user arrives. So,

$$Q_{k,block,2} = \sum_{i_1=0}^{N-1} Q_1(i_1) Q_2(N - i_1). \tag{5}$$

$v\text{-}th$ case $(3 \le v \le k)$: all resources are currently occupied by the users of which the lowest priority is v. Then new class k arrival users will be blocked. Using the iterative method, the respective probability is obtained as

$$Q_{k,block,v} = \left\{ \prod_{j=1}^{v-1} \left[\sum_{i_j=0}^{N-\sum_{l=1}^{j-1} i_l - 1} Q_j(i_j) \right] \right\} Q_v(N - \sum_{j=1}^{v-1} i_j). \tag{6}$$

where $Q_j(i_j)$ $(1 \le j \le v)$ can be derived by node equations of the corresponding one-dimension Markov chain in level j,

$$\begin{Bmatrix} Q_j(i_j) * \varphi_j = Q_j([i+1]_j) * i * \mu_j , & 0 \le i \le N - \sum_{d=1}^{j-1} i_d - 2 \\ Q_j(i_j) * \varphi_j = Q_j([i+1]_j) * [\wedge_j + (N - R_j) * \mu_j], & i = N - \sum_{d=1}^{j-1} i_d - 1 \\ \sum_{i=1}^{N-\sum_{d=1}^{j-1} i_d} Q_j(i_j) = 1 , & \end{Bmatrix} \tag{7}$$

After getting the value of $Q_{k,block,v}$,$Q_{k,block}$ can be obtained:

 ISBN: 978-1-61208-322-3

$$Q_{k,block} = \sum_{v=1}^{k} Q_{k,block,v} \quad (8)$$

$Q_{k,preempt}$ denotes the probability of preemption states that all resources are occupied and the lowest priority of users being served is k. Then the being served class k users can be preempted/interrupted by the higher priority class arrivals. Note that on preemption states, all the new class k arrivals will be blocked directly, the preemption states belong to one case of block states for class k (i.e., $v = k$ of block states). Hence,

$$Q_{k,preempt} = Q_{k,block,k} \quad (9)$$

Substituting formulas (8) and (9) into (3), we obtain the closed form expression of $b(k)$, which are expressed by φ_j, μ_j $(1 \le j \le k)$ directly. Note that one-dimensional Markov chains are used to calculate the blocking probability of each class. The corresponding computation complexity is reduced dramatically compared with solving a multi-dimensional Markov chain. Next we will give the concrete models for two- and three-service systems, both of which clarify the detailed construction, blocking calculations and the excellent extensibility of the proposed model.

B. *Example I: the model of the two-service system*

As shown in Fig. 4, the model of the two-service system is built as two levels of one-dimensional Markov chains. The $1st$ level shows an $M/M/N/N$ Erlang loss model, which presents all service states of class 1 users. For any state i_1 $(0 \le i_1 < N)$ of class 1, the $2nd$ level has a respective one-dimensional Markov chain, which gives all possible states of class 2 when i_1 resource units are busy serving class 1 users currently. Hence the level 2 has N different one-dimensional Markov chains.

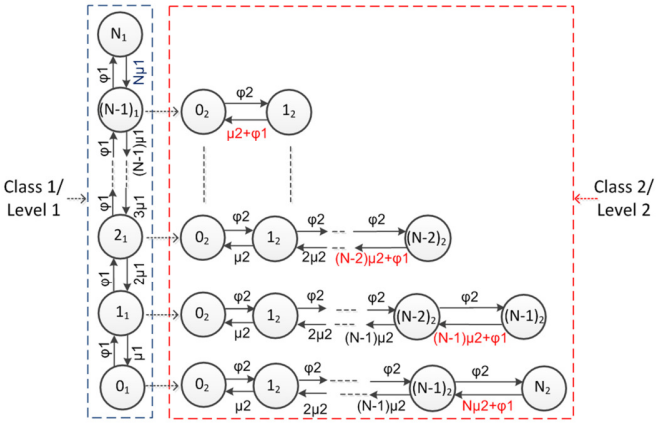

Figure 4. The model of the two-service system with capacity of N common resource units

When calculating the blocking probabilities, we use the closed form expressions directly. For class 1, its $b(1)$ is given by Erlang loss formula (i.e., formula (2)). For class 2, according to formulas (4) and (5), $Q_{2,block,1} = Q_1(N_1)$, $Q_{2,block,2} = \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1)$.

Since formulas (3), (8) and (9),

$$b(2) = Q_1(N_1) + \frac{(\varphi_1+\varphi_2)}{\varphi_2} * \sum_{i=0}^{N-1} Q_1(i_1) * Q_2(N-i_1) \quad (10)$$

where $Q_1(i_1)$, $Q_2(i_2)$ are directly obtained by formula (7).

C. *Example II: the model of the three-service system*

Fig. 5 shows the analytical model of the three-service system. Compared with the model of two-service system in Fig. 4, this model has one more level of one-dimensional Markov chains, which presents the service states of class 3 corresponding to all possible combinations of the states in first two levels. It is noticeable that, when we extend the model of two-service system into a new model of three-service system, we only need to add one more level of one-dimensional Markov chains. In addition, due to the preemptive scheduling, we also only need to calculate the blocking of the additional class 3, while the blocking expressions of the first two classes are not affected and kept unchanged. These shows the excellent extensibility of the proposed model.

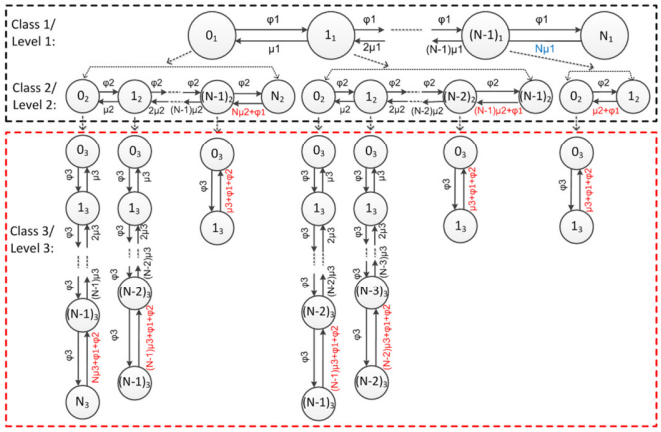

Figure 5. The model of the three-service system with capacity of N common resource units

For class 3 with the lowest priority, formulas (4)-(6) imply, $Q_{3,block,1} = Q_1(N_1)$, $Q_{3,block,2} = \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1)$, $Q_{3,block,3} = \sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-i_1-1} Q_1(i_1) * Q_2(i_2) * Q_3(N-i_1-i_2)$.

According to formulas (3), (8) and (9),

$$b(3) = Q_1(N_1) + \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1) + \left(\frac{\varphi_1+\varphi_2+\varphi_3}{\varphi_3}\right) * \sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-i_1-1} Q_1(i_1) * Q_2(i_2) * Q_3(N-i_1-i_2). \quad (11)$$

where $Q_1(i_1)$, $Q_2(i_2)$, $Q_3(i_3)$ are given by equations (7).

IV. Simulation and Analytical Results

In this section we evaluate the accuracy of the proposed model by simulations. Two- and three-service scenarios are considered. The simulator was built in the Discrete Event Modeling on Simula (DEMOS) software [13]. Ten independent simulations were performed for each parameter setting. For all simulation results we have plotted the error-bars giving the results with 95% confidence. The analytical results are obtained using formulas (1)-(11).

A. *Two-service system*

We consider a two-service communication system with 32 common resource units ($N = 32$). The total traffic ($A = A_1 + A_2$) offered by two service classes is varied from 0.1 to 1 ($0.1 \le A \le 1$). We use $S1$, $S2$ to denote the relative load value of two classes ($S1 = A_1/A$, $S2 = A_2/A$), and let $T1$, $T2$ to denote their mean holding times ($T1 = 1/\mu_1$, $T2 = 1/\mu_2$). In

this this we consider the same mean holding times of different service classes ($T1 = T2 = 1.184 * 10^{-6}$ s).

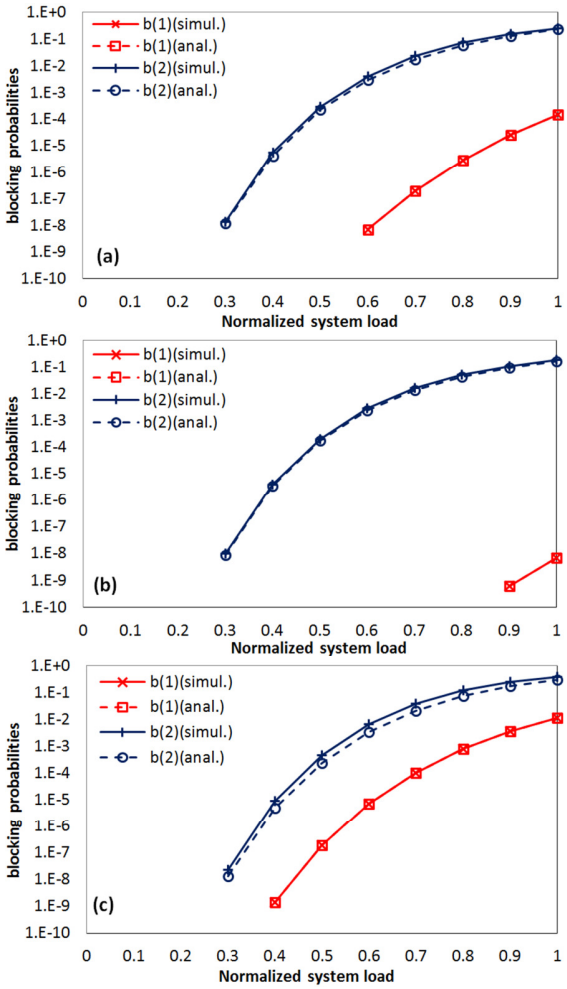

Figure 6. The blocking probabilities in a two-service system for different load allocations ($T1 = T2 = 1.184 * 10^{-6}$ s). (a). the same relative load values ($S1 = S2 = 0.5$). (b). the different relative load values ($S1 = 0.3, S2 = 0.7$). (c). the different relative load values ($S1 = 0.7, S2 = 0.3$).

Fig. 6 shows the blocking probabilities of two classes under different load allocations. We first keep $S1 = S2$ in Fig. 6(a), then change them as $S1 = 0.3, S2 = 0.7$ in Fig. 6(b) and $S1 = 0.7, S2 = 0.3$ in Fig. 6(c). Both simulation and analytical results are shown. The most important observation is that the analytical values approximate the simulation results very well under different system scenarios. We also observe that for class 1, both results overlap with each other completely for the same system load. This validates the accuracy of the analytical model, at least for the calculation of $b(1)$. Meanwhile, $b(1)$ only depends on the value of $S1$, it increases as $S1$ grows, as shown in Fig. 6(a) and (c). And it diminishes as $S1$ decreases, as shown in Fig. 6(a) and (b). This can be explained by the closed form expression of $b(1)$ in formula (2). Furthermore, although the analytical results of class 2 are very close to that of simulation for certain system load, it always produces a little smaller value, especially when $S1$ is larger than $S2$, as shown in Fig. 6(c). The reason is that when we consider the preemptive scheduling in analytical model, we use the arrival rate of class 1 to approximate its preemption probability on class 2 in the respective state. This can be seen from the outgoing transition probability of the last state in each Markov chain of level 2, as shown in Fig. 4. However, this approximation is not accurate, and the corresponding discrepancy will increases as the relative arrival rate of class 1 increases. Hence under the same system load Fig. 6(c) has larger discrepancy than Fig. 6(a) and (b), in which the discrepancy is so small and can be neglected.

B. *Three-service system*

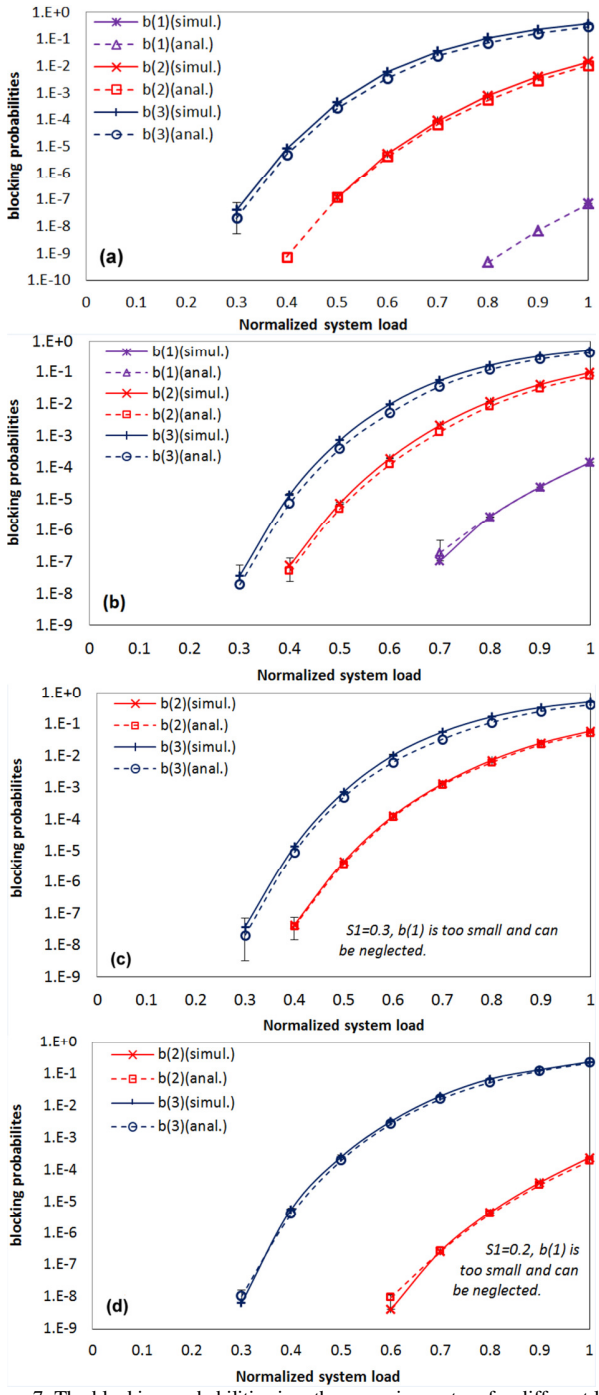

Figure 7. The blocking probabilities in a three-service system for different load allocations ($T1 = T2 = 1.184 * 10^{-6}$). (a). the same relative load values ($S1 = S2 = S3 = 1/3$). (b). the different relative load values ($S1 = 0.5, S2 =$

$0.3, S3 = 0.2$). (c). the different relative load values ($S1 = 0.3, S2 = 0.5, S3 = 0.2$). (d). the different relative load values ($S1 = 0.2, S2 = 0.3, S3 = 0.5$).

In this subsection we evaluate the proposed approximation model under a three-service scenario, i.e., class 1 has the highest priority and class 3 has the lowest priority. Same as in part A of this Section, we use $S1$, $S2$ and $S3$ ($S1 = A_1/A$, $S2 = A_2/A$, $S3 = A_3/A$, $A = A_1 + A_2 + A_3$) to denote the relative load value of three classes and let $T1, T2, T3$ ($T1 = T2 = T3 = 1.184 * 10^{-6}$ s) to denote their mean holding times. Fig. 7(a), (b), (c) and (d) present the results under different parameter settings. Both simulation and analytical results are shown. In order to further evaluate the accuracy of the propose model and get the influence of the different load allocations of three classes on the performance of the studied system, we first keep $S1 = S2 = S3$, the results are shown in Fig. 7(a), and then change the value of $S1: S2: S3$ as 5:3:2 (Fig. 7(b)), 3:5:2 (Fig. 7(c)) and 2:3:5 (Fig. 7(d)).

A number of observations can be done based on Fig. 7. The most important one is that the analytical values approximate the simulation results very well under different scenarios. This verifies the high accuracy of the proposed analytical model. We also observe that the analytical model always provides accurate $b(1)$ values for all consideration scenarios. Note that the discrepancies of class 1 in Fig. 7(b) and (d) resulted from the limited simulation times. Meanwhile, the $b(1)$ values only depend on the traffic load of class 1. As shown in Fig. 7(c) and (d), the $b(1)$ values are too small and can be neglected when $S1$ is not larger than 0.3. However, they will of course increase as $S1$ grows. In Fig. 7(b), when $S1 = 0.5$, $b(1)$ increases and the corresponding values are clearly shown. This can be explained by formula (2), the value of $b(1)$ is dominated by $S1$ for a constant N. For class 2 and 3, their blocking probabilities are only affected by the traffic pattern of the classes with same and higher priority, while not affected by the lower priority classes. Due to the operation of preemptive scheduling, the larger relative load value of one certain service class will lead to higher blocking probabilities of lower priority service classes. However, it cannot influence the blocking probabilities of higher priority service classes. As shown in Fig. 7(b) and (c), when $S1$ is increased to 0.5 while $S2$ is still 0.3, we can see the value of $b(2)$ increases over two orders of magnitude. However, comparing Fig. 7(b) and (d), when $S3$ is increased to 0.5 while keeping $S2$ unchanged, $b(2)$ decreases a lot due to the corresponding decrease in S1. For the lowest priority class 3 the blocking probability depends on the total load value of all other classes, independent of their relative load allocations. As shown in Fig. 7(b) and (c), we can see that the $b(3)$ value is kept the same under the same system load, even if the allocations of $S1$, $S2$ are different.

In addition, same as discussed in part A of this Section, although the analytical values approximate the simulation results very well under different scenarios, the proposed analytical model always produces smaller values for class 2 and 3. The reason is we made an important approximation for this model: for one service class, its preemption probability is equal to the arrival intensity of all higher priority classes. Hence this model offers high degree accurate blocking probabilities for the studied three-service system under small $S1$. Otherwise, it produces smaller values for both class 2 and 3, as shown in Fig. 7(a) and (b). Meanwhile, for class 3 with lowest priority, the corresponding discrepancy decreases as $(S1 + S2)$ diminishes. As shown in Fig. 7(d), when $S1 + S2 \geq 0.5$, the discrepancy is so small and can be neglected.

V. CONCLUSION

In this paper, we propose a novel analytical approximation model to investigate the performance of multi-service communication systems with preemptive scheduling. By using the conditional partitioning method, the proposed model builds multiple levels of one-dimensional Markov chains. Each level presents all possible service states of one service class. The corresponding blocking probability is calculated using the one-dimensional Markov chains of all higher levels as well as its own level. Its closed form expression can be derived directly and is shown in the paper. We also give the concrete models for both the two- and three-service case systems. Furthermore, the proposed model is evaluated by simulations. Both two- and three-service scenarios are considered. The results show that this model provides satisfactory approximation results under different scenarios. An additional observation is that for the lowest priority service class, its blocking probability depends on the total load value of all higher priority classes, independent of the allocation of their relative load values.

REFERENCES

[1] N. Stol, C. Raffaelli and M. Savi, "3-Level Integrated Hybrid Optical Network (3LIHON) to Meet Future QoS Requirements," *IEEE GLOBECOM 2011*, Houston, Texas, USA, Dec. 2011.

[2] C.J. Fidge, "Real-time schedulability tests for preemptive multitasking," *Computer Science*, vol. 14, no. 1, pp. 61-93, 1998.

[3] K. Lakshmanan, R. Rajkumar and J.P. Lehoczky, "Partitioned fixed-priority preemptive scheduling for multi-core processors," *In Proceedings of the 21st Euromicro Conference on Real-time Systems,* Dublin, 2009.

[4] S. Vestal, "Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance," *In Proceedings of the 28th IEEE Internationsl Real-time Systems Symposium,* Tucson, 2007.

[5] A. A. Fredericks, "Congestion in blocking systems-a simple approximation technique," *The Bell System Technical Journal*, vol. 59, no. 6, pp.805-827, 1980.

[6] E. Arthurs, J.S. Kaufman, Sizing a message store subject to blocking criteria, *In Proceeding of performance of data communications systems and their applications*, Amsterdam, 1981.

[7] J.W. Roberts, V. Mocci and I. Virtamo. "Broadband Network Teletraffic," Final Report of Action, Springer, Berlin, 1996.

[8] H. Øverby, N. Stol. "Evaluation of QoS differentiation Mechanism in Asynchronous Bufferless Optical Pakcet-Switched Networks," *IEEE Communication Magazine*, vol.44, pp. 52-57, 2006.

[9] H. Øverby, N. Stol. "Providing QoS in OPS/OBS networks with the Preemptive Drop Policy," *In Proceedings of the 3rd International Conference on Networking (ICN),* vol. 1, pp. 312-319, 2004.

[10] B. Kim, S. Lee, Y. Choi and Y. Cho, "An efficient preemption-based channel scheduling algorithm for service differentiation in OBS networks," *Computer Communications*, vol. 29, pp. 2348-2360, 2006.

[11] M. Stasiak, M. Glabowski, "A simple approximation of the link model with reservation by a one-dimensional Markov chain," *IEEE Performance Evaluation,* vol. 41, pp. 195-208, 2000.

[12] H. Akimaru, K. Kawashima, "Teletraffic theory and applications," Springer-Verlag, 1993.

[13] G. Birtwistle. "Demos-a system for Discrete Event Modelling on Simula," University of Sheffield, England S1 4DP, 2003.

 ISBN: 978-1-61208-322-3