Performance Modeling in Multi-Service Communications Systems with Preemptive Scheduling

Shuna Yang and Norvald Stol
Department of Telematics, Norwegian University of Science and Technology, Trondheim 7030, Norway
Email: shunayang2010@gmail.com; norvald.stol@item.ntnu.no

Abstract —This paper presents two separate Markov models to investigate blocking probabilities in multi-service communications systems, when preemptive scheduling is adopted to implement service differentiation. One is the generalized model, which is built as a multi-dimensional Markov chain, based on a variant of the multi-dimensional Erlang loss model. The other is the hierarchical model, which is constructed as a multi-level Markov chain, based on a combination of one-dimensional Erlang loss models. A detailed comparison when applying these two models to a general *R*-service communications system is presented. This validates the major advantages of the proposed hierarchical model: the closed form expressions of blocking probabilities; the dramatically reduced computational complexity; and the excellent scalability for analyzing larger systems. Furthermore, the analytical values are compared with simulation results for two- and three-service systems. Results show that the proposed hierarchical model provides a high degree of accuracy in the blocking probabilities under different scenarios, especially when the relative arrival rates of the lower priority classes are high.

Index Terms—Markov chain, preemptive scheduling, multi-service system, service differentiation, blocking probability

I. INTRODUCTION

Driven by increasing communication needs worldwide, a wide variety of services and applications will be brought into the future communications systems. 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], [2]. 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 applications, so that the corresponding service times could be predicted in advance [3]. 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 [4]-[6].

Manuscript received January 22, 2014; revised June 26, 2014.
Corresponding author email: shunayang2010@gmail.com.
doi:10.12720/jcm.9.6.448-460

In this paper we consider multi-service communications systems which integrate different kinds of applications together (some of them are safety-sensitive applications while some are safety-nonsensitive applications). Central to these systems is a service facility with multiple shared resource units (which may be interpreted according to the application under consideration as communication channels [7]-[9], computer memory sectors [10], time slots in a TDM bus [11], wavelength channels in a OPS/OBS (optical packet/burst switched) system [12]-[16], 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.

In order to evaluate the performance of the multi-service communications systems with preemptive scheduling, in terms of blocking probabilities, we present two different analytical models in this paper. One model is the generalized model. It is built based on a variant of multi-dimensional Erlang loss model, which is used to investigate the blocking in multi-service systems with the complete resource sharing policy [17], [18]. The main variation lies in that the generalized model considers the preemption scheduling policy, which introduces the transitions among boundary states. Here the boundary states denotes the service states where all shared resource units are occupied. Due to the preemptive scheduling, on boundary states the being served users with lower priority might be preempted/interrupted by the arriving higher priority classes users. The detailed blocking calculations of this model are given. For a system with R supported service classes and N shared common resource units, this model will introduce $O(N^R)$ states. Because of the need to solve the normalization constant (i.e. the node and global normalization equations, the number of these equations is equal to $(O(N^R)+1)$, the exact blocking probabilities are very difficult or prohibitive to be obtained when R or N is large. Hence this generalized model cannot be used for analyzing larger system which has practical meaning [19]-[21]. The other model is the hierarchical model. It is a novel approximation Markov model which was first proposed in our previous work [22]. According to the

priorities of service classes, this model builds multi-level of one-dimensional Markov chains. Each level presents all possible transmission states of the corresponding class. The blocking probability of each service class can be calculated separately. Compared with the generalized model, which has very high computational complexity and results in non-closed form expressions of blocking probabilities, the hierarchical model has several significant improvements. The most important is that this model avoids the limitation of the normalization constant. By using one-dimensional Markov chains to calculate the blocking probabilities, the computational complexity is reduced dramatically. Furthermore, the closed form expressions of the blocking probabilities can be derived separately and directly. For one specific service class, its blocking probability is expressed by the average arrival rates and holding time intensities of all classes with equal and higher priorities, the traffic patterns of the lower priority classes can be neglected. In addition, the proposed hierarchical model has excellent scalability for analyzing large systems supporting more service classes or resource units.

The remainder of this paper is organized as follows. In section II we present the traffic model of the studied system and the detailed operation of preemptive scheduling. In section III we first give the generalized model and the corresponding blocking calculation procedure, and then introduce a concrete model of a two-service communications system to clarify its calculations and limitations. Section IV proposes the hierarchical model and derives the closed form expressions of the blocking probabilities; it also presents two concrete models for two- and three-service systems. Section V compares two models in a two-service system under different scenarios. Section VI validates the accuracy of the hierarchical model by simulations in a three-service scenario. We conclude the paper in Section VII.

II. Preemptive Scheduling

As shown in Fig. 1(a), we consider the system with a capacity of N common resource units. The system serves 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. A class i user has a request size of one resource unit and an exponentially distributed holding time with mean value μ_i^{-1}. Thus, the average traffic load offered to the system by a class i arrival process is equal to: $A_i = \varphi_i/\mu_i$.

Fig. 1(b) 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 user arrival is served directly independent of its priority. However, if all resources are occupied currently, this new arrival should check its priority with that of the being served users. We assume the lowest priority of the being served users is i ($1 \le i \le R$) and the priority of this new arrival is j. If $j \ge i$, the 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, all the new arrivals will be blocked and no preemption/interruption will happen.

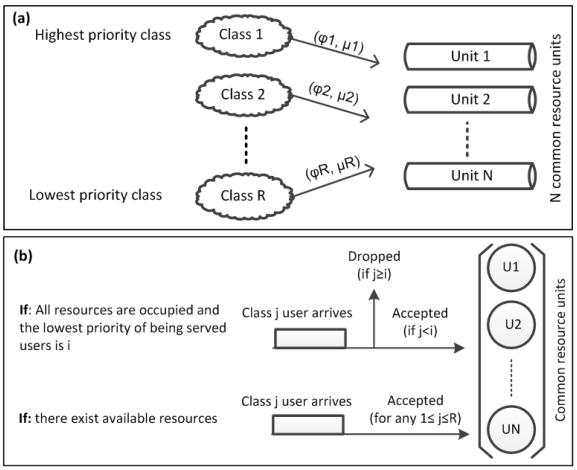

Fig. 1. (a) The traffic model of the studied system. (b) The operation of the preemptive scheduling.

In this paper we consider two distinct kinds of blocking probabilities: complete blocking probability and partial blocking probability. The former denotes the blocking probability introduced by the new user arrival which is blocked completely and directly. This happens when a new user arrives when all resources are occupied, and the being served users have higher or equal priority compared with this new arrival. Then this new user arrival is blocked completely and directly. The latter is the blocking probability given by the being served user which is preempted/interrupted during its holding time. This happens when all resources are occupied and the new user arrives, and the priority of this new arrival is higher than the lowest priority of the being served users. Then the being served user with the lowest priority will be preempted/interrupted by this new arrival. It is noticeable that, for class 1 users with the highest priority, only complete blocking probability exists; while for other classes, their corresponding blocking probabilities consist of both complete blocking probability and partial blocking probability.

III. The Generalized Markov Model of the Multi-Service System

In this section, we present the generalized model to study blocking probabilities in multi-service communications systems with preemptive scheduling. The traffic model is shown in section II. We first build the generalized model of a general R-service system and present the detailed blocking calculation procedure, and then give the concrete model of a two-service system as an example to clarify its calculations and limitations.

A. The Basic Recurrence Relations

We model the number of resource units held by each class in the system as a continuous time Markov chain. The state of the studied system is determined by the R-dimensional vector $X=(X_1,\ldots, X_i,\ldots, X_R)$ where, for $1\le i\le R$, X_i denotes the number of the resource units held by class i users in the system. Thus, in each state the total number of the busy resource units in the system is equal to $\sum_{i=1}^{i=R} X_i$.

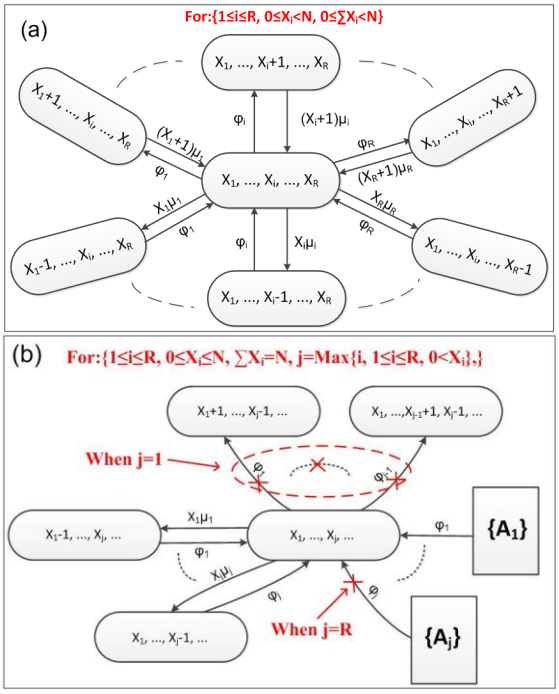

Fig. 2. The state diagram of the generalized model for a R-service communications system with capacity of N common resource units. (a). The transition diagram of normal states and (b). The transition diagram of boundary states.

For the states on which $0\le\sum_{i=1}^{i=R} X_i < N$, i.e. there still exist available resources in the system, the new user arrival will be accepted independent of its priority, no blocking or preemption happens. We call these states normal states. The transition diagram of such states for the studied R-service system is shown in Fig. 2(a). For the states on which $\sum_{i=1}^{i=R} X_i = N$, i.e. all resources are occupied, the users might be blocked or preempted due to the preemptive scheduling. We call these states boundary states. On boundary states, the new arrival will be blocked if its priority is equal or lower than the lowest priority of the being served users; meanwhile, the being served user might be preempted/interrupted if the higher priority users arrive during its service time. The transition diagram of such boundary states is given in Fig. 2(b).

As illustrated in Fig. 2(b), the state (X_1, …, X_j, …) ($1\le j\le R$) denotes the boundary state for which the last element which is larger than 0 is X_j. In this state all resource units are held by the users of which the lowest priority is j. If any of the being served users from class 1 to j has finished its service before the new user arrives, this state (X_1, …, X_j, …) will return to the corresponding normal state, like the state (X_1-1, …, X_j, …) as shown in Fig. 2(b) (i.e. the being served class 1 user finishes its service before the new user arrives). Due to the preemptive scheduling, the state (X_1, …, X_j, …) can transit to other boundary states when the being served class j user is preempted/interrupted (i.e. partial blocked) by any of the higher priority arriving users, like the state (X_1+1 …, X_j -1, …) in Fig. 2(b) (i.e. the being served class j user is preempted by class 1 user arrival). In addition, the other boundary states also can be transited into the studied state (X_1, …, X_j, …) because of the preemption. In Fig. 2(b) we use the sets $\{A_1\}$, …, $\{A_j\}$ to denote all these kinds of states. Set A_1 consists of all states which are transited because of the preemption by class 1 user, one example of these states is (X_1-1, …, X_j, 1,…). Set A_j is composed of all states which are transited due to the preemption by class j user, one of these examples is the state (X_1, …, X_j-1, 1,…). Note that when $j=1$, the state (X_1, …, X_j, …) ($1\le j\le R$) is same as (N, 0, …), all resources are held by only class 1 users currently, thus no preemption will happen; when $j=R$ the state (X_1, …, X_j, …) is same as (X_1, …, X_j, …, X_R), (X_R >0, $\sum_{i=1}^{i=R} X_i = N$), all resources are occupied and the lowest priority of the being served users is R, on this state the new class R user arrival will be blocked completely and no preemption is introduced by this new arriving class R user, hence in such case the set $\{A_j\}$ not exist, as shown in Fig. 2(b).

B. The Blocking Calculations Procedure

Section III.A presents the state diagram of the generalized model for the studied R-service system. In order to get the blocking probabilities, we have to find the probability of each system state and the transition probabilities among them, which can be solved by node equations with the normalization restriction. In the following, we use $Q(X_1, \ldots, X_i, \ldots, X_R)$ to denote the probability value of the state (X_1, …, X_i, …, X_R). The balance equations are listed as follows.

For normal states ($1\le i\le R$, $0\le X_i\le N$, $0\le\sum_{i=1}^{i=R} X_i < N$), the balance equation is written as

$$Q(\ldots, X_i, \ldots)(\sum_{i=1}^{i=R}\varphi_i+\sum_{i=1}^{i=R} X_i\mu_i)=\sum_{i=1}^{i=R} Q(\ldots, X_i+1, \ldots)\cdot (X_i+1)\mu_i+\sum_{i=1}^{i=R} Q(\ldots, X_i-1, \ldots)\varphi_i \quad (1)$$

For boundary states ($1\le i\le R$, $0\le X_i\le N$, $\sum_{i=1}^{i=R} X_i = N$, j=max{i, $1\le i\le R$, X_i >0}), the equation is

$$Q(\ldots, X_j, \ldots)(\sum_{i=1}^{j-1}\varphi_i+\sum_{i=1}^{j-1} X_i\mu_i)=\sum_{i=1}^{j}\varphi_i Q(A_i)+\sum_{i=1}^{j} Q(\ldots, X_i-1, \ldots)\varphi_i, \quad 1<j<R \quad (2)$$

$$Q(\ldots, X_j, \ldots)(\sum_{i=1}^{j-1} X_i\mu_i)=\sum_{i=1}^{j}\varphi_i Q(A_i)+\sum_{i=1}^{j} Q(\ldots, X_i-1, \ldots)\varphi_i, \quad j=1 \quad (3)$$

$$Q(\ldots, X_j, \ldots)(\sum_{i=1}^{j-1}\varphi_i+\sum_{i=1}^{j-1} X_i\mu_i)=\sum_{i=1}^{j-1}\varphi_i Q(A_i)+\sum_{i=1}^{j} Q(\ldots, X_i-1, \ldots)\varphi_i, \quad j=R \quad (4)$$

where $Q(A_i)$ denotes the summation of the probability value of each system state inside the set $\{A_i\}$.

The normalization restriction implies the summation of the probabilities of all system states is equal to 1. So

$$\sum_{X_i \in [0,N],\ i \in [1,R]} Q(X_1, \ldots X_i, \ldots, X_R) = 1. \quad (5)$$

The probability value of each system state can be obtained by Eqs. (1)-(5).

For class 1 with highest priority, blocking happens when all resources are held by only class 1 users. Its blocking probability ($b(1)$) is equal to its complete blocking probability ($b_c(1)$) and can be written as

$$b(1) = b_c(1) = Q(N, 0, \ldots, 0). \quad (6)$$

For any other class k ($1<k\leq R$), its blocking probability $b(k)$ consists of two parts: complete blocking probability ($b_c(k)$) and partial blocking probability ($b_p(k)$). As discussed in Section II, complete blocking happens on the states where all common resources are held by the users of which the lowest priority is higher than (i.e. numerically smaller than) or equal to k. We call these states blocked states. Partial blocking happens on the states where all resources are occupied by the users whose lowest priority is equal to k. We call these states preemption states. In the following, we use $Q_{k,block}$, $Q_{k,preempt}$ to denote the probability of blocked states and preemption states respectively. We also introduce Q_{all} to indicate the probability of all possible states of the studied system, which is equal to 1. According to the state diagram of the generalized model, we get

$$b_c(k) = \frac{\varphi_k(Q_{k,block})}{\varphi_k(Q_{all})} = Q_{k,block} \quad (7)$$

$$b_p(k) = \frac{\sum_{i=1}^{k-1} \varphi_i(Q_{k,preempt})}{\varphi_k(Q_{all})} = \frac{\sum_{i=1}^{k-1} \varphi_i}{\varphi_k} Q_{k,preempt} \quad (8)$$

$$b(k) = b_c(k) + b_p(k) = Q_{k,block} + \frac{\sum_{i=1}^{k-1} \varphi_i}{\varphi_k} Q_{k,preempt} \quad (9)$$

According to the discussion above, we list the blocking calculations procedure for any class k ($1\leq k\leq R$) as follows:

- Build the state diagram of the generalized model for the studied multi-service system (Fig. 2)
- List the node equations for all system states (Eqs. (1)-(4))
- List the normalization restriction equation (Eq. (5))
- Solve the node and normalization equations to get the probability value of each system state
- Find the blocked and preemption states for class k, and then get their corresponding probabilities ($Q_{k,block}$, $Q_{k,preempt}$)
- Determine the $b_c(k)$, $b_p(k)$ and $b(k)$ for any service class (Eqs. (7)-(9))

C. Example: The Generalized Model of the Two-Service System

Let the number of resource units held by class 1 and class 2 users currently be i and j, respectively. A pair (i, j) forms a two-dimensional Markov chain as shown in Fig. 3. In the following, we use $Q(i, j)$ to denote the probability value of the system state (i, j). Note that class 1 has priority over class 2.

In order to get blocking probabilities, we follow the order of calculations presented in section III.B. After building the generalized model, we list the corresponding node and normalization equations. Note that for the studied two-service system with capacity of N common resource units, the number of node equations is equal to that of system states, which is $(N+1)(N+2)/2$ as shown in Fig. 3.

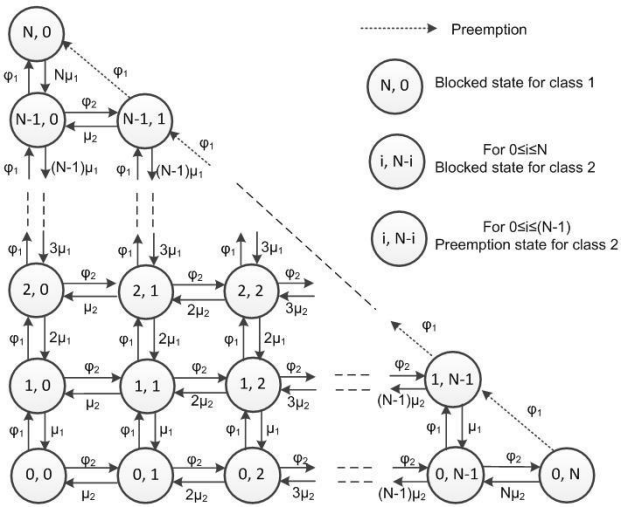

Fig. 3. The generalized model of the two-service system with capacity of N common resource units

Node equations for normal states ($0\leq i<N$, $0\leq j<N$, $0\leq i+j<N$) can be expressed as

$$Q(i, j)(\varphi_1 + \varphi_2 + i\mu_1 + j\mu_2) = Q(i-1, j)\varphi_1 + Q(i, j+1) \times (j+1)\mu_2 + Q(i+1, j)(i+1)\mu_1 + Q(i, j-1)\varphi_2 \quad (10)$$

Node equations for boundary states ($0\leq i<N$, $0\leq j<N$, $i+j=N$) are written as

$$Q(i, j)(\varphi_1 + i\mu_1 + j\mu_2) = (Q(i-1, j) + Q(i-1, j+1))\varphi_1 + Q(i, j-1)\varphi_2. \quad (11)$$

The normalization restriction implies

$$\sum_{i=0}^{N} \sum_{j=0}^{N} Q(i, j) = 1 \quad (12)$$

Substituting Eqs. (10) and (11) into Eq. (12), we can get the probability value of any state (i.e. $Q(i, j)$). The blocked and preemption states are shown in Fig. 3.

For class 1 with high priority, no partial blocking happens. Its blocking probability is

$$b(1) = b_c(1) = Q(N, 0). \quad (13)$$

For class 2, as illustrated in Fig. 3, complete blocking happens on all blocked states $(i, N-i)$ for $0\leq i\leq N$, on these states all the new arriving class 2 users are blocked completely. The partial blocking happens on all preemption states $(i, N-i)$for $0\leq i\leq(N-1)$, due to the lower priority compared with class 1, the being served class 2 users will be preempted/interrupted (i.e. lost) if the class 1 users arrive during their holding time. Thus

$$b_c(2) = Q_{2,block} = \sum_{i=0}^{N} Q(i,\ N-i)\,, \tag{14}$$

$$b_p(2) = \frac{\varphi_1}{\varphi_2} Q_{2,preempt} = \frac{\varphi_1}{\varphi_2} \sum_{i=0}^{N-1} Q(i,\ N\text{-}i)\,, \tag{15}$$

$$b(2) = b_c(2) + b_p(2) = Q(N,\ 0) + (1+\frac{\varphi_1}{\varphi_2}) \sum_{i=0}^{N-1} Q(i,\ N\text{-}i)\,. \tag{16}$$

In this part we build the generalized model to investigate blocking probabilities in multi-service systems with preemptive scheduling. As discussed above, in order to get blocking probabilities, we have to find the probability of each system state first, which are obtained by solving node and normalization equations. For the two-service system, whose model produces *(N+1)(N+2)/2* different system states, we have to solve the corresponding node and normalization equations (the number of these equations is *[(N+1)(N+2)/2+1]*). Considering a general *R*-service system with *N* common resource units, the model produces $O(N^R)$ states and the corresponding equations. The number of system states increases exponentially with that of supported service classes. This huge number of parameters and equations discourages people from using this kind of Markov model. Till now the existing research focuses on the two-dimensional generalized model, which only can be used for analyzing two-service systems [13], [15], [17], [18].

As discussed above, we conclude that the generalized model does not have practical application for analyzing large systems, especially the systems which supports more than two service classes or large number of common resource units. Next we will present a novel approximation Markov model, the hierarchical model, which employs a more direct and simple method to analyze the performance of a general multi-service system.

IV. The Hierarchical Markov Model of the Multi-Service System

In this section we propose the hierarchical model to evaluate the performance of the multi-service communications systems with preemptive scheduling. The traffic model is the same as shown in Section II. First, we give the hierarchical model of a general *R*-service system and present the detailed derivation of the closed form expressions of blocking probabilities. Then we show the concrete models of the two- and three-service systems as examples to clarify its construction and calculations.

A. The Basic Recurrence Relations

We model the number of the resource units occupied by each class as a continuous time Markov chain. For the *R*-service communications system, according to the priority of each class, the model is built from the *1*st/top to the *R*th/bottom level as shown in Fig. 4. Each level presents all possible service states of the corresponding class. The *1*st/top level gives all states of the class 1 users while the *R*th/bottom level presents all states of the class *R*. In Fig. 4, 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.

For class 1 with the highest priority, blocking only happens when all resources are currently occupied by other class 1 users, hence the studied system can be modeled as an *M/M/N/N* loss system as shown in the *1*st/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 possible states of class 2. However, when i_1 is equal to *N*, i.e., all common 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 in the 1st/top level and i_2 in the 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 service 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.

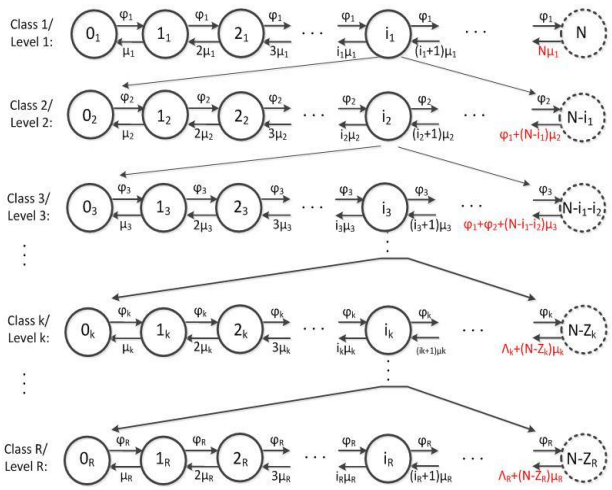

Fig. 4. The hierarchical model of the *R*-service system with capacity of *N* common resource units

Using the iterative method, we build all conditional one-dimensional Markov chains in each level. Fig. 4 shows only one but 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 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 hierarchical model shows excellent scalability compared with the generalized model.

In Fig. 4, φ_k and μ_k denote the average arrival and holding time intensity of service class k respectively. Note that 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 as shown in Fig. 4, 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 preempted/interrupted 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*), Λ_k and Z_k of the k-th level in Fig. 4 are defined as

$$\Lambda_k = \sum_{j=1}^{k-1} \varphi_j, \qquad Z_k = \sum_{j=1}^{k-1} i_j \tag{17}$$

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

B. The Closed-From Expressions of Blocking Probabilities

According to the hierarchical model in Fig. 4, the blocking probability of each service class can be calculated level by level. Due to the preemptive scheduling, for any class k ($1 \le k \le R$), its $b_c(k)$, $b_p(k)$ and $b(k)$ only depend on the traffic pattern of classes with equal or higher priority, while not affected by the performance of classes with lower priority. We can derive the blocking probability of each class from the highest to lowest priority.

As illustrated in Fig. 4, the blocking probability ($b(1)$) of class 1 is given directly by the *M/M/N/N* Erlang loss formula [23]

$$b(1) = Q_1(N_1) = \frac{A_1^N / N!}{\sum_{v=0}^{N} A_1^v / v!} \tag{18}$$

For any class k ($1 < k \le R$), as discussed in section III. B, its $b(k)$ consists of $b_c(k)$ (i.e. the complete blocking probability of class k) and $b_p(k)$ (i.e. the partial blocking probability of class k), and the former happens in all blocked states while the latter happens in all preemption states. We have to find all possible blocked and preemption states and their corresponding probabilities ($Q_{k,block}$, $Q_{k,preempt}$).

Blocked 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.

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

$$Q_{k,block,1} = Q_1(N_1) \tag{19}$$

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

$$Q_{k,block,2} = \sum_{i_1=0}^{N-1} Q_1(i_1) Q_2(N - i_1) \tag{20}$$

The v-th case ($2 < v \le k$): all resources are currently occupied by the users of which the lowest priority is v. Then new class k arrivals will be completely blocked. Using the iterative method, its probability is

$$Q_{k,block,v} = \left\{ \prod_{j=1}^{v-1} \left[\sum_{i_j=0}^{N-\sum_{i=1}^{j-1} i_j - 1} Q_j(i_j) \right] \right\} Q_v(N - \sum_{j=1}^{v-1} i_j) \tag{21}$$

where $Q_j(i_j)$ ($1 < j \le v$) is derived by the corresponding one-dimension Markov chain in level j, which is listed as

$$\begin{Bmatrix} Q_j(i_j)\varphi_j = Q_j((i+1)_j) i \mu_j, & 0 \le i \le N - \sum_{d=1}^{j-1} i_d - 2 \\ Q_j(i_j)\varphi_j = Q_j((i+1)_j)\left[\Lambda_j + (N - R_j)\mu_j\right], & i = N - \sum_{d=1}^{j-1} i_d - 2 \\ \sum_{i=1}^{N-\sum_{d=1}^{j-1} i_d} Q_j(i_j) = 1 \end{Bmatrix} \tag{22}$$

Hence, $Q_{k,block}$ can be obtained as

$$Q_{k,block} = \sum_{v=1}^{k} Q_{k,block,v} \tag{23}$$

$Q_{k,preempt}$ 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 blocked states for class k (i.e., v=k of blocked states). Hence

$$Q_{k,preempt} = Q_{k,block,k} \tag{24}$$

We also introduce Q_{all} in this part to denote the probability of all possible states of the studied system, which is equal to 1. Therefore

$$b_c(k) = \frac{\varphi_k(Q_{k,block})}{\varphi_k(Q_{all})} = Q_{k,block} = \sum_{v=1}^{k} Q_{k,block,v} \tag{25}$$

$$b_p(k) = \frac{\sum_{j=1}^{k-1} \varphi_j Q_{k,preempt}}{\varphi_k(Q_{all})} = \frac{\sum_{j=1}^{k-1} \varphi_j}{\varphi_k} Q_{k,block,k} \tag{26}$$

$$b(k) = b_p(k) + b_c(k) = \sum_{v=1}^{k} Q_{k,block,v} + \frac{\sum_{j=1}^{k-1} \varphi_j}{\varphi_k} Q_{k,block,k} \tag{27}$$

Substituting Eqs. (19)-(21) into Eqs. (25)-(27), we get the closed form expressions of $b_c(k)$, $b_p(k)$and $b(k)$, which are expressed by φ_j, μ_j ($1 \le j \le k$) directly. Note that these expressions validate one important property of the preemptive service differentiation scheme: blocking probability of one service class only depends on the traffic pattern of the classes with equal and higher priorities, while not affected by the performance of lower priority classes. In addition, Eq. (22) implies that one-dimensional Erlang loss models are used to calculate the blocking probability of each service class. The corresponding computational complexity is reduced dramatically compared with solving a multi-dimensional Markov chain in the generalized model.

Next we will introduce the concrete models for two- and three-service communications systems, both of which clarify the detailed blocking calculations and excellent scalability of the proposed hierarchical model.

C. Example I: The Hierarchical Model of the Two-Service System

As shown in Fig. 5, the hierarchical model of the two-service system is built as two levels of one-dimensional Markov chains. The *1*st/top level shows an *M/M/N/N* Erlang loss model, which presents all possible service states of class 1 users. For any state i_1 ($0 \le i_1 < N$) of service class 1, the 2nd level has a respective conditional 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 conditional one-dimensional Markov chains.

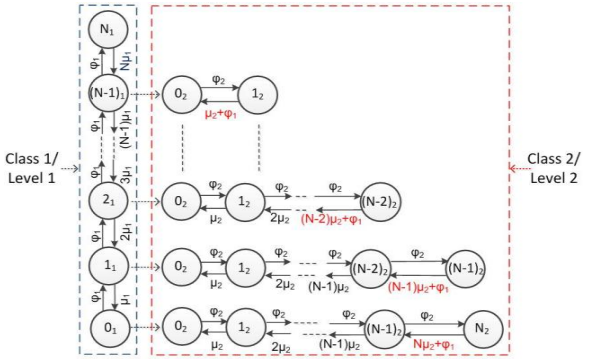

Fig. 5. The hierarchical model of the two-service system with capacity of *N* common resource units

When calculating the blocking probabilities, we use the closed form expressions directly. For class 1, its *b(1)* is given by the Erlang loss formula (i.e., Eq. (18)) directly. For class 2, according to Eqs. (19) and (20), $Q_{2,block,1} = Q_1(N_1)$, $Q_{2,block,2} = \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1)$.

Since Eqs. (23)-(27),

$$b_c(2) = \sum_{v=1}^{2} Q_{2,block,v} = Q_1(N_1) + \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1) \quad (28)$$

$$b_p(2) = \frac{\varphi_1}{\varphi_2} Q_{2,block,2} = \frac{\varphi_1}{\varphi_2} \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1) \quad (29)$$

$$b(2) = Q_1(N_1) + \frac{(\varphi_1+\varphi_2)}{\varphi_2} \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1) \quad (30)$$

where $Q_1(i_1)$, $Q_2(i_2)$ is obtained directly by Eq. (22).

D. Example II: The Hierarchical Model of the Three-Service System

For the three-service communications system, the hierarchical model is built as shown in Fig. 6. Compared with the model of two-service system in Fig. 5, 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. Note that, for the three-service system, the corresponding model only adds one more level of conditional one-dimensional Markov chains to that of the two-service system. This clearly shows the excellent scalability of the proposed hierarchical model.

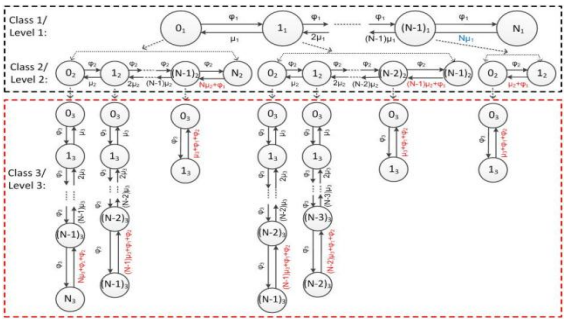

Fig. 6. The hierarchical model of the three-service system with capacity of *N* common resource units

As discussed in Section IV.A, due to the preemptive scheduling, the blocking probabilities of class 1 and 2 are not affected by class 3. They are shown in Eqs. (18), (28)-(30). For the studied three-service system, we only need to calculate $b_c(3)$, $b_p(3)$ and $b(3)$ of class 3.

For class 3 with the lowest priority, Eqs. (19)-(21) imply $Q_{3,block,1} = Q_1(N_1)$, $Q_{3,block,2} = \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1)$, $Q_{3,block,3} = \sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-i_1-1} Q_1(i_1)Q_2(i_2)Q_3(N-i_1-i_2)$.

Since Eqs. (25)-(27),

$$b_c(3) = Q_{3,block} = Q_1(N_1) + \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1) + \sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-i_1-1} Q_1(i_1)Q_2(i_2)Q_3(N-i_1-i_2), \quad (31)$$

$$b_p(3) = \frac{\sum_{j=1}^{2}\varphi_j}{\varphi_3} Q_{3,block,3} = \frac{\varphi_1+\varphi_2}{\varphi_3} \sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-i_1-1} Q_1(i_1)Q_2(i_2)Q_3(N-i_1-i_2), \quad (32)$$

$$b(3) = Q_1(N_1) + \sum_{i_1=0}^{N-1} Q_1(i_1)Q_2(N-i_1) + \frac{\varphi_1+\varphi_2+\varphi_3}{\varphi_3} \sum_{i_1=0}^{N-1}\sum_{i_2=0}^{N-i_1-1} Q_1(i_1)Q_2(i_2)Q_3(N-i_1-i_2). \quad (33)$$

where $Q_1(i_1)$, $Q_2(i_2)$, $Q_3(i_3)$ are given directly by Eq. (22).

V. Comparision of The Simulation and Calculation Results

In this part we will evaluate the accuracy of the proposed hierarchical model. Since the generalized model are widely used for analyzing two-service communications systems and its accuracy has been verified in references [13], [15], we compare the analytical results from both models for two-service system first. Then we extend the studied system into a three-service communications system, the analytical values from the hierarchical model are compared with the simulation results. Many different kinds of scenarios are considered.

A. The Analytical Results of a Two-Service System

We consider a two-service communications 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 S_1, S_2 to denote the relative load value of two classes ($S_1=A_1/A$, $S_2=A_2/A$), and let T_1, T_2 to denote their mean holding times ($T_1=1/\mu_1$, $T_2=1/\mu_2$). Fig. 7-11 show the results under different parameter settings, (G.) denotes the results from the generalized

Markov Model and (H.) indicates them from the proposed hierarchical model. Unless stated otherwise, these symbols are used in the following results. In order to evaluate the accuracy of the hierarchical model completely, we first set the value of S_1:S_2 as 1:9 (Fig. 8) and 9:1 (Fig. 9) while keeping T_1=T_2, and then change the value of T_1:T_2 as 10:1 (Fig. 10) and 1:10 (Fig. 11) while keeping S_1=S_2 respectively. Fig. 7 shows the reference case in which S_1=S_2 and T_1=T_2. The main observation is that the results from the hierarchical model approximate that of the generalized model very well under different scenarios.

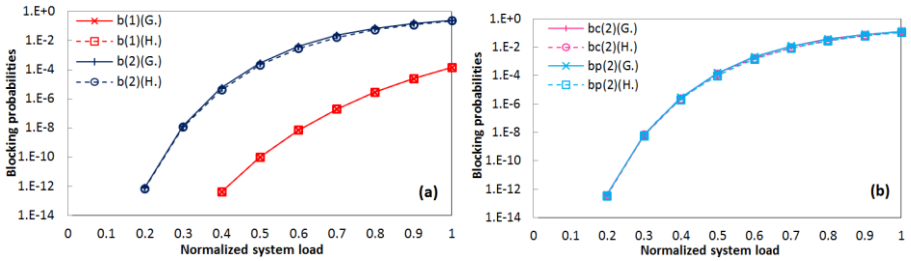

Fig. 7. Comparison of two models for a two-service system (S_1=S_2=0.5, T_1=T_2=*1.184e-6s*). (a). The blocking probabilities of class 1 and 2 (*b(1)*, *b(2)*). (b). The complete and partial blocking probabilities of class 2 ($b_c(2)$, $b_p(2)$).

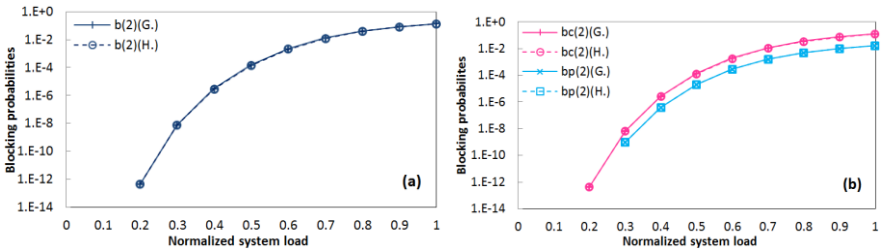

Fig. 8. Comparison of two models for a two-service system (S_1=0.1, S_2=0.9, T_1=T_2=*1.184e-6s*). (a). The blocking probabilities of class 1 and 2 (*b(1)*, *b(2)*). (b). The complete and partial blocking probabilities of class 2 ($b_c(2)$, $b_p(2)$).

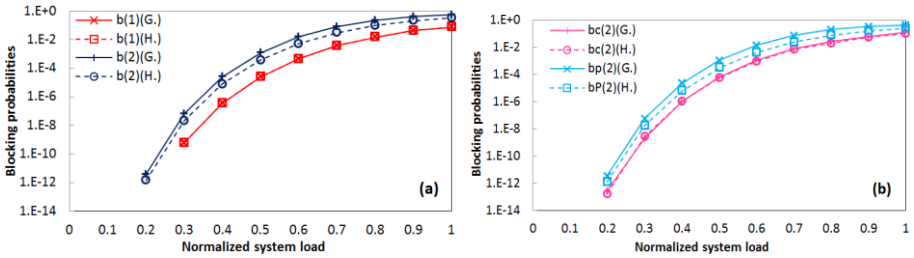

Fig. 9. Comparison of two models for a two-service system (S_1=0.9, S_2=0.1, T_1=T_2=*1.184e-6s*). (a). The blocking probabilities of class 1 and 2 (*b(1)*, *b(2)*). (b). The complete and partial blocking probabilities of class 2 ($b_c(2)$, $b_p(2)$).

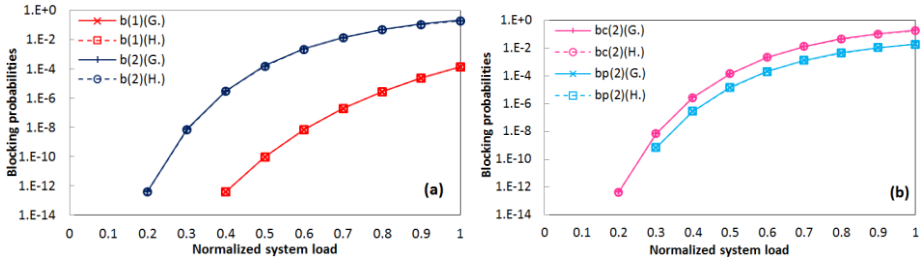

Fig. 10. Comparison of two models for a two-service system (S_1=S_2=0.5, T_1=$10T_2$=*1.184e-5s*). (a). The blocking probabilities of class 1 and 2 (*b(1)*, *b(2)*). (b). The complete and partial blocking probabilities of class 2 ($b_c(2)$, $b_p(2)$).

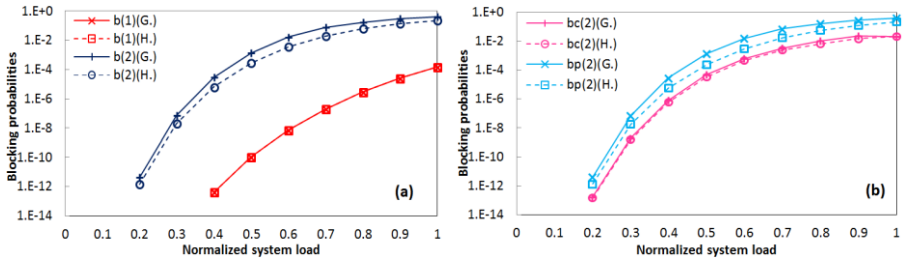

Fig. 11. Comparison of two models for a two-service system (S_1=S_2=0.5, T_2=$10T_1$=*1.184e-5s*). (a). The blocking probabilities of class 1 and 2 (*b(1)*, *b(2)*). (b). The complete and partial blocking probabilities of class 2 ($b_c(2)$, $b_p(2)$).

As shown in Fig. 7-11, for class 1 with high priority, both models give the same $b(1)$ values under different scenarios. As discussed in Section IV, the $b(1)$ value from the hierarchical model is got by Erlang loss formula directly. However, in the generalized model, it has to be calculated by complex computation according to the order of calculations. We can conclude that the hierarchical model reduces the computation complexity significantly, also for the calculation of the $b(1)$ value. In addition, the $b(1)$ value only depends on the traffic load offered by class 1 while not affected by T_1. This can be explained by Eq. (18). As shown in Fig. 7-9, $b(1)$will decrease as S_1 decreases or S_2 increases while not be influenced by T_1 values as in Fig. 10-11.

For class 2 with low priority, the blocking values of $b(2)$, $b_c(2)$, $b_p(2)$ from two models perfectly coincide with each other, as shown in Fig. 7, 8 and 10. When $S_1 \leq S_2$ under $T_1=T_2$ or $T_1 \geq T_2$ under $S_1=S_2$, a lot of additional cases have also been checked, all of which confirms that similar accuracy can be obtained for the hierarchical model. As shown in Fig. 9 and 11, although the analytical values of the hierarchical model are very close to that of the generalized model, it always produces smaller $b(2)$, $b_p(2)$. A lot of additional cases have been checked (for $S_1 \geq S_2$ under $T_1=T_2$ or $T_1 \leq T_2$ under $S_1=S_2$), the results show that this discrepancy decreases as S_1:S_2 or T_2:T_1 decreases. Another observation is that this discrepancy is mainly from the partial blocking probability $b_p(2)$. As shown in Fig. 9 and 11, when S_1:S_2 is very high or T_2 is much larger than T_1, the value of $b_p(2)$ will dominate $b(2)$and the corresponding discrepancy will determine the total discreapancy of the blocking probability of class 2. The reason of this discrepancy will be shown later. Furthermore, the partial blocking value of class 2 will increase as its relative load value increase. As shown in Fig. 9 and 10, $b_p(2)$ is much larger than $b_c(2)$ and dominates $b(2)$. Otherwise, in Fig. 8 and 10 for which class 2 has the higher relative load value and heavier arrival intensity, $b_c(2)$ is much larger and dominates $b(2)$. In addition, from the results in Fig. 7-11, we found both models provide very similar $b_c(2)$ under different scenarios. These results show that the proposed hierarchical model provides highly accurate complete blocking probability values (i.e. $b_c(2)$) of the studied system.

According to the results in Fig. 7-11, we can conclude the applicability of the proposed hierarchical model as follows.

Conclusion: For two-service communications systems, the hierarchical model provides accurate blocking probabilities for class 1 ($b(1)$). However for class 2, it gives highly accurate blocking values ($b(2)$, $b_c(2)$, $b_p(2)$) when the arrival rate of class 1 is not larger than that of class 2 (i.e. $\varphi_1 \leq \varphi_2$); otherwise although the results from this model approximate the blocking probabilities of class 2 (i.e. ($b(2)$, $b_c(2)$) very well, it always produces smaller values, and this discrepancy mainly comes from $b_p(2)$. In addition, this model provides highly accurate complete blocking probability for class 2 (i.e. $b_c(2)$).

Reason: For the proposed hierarchical model, we model the service state of class 1 as an *M/M/N/N* loss model. Because of the high priority of class 1, its blocking value is not affected by the performance of class 2. Hence the hierarchical model provides the accurate $b(1)$ values under different scenarios, which is obtained by Erlang loss formula directly. For class 2, we consider the preemptive scheduling when building each conditional one-dimensional Markov chain. The outgoing transition probability of the last state in level 2 introduces the arrival rate of class 1(i.e. φ_1), which is used for approximating its preemption probability on class 2. However, this approximation is not accurate. When the preemption happens, the current service states of both classes will transit at the same time: the service state of class 1 transits into the state (i_1+1) in level 1; The state of class 2 (i.e. state (N-i_1)) transits into the state (N-i_1-1) of another conditional Markov chain corresponding to the state (i_1+1) of class 1. In the hierarchical model, we cut off the relations among these conditional Markov chains and consider them separately, thus the discrepancy occurs. This discrepancy mainly results from the approximation of $b_p(2)$ and its value becomes larger as φ_1 increases. Since $T_1=1/\mu_1$ and $A_1=AS_1=\varphi_1/\mu_1$, φ_1 is equal to AS_1/T_1, the value of φ_1 will increase if we increase S_1 or decrease T_1. Accordingly, the discrepancy of $b_p(2)$ values increases as S_1 increases or T_1 decreases. They also lead to the discrepancy of $b(2)$ values. All of these are validated by the results shown in Fig. 7-11.

B. The Simulation and Analytical Results for a Three-Service System.

In this subsection we evaluate the proposed hierarchical model under a 3-service scenario, i.e., class 1 has the highest priority and class 3 has the lowest priority. Fig. 12-18 presents the results under different parameter settings. (S.) denotes the results from the simulation model. (H.) denotes the analytical results. The simulator is built using the Discrete Event Modeling on Simula (DEMOS) software [24]. Ten independent simulations are performed for each parameter setting. For all simulation results we have plotted the average values with error-bars giving the results with 95% confidence (in some cases these are however too narrow to be visible). As in Section 5.1, we use S_1, S_2 and S_3 ($S_1=A_1/A$, $S_2=A_2/A$, $S_3=A_3/A$, $A=A_1+A_2+A_3$) to denote the relative load value of three classes and let T_1, T_2, T_3 ($T_1=1/\mu_1$, $T_2=1/\mu_2$, $T_3=1/\mu_3$) to denote their mean holding times. In order to evaluate the hierarchical model completely, we first give the reference case in which $S_1=S_2=S_3$ and $T_1=T_2=T_3$ (Fig. 12), then we set the value of S_1:S_2:S_3 as 2:2:6 (Fig. 13), 2:6:2 (Fig. 14) and 6:2:2 (Fig. 15) while keeping $T_1=T_2=T_3$. Furthermore we adjust T_1:T_2:T_3 as 5:1:1 (Fig. 16), 1:5:1 (Fig. 17) and 1:1:5 (Fig. 18) while keeping $S_1=S_2=S_3$. Both analytical values and simulation results are shown. The main observation is that the results from the hierarchical model approximate that of simulations very well under different scenarios.

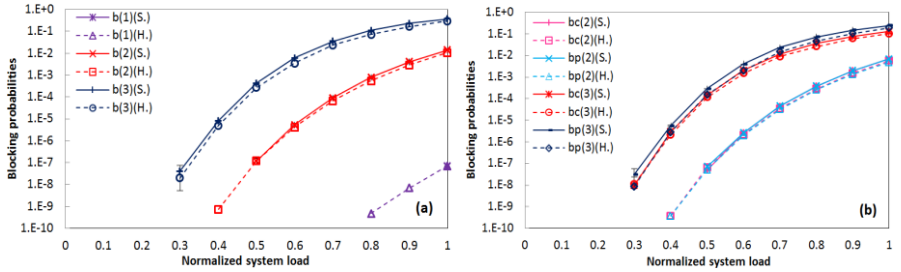

Fig. 12. Comparison of analytical and simulation results for the three-service system ($S_1=S_2=S_3=1/3$, $T_1=T_2=T_3=1.184e\text{-}6s$). (a). The blocking probabilities of class 1, 2 and 3 ($b(1)$, $b(2)$, $b(3)$). (b). The complete and partial blocking probabilities of class 2 and 3 ($b_c(2)$, $b_p(2)$, $b_c(3)$, $b_p(3)$).

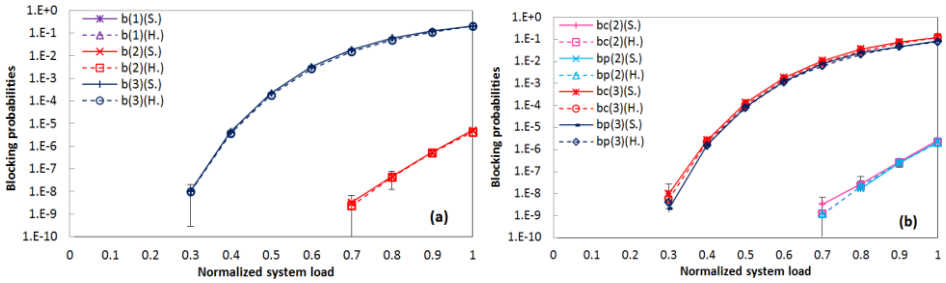

Fig. 13. Comparison of analytical and simulation results for the three-service system ($S_1=0.2$, $S_2=0.2$, $S_3=0.6$, $T_1=T_2=T_3=1.184e\text{-}6s$). (a). The blocking probabilities of class 1, 2 and 3 ($b(1)$, $b(2)$, $b(3)$). (b). The complete and partial blocking probabilities of class 2 and 3 ($b_c(2)$, $b_p(2)$, $b_c(3)$, $b_p(3)$).

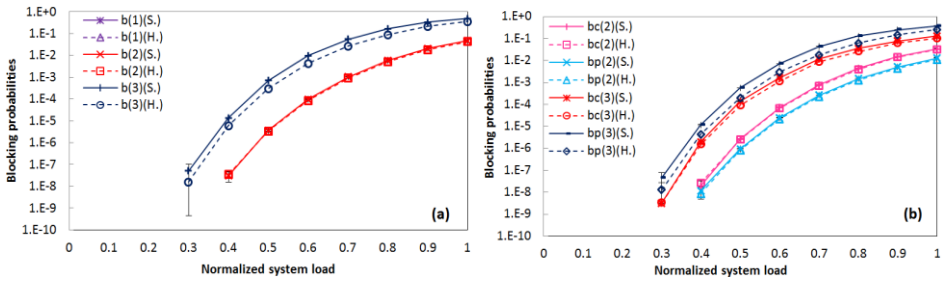

Fig. 14. Comparison of analytical and simulation results for the three-service system ($S_1=0.2$, $S_2=0.6$, $S_3=0.2$, $T_1=T_2=T_3=1.184e\text{-}6s$). (a). The blocking probabilities of class 1, 2 and 3 ($b(1)$, $b(2)$, $b(3)$). (b). The complete and partial blocking probabilities of class 2 and 3 ($b_c(2)$, $b_p(2)$, $b_c(3)$, $b_p(3)$).

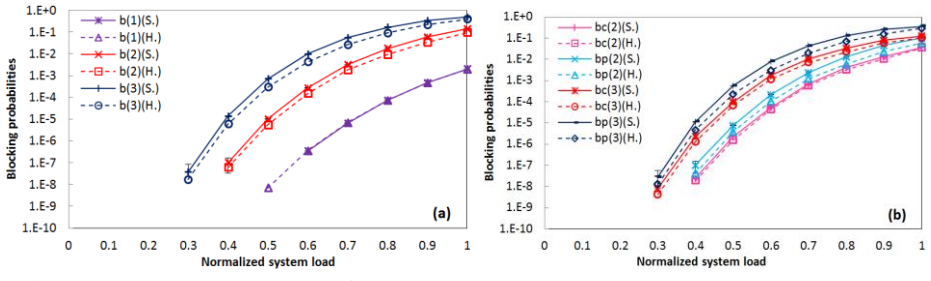

Fig. 15. Comparison of analytical and simulation results for the three-service system ($S_1=0.6$, $S_2=0.2$, $S_3=0.2$, $T_1=T_2=T_3=1.184e\text{-}6s$). (a). The blocking probabilities of class 1, 2 and 3 ($b(1)$, $b(2)$, $b(3)$). (b). The complete and partial blocking probabilities of class 2 and 3 ($b_c(2)$, $b_p(2)$, $b_c(3)$, $b_p(3)$).

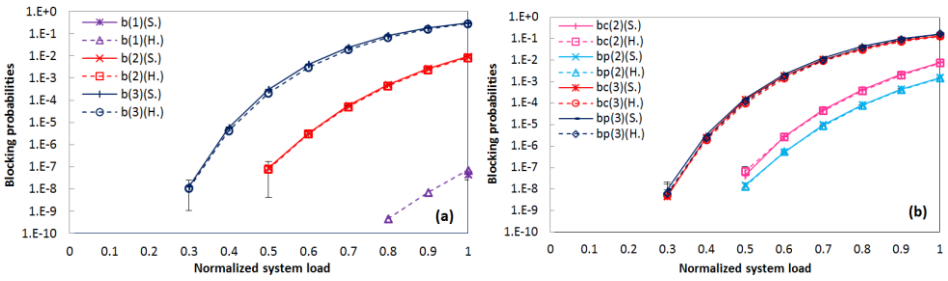

Fig. 16. Comparison of analytical and simulation results for the three-service system ($S_1=S_2=S_3=1/3$, $T_1=5T_2=5T_3=1.184e\text{-}6s$). (a). The blocking probabilities of class 1, 2 and 3 ($b(1)$, $b(2)$, $b(3)$). (b). The complete and partial blocking probabilities of class 2 and 3 ($b_c(2)$, $b_p(2)$, $b_c(3)$, $b_p(3)$).

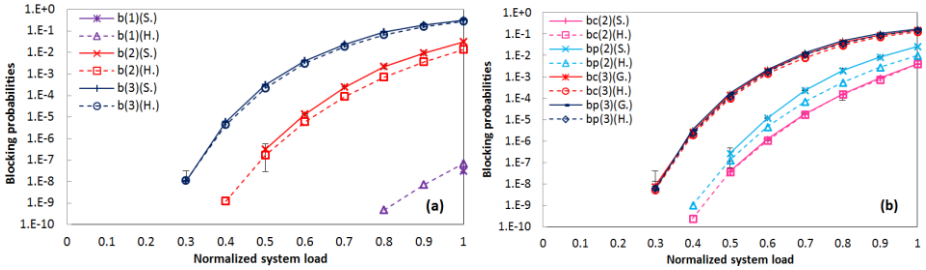

Fig. 17. Comparison of analytical and simulation results for the three-service system (S_1=S_2=S_3=1/3, T_2=5T_1=5T_3=*1.184e-6s*). (a). The blocking probabilities of class 1, 2 and 3 (*b(1), b(2), b(3)*). (b). The complete and partial blocking probabilities of class 2 and 3 ($b_c(2)$, $b_p(2)$, $b_c(3)$, $b_p(3)$).

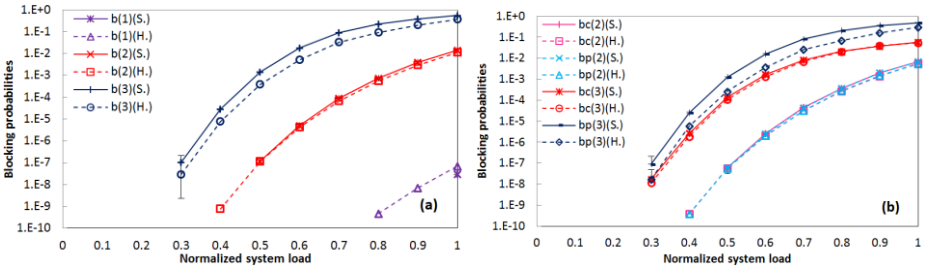

Fig. 18. Comparison of analytical and simulation results for the three-service system (S_1=S_2=S_3=1/3, T_3=5T_1=5T_2=*1.184e-6s*). (a). The blocking probabilities of class 1, 2 and 3 (*b(1), b(2), b(3)*). (b). The complete and partial blocking probabilities of class 2 and 3 ($b_c(2)$, $b_p(2)$, $b_c(3)$, $b_p(3)$).

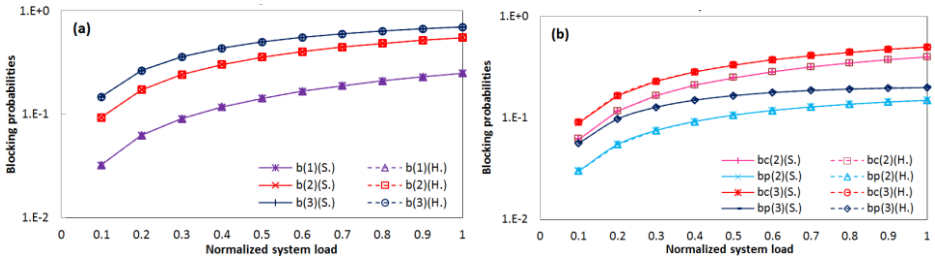

Fig. 19. Evaluation of the hierarchical model for a three-service system with capacity of only one common resource unit (*N=1*).

Results from Fig. 12-18 show that the proposed hierarchical model provides the accurate *b(1)* values under different scenarios, and its value only depends on its own traffic load. This can be explained by formula (18), which is expressed by A_1 and N. The value of *b(1)* increases a lot when increasing S_1 from 1/3 to 0.6 (as shown in Fig. 12 and 15), and it decrease when decreasing S_1 from 1/3 to 0.2 (as shown Fig. 12 and 13/14); meanwhile the *b(1)* value is not affected by T_1 when keeping S_1 unchanged (as shown in Fig. 16-18). For other classes, the hierarchical model offers high degree accurate blocking probabilities under small relative load value of the higher priority classes for T_1=T_2=T_3 or large relative holding time of the higher priority classes for S_1=S_2=S_3. Otherwise, even if the analytical results are very close to that of simulations, it always produces smaller values. For instance, considering class 2, the hierarchical model provides the highly accurate blocking probability under small S_1 (as shown in Fig. 12-14) or large T_1 (as shown in Fig. 16). Otherwise this model gives the smaller blocking probabilities as shown in Fig. 15 (S_1=*0.6*) and Fig. 17 (T_1=*5×1.184e-6s*). When considering class 3, the accuracy of its blocking probabilities from this model depends on the performance of both classes 1 and 2. As shown in Fig. 12-18, the hierarchical model provides highly accurate *b(3)* values when S_1+S_2 ≤2/3(Fig. 12 and 13) or smaller T_3 compared with T_1 or T_2 (as shown Fig. 16 and 17), otherwise the analytical values of *b(3)* is smaller (Fig. 14, 15, 18). Furthermore, we find the discrepancy between the analytical and simulation results mainly comes from the partial blocking probability. As shown in Fig. 14-15 and Fig. 17-18, the discrepancy from the partial blocking probability dominates that of the corresponding total blocking probability. The reason of this is described in Section V.A, which shows that the approximation, which approximates the arrival intensity of all higher priority classes as its preemption probability, is not accurate. In addition, we also observe that the hierarchical model always provides the highly accurate complete blocking probabilities for the studied system. As shown in Fig. 12-18, the hierarchical model always provides accurate values of $b_c(2)$ and $b_c(3)$ under different scenarios. Fig. 12-18 also validates one important property of preemptive scheduling: for the service class with the lowest priority, its blocking probability depends on the total load value of all higher priority classes, independent of their load allocation. As shown in Fig. 14 and 15, the value of *b(3)* is not affected by the variation of the allocation between S_1 and S_2.

According to the results from Fig. 12-15, we observe that the blocking probabilities of three classes are mainly depends on their load allocations. However, if we vary the mean holding times of them while keeping the same load allocation, the blocking probability of each class will keep quite stable under the same system load, as shown in

Fig. 15(a)-18(a). It is noticeable that the allocation of complete and partial blockings changes dramatically. The reason for this is the holding time of one class determines its preemption/interrupted probability, the partial blocking value of one class (i.e. the probability of being preempt/interrupted) increases as its holding time increases, like $b_p(2)$ in Fig. 17(b) and $b_p(3)$ in Fig. 18(b).

C. Results for a Three-Service System with Only One Common Resource Unit.

Fig. 19 presents the simulation results and blocking probabilities from the proposed hierarchical model for the case that the studied system has only one common resource unit, the parameter settings (except that $N=1$) are same as in Fig. 12. Both results for three classes are shown. The most important observation is that the blocking probabilities (both complete blocking and partial blocking probabilities) from the hierarchical model completely coincide with that of the simulation model. No discrepancy exists between them. The reason is that the approximation made in hierarchical model is accurate for the studied system in case of only one common resource unit. For instance, for a 3-service system with $N=1$, we use the arrival rate intensities of class 1 and 2 to approximate their preemption probability on class 3. Since N is equal to 1, on the boundary state (i.e. the resource is occupied by class 3 user), if any higher priority class user (i.e. the user of class 1 or 2) arrives, the being served class 3 user will be preempted/interrupted, its service state will go back to state (0) while the state of this higher priority arriving user will transit to state (1). This is in accord with our hierarchical model. Hence the hierarchical model provides the accurate blocking probabilities. In addition, for the studied system with only one common resource for all classes, the proposed model provides accurate blocking probability for each class with much simpler calculation. For a R-service system, the generalized model is a R-dimensional Markov model. Because of the normalization limitation, we need to solve (R+1) equations to get those R parameters (i.e. the probability value of each system state). However, using the hierarchical model, we can get the blocking probability of each class by the closed form expression directly, independent of the number of the service classes supported by the system.

VI. Conclusions

In this paper two different Markov models, which are used for analyzing the performance of the multi-service communication systems with preemptive scheduling, have been presented. One model is the generalized model-based on a variant of the multi-dimensional Erlang loss model. Due to the complexity of finding the normalization constant, existing research using this model is limited to analyzing small systems with only two service classes. The other model is a novel approximate hierarchical model-based on the combination of conditional one-dimensional Markov chains. It is built as multiple levels of one-dimensional Markov chains. Each level presents all possible service states for one class. This model analyzes the performance of each service class directly and separately, the corresponding computational complexity is reduced dramatically. Furthermore, this model can be applied to a general multi-service scenario. Closed form expressions of the blocking probabilities are also given by this hierarchical model.

In order to investigate the accuracy and the applicability of the proposed hierarchical model, we compare the analytical results with that of the generalized model and with simulations, in two- and three-service systems respectively. Results show that the accurate blocking probability of class 1 can be got by the Erlang loss formula directly in our hierarchical model. For other classes, this model gives highly accurate blocking probabilities under different scenarios, especially when the relative arrival rates of the lower priority classes are higher. Otherwise, although this model always offers satisfactory results, it produces smaller values. This discrepancy mainly comes from the corresponding partial blocking probability. The reason for this is given in the paper. In addition, this hierarchical model provides accurate blocking probabilities for all classes in case of the system with only one common resource unit.

References

[1] N. Stol, C. Raffaelli, and M. Savi, "3-Level integrated hybrid optical network (3LIHON) to meet future QOS requirements," in *Proc. IEEE Global Telecommunication Conference*, 2011, pp. 1-6.

[2] H. Kopetz, *Real-Time Systems*, Springer-New York, 2009.

[3] C. J. Fidge, "Real-time scheduling theory: A historical perspective," *Real-Time Systems*, vol. 28, no. 2-3, 2004, pp. 101-155.

[4] K. Lakshmanan, R. Rajkumar, and J. P. Lehoczky, "Partitioned fixed-priority preemptive scheduling for multi-core processors," in *Proc. 21st Euromicro Conference on Real-time Systems*, 2009, pp. 239-248.

[5] S. Vestal, "Preemptive scheduling of multi-criticality systems with varying degrees of execution time assurance," in *Proc. 28th IEEE International Real-time Systems Symposium*, 2007, pp. 239-243.

[6] N. Scaife and P. Caspi, "Integrating model-based design and preemptive scheduling in mixed time- and event-triggered systems," in *Proc. 16th Euromicro Conference on Real-time Systems*, 2004, pp. 119-126.

[7] M. C. Necker, "A comparison of scheduling mechanisms for service class differentiation in HSDPA networks," *Journal of Electronics and Communications*, vol. 60, no. 2, pp. 136-141, 2006.

[8] C. K. Tham, Q. Yao, and Y. Jiang, "A multi-class probabilistic priority scheduling discipline for differentiated services networks," *Computer Communications*, vol. 25, no. 17, pp. 1487-1496, 2002.

[9] J. C. Chuang and M. A. Sirbu, "Distributed network storage service with quality-of-service guarantees," *Journal of Network and Computer Applications*, vol. 23, no. 3, pp. 163-185, 2000.

[10] G. Romanazzi, P. K. Jimack, and C. E. Goodyer, "Reliable performance prediction for parallel scientific software in a multi-

cluster grid environment," in *Proc. 6th Engineering Computational Technology Conference*, 2008, pp. 1-14.
[11] R. Stefan and K. Goossens, "A TDM slot allocation flow based on multipath routing in NoCs," *Journal of Microprocessors & Microsystems*, vol. 35, no. 2, pp. 130-138, 2011.
[12] H. Øverby and N. Stol, "A teletraffic model for service differentiation in OPS networks," in *Proc. 8th Optoelectronic and Communications Conference*, 2003.
[13] H. Øverby and N. Stol, "Quality of Service in asychronous bufferless optical packet switched networks," *IEEE Telecommunication Systems*, vol. 27, pp. 151-179, 2004.
[14] L. Yang, Y. Jiang, and S. Jiang, "A probabilistic preemptive scheme for providing service differentiation in OBS networks," in *Proc. IEEE Global Telecommunication Conference*, 2003.
[15] H. Øverby and N. Stol, "Providing QoS in OPS/OBS networks with the preemptive drop policy," in *Proc. 3rd International Conference on Networking*, 2004, pp. 312-319.
[16] H. C. Cankaya, S. Charcranoon, and T. S. E. Bawab, "A preemptive scheduling technique for OBS networks with service differentiation," in *Proc. IEEE Global Telecommunications Conference*, 2003, pp. 2704-2708.
[17] B. C. Kim, S. H. Lee, Y. S. Choe, and Y. Z. Cho, "An efficient preemption-based channel scheduling algorithm for service differentiation in OBS networks," *IEEE Computer Communications*, vol. 29, pp. 2348-2360, 2006.
[18] H. Øverby and N. Stol, "Evaluation of QoS differentiation mechanism in asynchronous bufferless Optical Pakcet-Switched networks," *IEEE Communication Magazine*, vol. 44, pp. 52-57, 2006.
[19] M. Stasiak and 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.
[20] M. Glabowski, A. Kaliszan, and M. Stasiak, "Modeling product-form state-dependent systems with BPP traffic," *Journal of Performance Evaluation*, vol. 2010, no. 67, pp. 174-197, 2010.
[21] M. Stasiak, "An approximate model of a switching network carrying mixture of different multichannel traffic streams," *IEEE Transactions on Communications*, vol. 41, no. 6, pp. 836-840, 1993.
[22] S. Yang, N. Stol, and H. Overby, "A novel approach in modeling multi-service optical packet switched networks with the preemptive drop policy," in *Proc. IEEE ONDM Conference*, 2013, pp. 88-93.
[23] H. Akimaru and K. Kawashima, *Teletraffic Theory and Applications*, Springer-Verlag, 1993.
[24] G. Birtwistle, Demos-a system for Discrete Event Modelling on Simula, University of Sheffield, England S1 4DP, 2003.

Shuna Yang was born in Henan Province, China, in 1982. She received the B.S. degree from Hangzhou Dianzi University, Hangzhou, in 2007 and the M.S. degree from Zhejiang University, Hangzhou, in 2010, both in electrical and communication engineering. She is currently pursuing the Ph.D. degree with the Department of telematics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. Her research interests include Optical switching architectures and performance evaluation.

Norvald Stol was born in Karmøy, Norway, in 1960. He received the Siv.ing. and Dr.ing. degrees from the Department of telematics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 1983 and 1996, respectively. He is currently an Associate professor at the Department of telematics, NTNU. His research interests include optical switching, network architectures, and performance and dependability modelling.