Analysis of Opportunistic Spectrum Sharing with Markovian Arrivals and Phase-type Service

Shensheng Tang and Brian L. Mark
Dept. of Electrical and Computer Eng.
George Mason University
4400 University Drive, MS 1G5
Fairfax, VA 22030
tel: 703-993-4069, fax: 703-993-1601
email: { stang1 , bmark }@gmu.edu

NAPL Technical Report
Number: TR-GMU-NAPL-Y08-N2
Date: December 15, 2008

GMU Network Architecture and Performance Laboratory (NAPL)

Abstract

We develop a general framework for analyzing the performance of an opportunistic spectrum sharing (OSS) wireless system at the session level with Markovian arrivals and phase-type service times. The OSS system consists of primary or licensed users of the spectrum and secondary users that sense the channel status and opportunistically share the spectrum resources with the primary users in a coverage area. When a secondary user with an active session detects an arrival of a primary session in its current channel, the secondary user leaves the channel quickly and switches to an idle channel, if one is available, to continue the session. Otherwise, the secondary session is preempted and moved to a preemption queue. The OSS system is modeled by a multi-dimensional Markov process. We derive explicit expressions for the related transition rate matrices using matrix-analytic methods. We also obtain expressions for several performance measures of interest, and present both analytic and simulation results in terms of these performance measures. The proposed OSS model encompasses a large class of specific models as special cases, and should be useful for modeling and performance evaluation of future opportunistic spectrum sharing systems.

Index Terms

Opportunistic spectrum sharing (OSS), Primary users, Secondary users, Multi-dimensional Markov process, MAP, PH distribution.

I. INTRODUCTION

Studies of wireless spectrum usage [1], [2] have shown that large portions of the allocated spectrum are highly underutilized. Frequency agile radios (FARs) are cognitive radios that are capable of detecting idle frequency channels and opportunistically making use of them without causing harmful interference to the primary users [3], [4]. In a scenario of opportunistic spectrum sharing (OSS), the FARs are called *secondary* users and the owners of the allocated spectrum are the *primary* users. By allowing secondary users to reclaim idle channels, much higher spectrum efficiency can be achieved [5]. Opportunistic spectrum sharing techniques offer the potential for higher spectrum reuse in commercial, government, and military applications.

In the OSS system, the spectrum usage of the secondary users is contingent on the requirement that the interference to the primary users must be limited to a certain threshold. A number of opportunistic spectrum access (OSA) schemes have been developed recently in the literature [6]–[10]. In [6], a sensing-based approach is studied for channel selection in spectrum agile communication systems. In [7], a multi-channel OFDMA technique is proposed for OSA networks. In [8], an admission control algorithm in conjunction with a power control scheme is proposed for cognitive wireless networks such that QoS requirements of all admitted secondary users are satisfied. In [9], an adaptive spectrum detection mechanism based on Bayes criterion is proposed for cognitive radio networks in dynamic traffic environments. In [10], collaborative sensing is investigated as a means to improve the performance of sensing-based opportunistic spectrum access in fading channels.

To obtain analytical results for performance evaluation in wireless cellular networks, and even in general communications networks, interarrival and holding time variables are often assumed to be independent and exponentially distributed. However, recent field and simulation studies have suggested that the exponential assumption may not be appropriate for many wireless and cellular systems [11], [12], especially for those based on IP traffic. Internet traffic typically exhibits burstiness and self-similar characteristics at multiple time scales, which cannot be captured by Poisson models. In [13], Poisson models are observed to severely underestimate the burstiness of TCP traffic taken from trace data over a wide range of time scales from the packet level to the session level. In [14], an MMPP (Markov Modulated Poisson Process) traffic model is proposed that mimics the hierarchical behavior of the packet generation process by Internet users. Such Markovian models (cf. [15], [16]) are able to approximately capture the LRD (Long Range Dependence) characteristics of Internet traffic, since the effect of LRD on queueing performance becomes negligible beyond a finite number of time-scales.

In the present paper[1], we analyze the performance an OSS wireless system at the call or session level[2] Call arrivals are modeled by a Markovian arrival process (MAP), which captures the correlation of interarrival times among primary users, among secondary users, as well as between the two types of users. The Markovian arrival process (MAP) has been found to provide a good representation for bursty and correlated traffic arising in modern networks [17]–[22]. The MAP encompasses a rich class of point processes as special cases, including the Poisson process, the MMPP, the PH-renewal process, etc. Channel holding times are modeled using the phase-type (PH) distribution, which can be characterized as the absorption time of a Markov process with one absorbing state. Each state of the Markov process represents one of the phases. The PH distribution generalizes a large class of useful distributions, including the exponential, Erlang, hyper-exponential, hypoexponential, and Coxian distributions. Applications of the PH distribution to modeling service times can be found in [22], [23].

The general OSS system model proposed here is represented by a multi-dimensional Markov process. We derive explicit analytical results for the relevant call level system performance metrics using matrix-analytic methods (cf. [20], [21], [24]). Matrix-analytic methods have also been used for performance analysis of telecommunication systems at the packet level (cf. [25], [26]). The remainder of the paper is organized as follows. Section II describes the OSS system model. Section III develops a multi-dimensional Markovian model and constructs a set of matrix expressions to evaluate the system performance. Section IV develops a recursive solution for the multi-dimensional Markovian model. Special cases of the model are discussed in Section II. Section VI derives several performance measures of interest and discusses their application to network design. Section VII presents numerical results in terms of the obtained performance measures. Finally, Section VIII concludes the paper.

II. System Model

Consider a wireless network operating over a given service area. The network owns the license for spectrum usage and hence is referred to as the *primary system*. The users of this network are the *primary users*. Calls generated by primary users constitute the primary traffic (PT) stream. Next, consider a *secondary* wireless network in the same service area, which opportunistically shares spectrum resource with the primary system. Calls generated by secondary users constitute the secondary traffic (ST) stream. The system consisting of the primary and secondary systems is called an opportunistic spectrum sharing (OSS) system [5]. The proposed OSS system model can be applied to both infrastructured and infrastructureless wireless networks.

In the OSS system, the spectrum availability for the secondary users depends on the spectrum occupancy of the

[1]This work was supported in part by the National Science Foundation under Grants CNS-0520151. Part of this work has been presented in [17].

[2]In this paper, the terms “call” and “session” are used interchangeably.

primary users. A distinct feature of a well-designed OSS system is that the secondary users have the capability to sense channel usage and switch between different channels using appropriate mechanisms, without causing harmful interference to the primary users. This switching between different channels is sometimes called "spectrum mobility" or "spectrum handoff." Such functionality could be realized with the help of cognitive radios [3]. Secondary users detect the presence or absence of signals from primary users and maintain records of the channel occupancy status. The detection mechanism may involve collaboration with other secondary users (cf. [9]) and/or information exchange with a base station (BS) associated with the secondary system.

In the proposed model, we assume a perfect signal detection mechanism[3]. Secondary users opportunistically access the channels that are free. If a secondary call arrives when all channels are occupied, the call is considered to be blocked from the system. On the other hand, when a secondary user in service detects an arrival of a PT call in its current channel, it immediately leaves the channel and switches to an idle channel, if one is available, to continue the call. If at that time all the channels are occupied, the ST call is *preempted* and placed in a queue, which we refer to as the *preemption queue*. The ST call remains in the preemption queue until either the waiting time of the call expires (and the call leaves the system) or a PT/ST call releases a channel. In the latter case, the ST call at the head of the queue immediately occupies the vacated channel. In the primary system, the PT calls operate as if there are no ST calls in the OSS system. When a PT call arrives, it occupies a free channel if one is available; otherwise, it is blocked.

The aggregate arrival process, consisting of PT and ST call arrivals, to the OSS system is modeled by a general MAP, which can capture correlation between interarrival times. The MAP is a generalization of the Poisson process in which the arrivals are governed by an underlying m-state Markov chain. Let g_{ij}^0, $i \neq j$, $1 \leq i, j \leq m$, be the transition rate from state i to state j in the underlying Markov chain without an arrival. Let g_{ij}^P and g_{ij}^S, $1 \leq i, j \leq m$, be the transition rates from state i to state j in the underlying Markov chain with a PT call arrival and an ST call arrival, respectively. The matrix $\mathbf{G}_0 = [g_{ij}^0]$ has nonnegative off-diagonal and negative diagonal elements. Both the matrices $\mathbf{G}_P = [g_{ij}^P]$ and $\mathbf{G}_S = [g_{ij}^S]$ have nonnegative elements. The matrix $\mathbf{G} = \mathbf{G}_0 + \mathbf{G}_P + \mathbf{G}_S$ is the irreducible infinitesimal generator of the m-state Markov chain and the sojourn time in state i is exponentially distributed with parameter λ_i, $1 \leq i \leq m$. At the end of the sojourn in state i, there are three possible transitions (cf. [22]):

- to state j with probability q_{ij}^0, $j \neq i$, $1 \leq i, j \leq m$, without any call arrival;
- to state j with probability q_{ij}^P, $1 \leq i, j \leq m$, with a PT call arrival;
- to state j with probability q_{ij}^S, $1 \leq i, j \leq m$, with a ST call arrival.

[3]Unreliable spectrum sensing is considered in [27], but with Poisson arrivals and exponential service.

For each fixed i, the following relation holds:

$$\sum_{j=1(j\neq i)}^{m} q_{ij}^0 + \sum_{j=1}^{m} q_{ij}^P + \sum_{j=1}^{m} q_{ij}^S = 1. \tag{1}$$

Further, we have $g_{ij}^0 = \lambda_i q_{ij}^0$ for $j \neq i$, $g_{ij}^P = \lambda_i q_{ij}^P$, $g_{ij}^S = \lambda_i q_{ij}^S$ and $g_{ii}^0 = -\lambda_i$, where $1 \leq i, j \leq m$. Note that $(\mathbf{G}_0 + \mathbf{G}_P + \mathbf{G}_S)\mathbf{e} = \mathbf{0}$ holds, where $\mathbf{e}$ is a column vector of 1's.

Let $\boldsymbol{\pi}$ be the stationary probability vector of the generator $\mathbf{G}$. Then we have $\boldsymbol{\pi}\mathbf{G} = 0$ and $\boldsymbol{\pi}\mathbf{e} = \mathbf{1}$. The arrival rate of the MAP is $\lambda_P = \boldsymbol{\pi}\mathbf{G}_P\mathbf{e}$ for PT calls, and $\lambda_S = \boldsymbol{\pi}\mathbf{G}_S\mathbf{e}$ for ST calls. For special cases when $m = 1$, the MAP reduces to a Poisson process with rate λ_1, consisting of two independent Poisson arrival processes with rates $\lambda_1 q_{11}^P$ and $\lambda_1 q_{11}^S$, respectively. When both $\mathbf{G}_P$ and $\mathbf{G}_S$ are diagonal matrices, the MAP reduces to an MMPP, which has been extensively used to describe superposition of data or packetized voice streams [18], [19].

The channel holding times of PT and ST calls, denoted by Z_P and Z_S, respectively, are assumed to follow phase-type (PH) distributions with n phases, denoted by PH$(\boldsymbol{\alpha}_P, \mathbf{T}_P)$ and PH$(\boldsymbol{\alpha}_S, \mathbf{T}_S)$, respectively, where $\boldsymbol{\alpha}_P$ and $\boldsymbol{\alpha}_S$ are row vectors of dimension n, and $\mathbf{T}_P$ and $\mathbf{T}_S$ are square matrices of dimension n [20], [21]. The distribution function of Z_P is given by

$$F_{Z_P}(t) = P\{Z_P \leq t\} = 1 - \boldsymbol{\alpha}_P \exp(\mathbf{T}_P t)\boldsymbol{e}. \tag{2}$$

The exit vector is $\mathbf{T_P^0} = -\mathbf{T}_P\boldsymbol{e}$, where, similarly, $\boldsymbol{e}$ is a column vector of all 1's of appropriate size. If we denote state $n+1$ as the absorbing state with initial probability α_{n+1}^P, then we have $\boldsymbol{\alpha}_P\boldsymbol{e} + \alpha_{n+1}^P = 1$, where $[\boldsymbol{\alpha}_P, \alpha_{n+1}^P]$ is the initial probability vector. Similarly, the distribution function of Z_S is given by

$$F_{Z_S}(t) = P\{Z_S \leq t\} = 1 - \boldsymbol{\alpha}_S \exp(\mathbf{T}_S t)\boldsymbol{e}, \tag{3}$$

with $\mathbf{T_S^0} = -\mathbf{T}_S\boldsymbol{e}$ and $\boldsymbol{\alpha}_S\boldsymbol{e} + \alpha_{n+1}^S = 1$.

As mentioned above, when an ongoing ST call vacates its channel and no other channels are available, it joins the preemption queue. The maximum waiting time of an ST call in the preemption queue, denoted by $\Psi_{\max}$, is assumed to follow a PH distribution with r phases, denoted by PH$(\boldsymbol{\theta}, \boldsymbol{\Theta})$, where $\boldsymbol{\theta}$ is a row vector of dimension r, and $\boldsymbol{\Theta}$ is a square matrix of dimension r. The distribution function of $\Psi_{\max}$ is given by

$$F_{\Psi_{\max}}(t) = P\{\Psi_{\max} \leq t\} = 1 - \boldsymbol{\theta} \exp(\boldsymbol{\Theta} t)\boldsymbol{e}, \tag{4}$$

with $\boldsymbol{\Theta^0} = -\boldsymbol{\Theta}\mathbf{e}$ and $\boldsymbol{\theta}\boldsymbol{e} + \theta_{r+1} = 1$.

III. Analysis and Matrix Construction

We define a set of tuples

$$\mathcal{S}^- \triangleq \{(i, i_p, u, \boldsymbol{S}_{i_P}, \boldsymbol{S}_{i_S}, \boldsymbol{S}_{i_Q})\},$$

where $1 \leq i \leq 2M$; $0 \leq i_P \leq M$; $1 \leq u \leq m$. The element $\boldsymbol{S}_{i_P}$ is defined to be empty if $i_P = 0$; otherwise, $\boldsymbol{S}_{i_P} \triangleq (S_1^P, S_2^P, ..., S_{i_P}^P)$ with $1 \leq S_j^P \leq n, 1 \leq j \leq i_P$. Define $\widehat{i_M} \triangleq \min\{i, M\}$. The element $\boldsymbol{S}_{i_S}$ is defined to be empty if $\widehat{i_M} - i_P = 0$; otherwise, $\boldsymbol{S}_{i_S} \triangleq (S_1^S, S_2^S, ..., S_{\widehat{i_M}-i_P}^S)$, with $1 \leq S_k^S \leq n, 1 \leq k \leq \widehat{i_M} - i_P$. Finally, the element $\boldsymbol{S}_{i_Q}$ is defined to be empty if $i - \widehat{i_M} = 0$; otherwise, $\boldsymbol{S}_{i_Q} \triangleq (S_1^Q, S_2^Q, ..., S_{i-\widehat{i_M}}^Q)$, with $1 \leq S_l^Q \leq n, 1 \leq l \leq i - \widehat{i_M}$. We then consider a stochastic process $\{X(t) : t \geq 0\}$, defined on the state space $\mathcal{S} = \{(0, u) : 1 \leq u \leq m\} \cup \mathcal{S}^-$.

The state $(0, u)$ represents the state with no call in the system and the arrival process is in phase u. The element i represents the total number of calls in service and preempted ST calls in the preemption queue. Since the state $(0, u)$ represents an empty system, the state $(i, i_p, u, \boldsymbol{S}_{i_P}, \boldsymbol{S}_{i_S}, \boldsymbol{S}_{i_Q})$ implies that there is at least one call in the system. When an ST node detects an arrival of PT call in its current channel and at that time all the channels are occupied, the ST call becomes a *preempted ST call* and moves to the preemption queue. Clearly, the maximum number of preempted ST calls is M, which corresponds to the limiting case that all the M ongoing calls are ST calls and are eventually preempted to the preemption queue due to the arrivals of PT calls. Thus, we have $1 \leq i \leq 2M$.

In state $(i, i_P, u, \boldsymbol{S}_{i_P}, \boldsymbol{S}_{i_S}, \boldsymbol{S}_{i_Q})$, there are totally i calls in the system (including the ST calls in the preemption queue), $\widehat{i_M}$ calls in service with i_P calls being PT calls, and $i - \widehat{i_M}$ ST calls in the queue. The arrival process is in phase u, the jth ($0 \leq j \leq i_P$) PT call among the i_P PT calls is being served in phase S_j^P; the kth ($1 \leq k \leq \widehat{i_M} - i_P$) ST call among the $\widehat{i_M} - i_P$ ST calls is being served in phase S_k^S, and the lth ($1 \leq l \leq i - \widehat{i_M}$) preempted ST call among the $i - \widehat{i_M}$ preempted ST calls in the preemption queue was served in phase S_l^Q immediately before it was preempted. Note that in state $(i, i_p, u, \boldsymbol{S}_{i_P}, \boldsymbol{S}_{i_S}, \boldsymbol{S}_{i_Q})$ with $i = 2M$, the element $\boldsymbol{S}_{i_S}$ is empty since in this case $i_P = M$ (all M channels are occupied by PT calls). When $1 \leq i \leq M$, $\boldsymbol{S}_{i_Q}$ is empty since in this case there are no preempted ST calls in the system. One can show that $\{X(t)\}$ is a continuous-time Markov process with infinitesimal generator

$$\mathbf{Q} = \begin{bmatrix} \mathbf{E}_0 & \mathbf{B}_0 & & & & & \\ \mathbf{D}_1 & \mathbf{E}_1 & \mathbf{B}_1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & & \mathbf{D}_{2M-1} & \mathbf{E}_{2M-1} & \mathbf{B}_{2M-1} \\ & & & & \mathbf{D}_{2M} & \mathbf{E}_{2M} \end{bmatrix}, \tag{5}$$

where $\mathbf{E}_i$ ($0 \leq i \leq 2M$) is a matrix representing the absence of transitions from the state in which there are i calls in the system; $\mathbf{B}_i$ ($0 \leq i \leq 2M - 1$) is a matrix representing the transition rates due to the arrival of a PT or ST call when there are i calls in the system; and $\mathbf{D}_i$ ($1 \leq i \leq 2M$) denotes a departure of a call when there are i calls in the system.

Next, we introduce the following notations (cf. [22]):

- $\mathbf{I}_k$ is an identity matrix of dimension k and $\mathbf{I}(k,s) = \underbrace{\mathbf{I}_k \otimes \cdots \otimes \mathbf{I}_k}_{s}$ $(k, s = 1, 2, \cdots)$ is the Kronecker product of s identity matrices $\mathbf{I}_k$; $\mathbf{I}(k,0) \triangleq 1$.
- $\mathbf{W}_P(s) = \underbrace{\mathbf{T}_P \oplus \cdots \oplus \mathbf{T}_P}_{s}$ is the Kronecker sum of s square matrices $\mathbf{T}_P$, which represents the service phases of the s PT calls that are in service; $\mathbf{W}_P(0) \triangleq 0$.
- $\mathbf{W}_S(s) = \underbrace{\mathbf{T}_S \oplus \cdots \oplus \mathbf{T}_S}_{s}$ is the Kronecker sum of s square matrices $\mathbf{T}_S$, which represents the service phases of the s ST calls that are in service; $\mathbf{W}_S(0) \triangleq 0$.
- $\mathbf{W}_Q(s) = \underbrace{\mathbf{T}_S \oplus \cdots \oplus \mathbf{T}_S}_{s}$ is the Kronecker sum of s square matrices $\mathbf{T}_S$, which represents the service phases of the s preempted ST calls immediately before they are preempted; $\mathbf{W}_Q(0) \triangleq 0$. Note that $\mathbf{W}_Q(s)$ has the same form as $\mathbf{W}_S(s)$.
- $\mathbf{V}_P(s) = \sum_{k=0}^{s-1} \mathbf{I}(n,k) \otimes \mathbf{T_P^0} \otimes \mathbf{I}(n, s-k-1)$ represents service completion of one of the s PT calls in service.
- $\mathbf{V}_S(s) = \sum_{k=0}^{s-1} \mathbf{I}(n,k) \otimes \mathbf{T_S^0} \otimes \mathbf{I}(n, s-k-1)$ represents service completion of one of the s ST calls in service.

Expressions for these matrices are derived below.

A. Construction of Block Matrices $\mathbf{B}_i$

If $0 \le i \le M-1$, $\mathbf{B}_i$ is an $(i+1) \times (i+2)$ block matrix given by

$$\mathbf{B}_i = \begin{bmatrix} \mathbf{B}_{i,0}^P & \mathbf{B}_{i,0}^S & & & \\ & \mathbf{B}_{i,1}^P & \mathbf{B}_{i,1}^S & & \\ & & \ddots & \ddots & \\ & & & \mathbf{B}_{i,i}^P & \mathbf{B}_{i,i}^S \end{bmatrix}, \tag{6}$$

where $\mathbf{B}_{i,j}^P = \mathbf{G}_P \otimes \mathbf{I}(n,j) \otimes \boldsymbol{\alpha}_P \otimes \mathbf{I}(n, i-j)$, $0 \le j \le i$, represents the transition rates corresponding to the arrival of a PT call when there are i calls in the system, among which j are PT calls; and $\mathbf{B}_{i,j}^S = \mathbf{G}_S \otimes \mathbf{I}(n,j) \otimes \mathbf{I}(n, i-j) \otimes \boldsymbol{\alpha}_S$, $0 \le j \le i$, represents transition rates corresponding to the arrival of an ST call when there are i calls in the system, among which j are PT calls. Clearly, when $i = 0$, we have $\mathbf{B}_0 = [\mathbf{G}_P \otimes \boldsymbol{\alpha}_P \quad \mathbf{G}_S \otimes \boldsymbol{\alpha}_S]$.

If $M \le i \le 2M-1$, $\mathbf{B}_i$ is a $(2M-i+1) \times (2M-i)$ block matrix given by[4]

[4]Note that in the matrix $\mathbf{B}_i$ and the following $\mathbf{D}_i$, $\mathbf{0}$ denotes a submatrix with the same size to its associated $B_{i,j}^P$ and $D_{i,j}^P$ (or $D_{i,j}^S$), respectively.

$$\mathbf{B}_i = \begin{bmatrix} \mathbf{B}^P_{i,i-M} & & & \\ & \mathbf{B}^P_{i,i-M+1} & & \\ & & \ddots & \\ & & & \mathbf{B}^P_{i,M-1} \\ & & & \mathbf{0} \end{bmatrix}, \tag{7}$$

where $\mathbf{B}^P_{i,j} = \mathbf{G}_P \otimes \mathbf{I}(n,j) \otimes \boldsymbol{\alpha}_P \otimes \mathbf{I}(n,i-j)$, $i-M \le j \le M-1$, represents state transition rates corresponding to the arrival of a PT call when there are i calls in the system (including the preempted ST calls in the preemption queue), among which j are PT calls.

B. Construction of Block Matrices $\mathbf{D}_i$

If $1 \le i \le M$, $\mathbf{D}_i$ is an $(i+1) \times i$ block matrix given by

$$\mathbf{D}_i = \begin{bmatrix} \mathbf{D}^S_{i,0} & & & \\ \mathbf{D}^P_{i,1} & \mathbf{D}^S_{i,1} & & \\ & \ddots & \ddots & \\ & & \mathbf{D}^P_{i,i-1} & \mathbf{D}^S_{i,i-1} \\ & & & \mathbf{D}^P_{i,i} \end{bmatrix}, \tag{8}$$

where in each row of $\mathbf{D}_i$, the element $\mathbf{D}^P_{i,j} = \mathbf{I}_m \otimes \mathbf{V}_P(j) \otimes \mathbf{I}(n,i-j)$, $1 \le j \le i$, represents transition rates corresponding to a PT call departure when there are i calls in the system, among which j are PT calls; and the element $\mathbf{D}^S_{i,j} = \mathbf{I}_m \otimes \mathbf{I}(n,j) \otimes \mathbf{V}_S(i-j)$, $0 \le j \le i-1$, represents transition rates corresponding to the departure of an ST call when there are i calls in the system, among which j are PT calls. Clearly, when $i = 1$, we have

$$\mathbf{D}_1 = \begin{bmatrix} \mathbf{I}_m \otimes \mathbf{T}^{\mathbf{0}}_{\mathbf{S}} \\ \mathbf{I}_m \otimes \mathbf{T}^{\mathbf{0}}_{\mathbf{P}} \end{bmatrix}.$$

If $M+1 \le i \le 2M$, $\mathbf{D}_i$ is a $(2M-i+1) \times (2M-i+2)$ block matrix given by

$$\mathbf{D}_i = \begin{bmatrix} \mathbf{D}^P_{i,i-M} & \mathbf{D}^S_{i,i-M} & & & \\ & \mathbf{D}^P_{i,i-M+1} & \mathbf{D}^S_{i,i-M+1} & & \\ & & \ddots & \ddots & \\ & & & \mathbf{D}^P_{i,M-1} & \mathbf{D}^S_{i,M-1} \\ & & & & \mathbf{D}^P_{i,M} \quad \mathbf{0} \end{bmatrix}, \tag{9}$$

where in each row of $\mathbf{D}_i$, the element $\mathbf{D}^P_{i,j} = \mathbf{I}_m \otimes \mathbf{V}_P(j) \otimes \mathbf{I}(n,i-j)$, $i-M \le j \le M$, represents state transition rates corresponding to the departure of a PT call when there are i calls in the system with M calls receiving service, among which j are PT calls; and the element $\mathbf{D}^S_{i,j} = \mathbf{I}_m \otimes \mathbf{I}(n,j) \otimes \mathbf{V}_S(i-j)$, $i-M \le j \le M-1$, represents transition rates corresponding to the departure of an ST call when there are i calls in the system with M calls receiving service, among which j are PT calls.

C. Construction of Block Matrices $\mathbf{E}_i$

If $0 \le i \le M$, $\mathbf{E}_i$ is an $(i+1) \times (i+1)$ block diagonal matrix given by

$$\mathbf{E}_i = \text{diag}\{\mathbf{E}_{i,0}, \mathbf{E}_{i,1}, \ldots, \mathbf{E}_{i,i}\}. \tag{10}$$

where $\mathbf{E}_{i,j}$, $0 \le j \le i$, represents the absence of state transitions when there are i calls receiving service, among which j are PT calls, and is given by

$$\mathbf{E}_{i,j} = \begin{cases} \mathbf{G}_0 \oplus \mathbf{W}_P(j) \oplus \mathbf{W}_S(i-j), & 0 \le i \le M-1,\ 0 \le j \le i; \\ (\mathbf{G}_0 + \mathbf{G}_S) \oplus \mathbf{W}_P(j) \oplus \mathbf{W}_S(i-j), & i = M,\ 0 \le j \le M-1; \\ (\mathbf{G}_0 + \mathbf{G}_S + \mathbf{G}_P) \oplus \mathbf{W}_P(j) \oplus \mathbf{W}_S(i-j), & i = M,\ j = M. \end{cases} \tag{11}$$

Clearly, when $i = 0$, we have $\mathbf{E}_0 = \mathbf{E}_{0,0} = \mathbf{G}_0$. If $M+1 \le i \le 2M$, $\mathbf{E}_i$ is a $(2M-i+1) \times (2M-i+1)$ block diagonal matrix given by

$$\mathbf{E}_i = \text{diag}\{\mathbf{E}_{i,i-M}, \mathbf{E}_{i,i-M+1}, \ldots, \mathbf{E}_{i,M}\}, \tag{12}$$

where $\mathbf{E}_{i,j}$, $i - M \le j \le M$, represents the absence of state transitions when there are i calls in the system with M calls receiving service, among which at least j calls are PT calls, and is given by

$$\mathbf{E}_{i,j} = \begin{cases} (\mathbf{G}_0 + \mathbf{G}_S) \oplus \mathbf{W}_P(j) \oplus \mathbf{W}_S(M-j) \oplus \mathbf{W}_Q(i-M), & M+1 \le i \le 2M-1,\ i-M \le j \le M-1; \\ (\mathbf{G}_0 + \mathbf{G}_S + \mathbf{G}_P) \oplus \mathbf{W}_P(j) \oplus \mathbf{W}_S(M-j) \oplus \mathbf{W}_Q(i-M), & M+1 \le i \le 2M,\ j = M. \end{cases} \tag{13}$$

IV. Computation of Stationary State Probability Vector

In this section, we derive the stationary state probability vector of the Markov process $\{X(t)\}$ and provide a recursive computational algorithm. Let $p(0,u)$ and $p(i, i_P, u, \boldsymbol{S}_{i_P}, \boldsymbol{S}_{i_S}, \boldsymbol{S}_{i_Q})$ denote the stationary probability of the system in states $(0,u)$ and $(i, i_P, u, \boldsymbol{S}_{i_P}, \boldsymbol{S}_{i_S}, \boldsymbol{S}_{i_Q})$, respectively. Let $\boldsymbol{p}_i$, $0 \le i \le 2M$, be the stationary state probability vector of the system in equilibrium when there are i calls in the system. The sequence of elements in the vectors $\boldsymbol{p}_i$, $0 \le i \le 2M$, is ordered lexicographically based on the probabilities $p(0,u)$ and $p(i, i_P, u, \boldsymbol{S}_{i_P}, \boldsymbol{S}_{i_S}, \boldsymbol{S}_{i_Q})$. For example, when $i = 0$, $\boldsymbol{p}_0 = (p(0,1), p(0,2), \cdots, p(0,m))$. Thus, the stationary probability vector of the system is $\boldsymbol{P} = (\boldsymbol{p}_0, \boldsymbol{p}_1, \cdots, \boldsymbol{p}_M, \boldsymbol{p}_{M+1}, \cdots, \boldsymbol{p}_{2M})$, where $\boldsymbol{p}_i$, $0 \le i \le M$, is a probability vector of level i with dimension $(i+1)mn^i$, and $\boldsymbol{p}_i$, $M+1 \le i \le 2M$, is a probability vector of level i with dimension $(2M-i+1)mn^i$. From the equilibrium conditions $\boldsymbol{PQ} = \boldsymbol{0}$ (where $\boldsymbol{0}$ is a row vector of all zeros of appropriate dimension) and $\boldsymbol{Pe} = 1$, we have

$$\begin{aligned} &\boldsymbol{p}_0\mathbf{E}_0 + \boldsymbol{p}_1\mathbf{D}_1 = \boldsymbol{0}; \quad \boldsymbol{p}_{2M-1}\mathbf{B}_{2M-1} + \boldsymbol{p}_{2M}\mathbf{E}_{2M} = \boldsymbol{0}; \\ &\qquad \boldsymbol{p}_{i-1}\mathbf{B}_{i-1} + \boldsymbol{p}_i\mathbf{E}_i + \boldsymbol{p}_{i+1}\mathbf{D}_{i+1} = \boldsymbol{0}, \quad 1 \le i \le 2M-1. \end{aligned} \tag{14}$$

Solving the above equations, we obtain the following recurrence formula:

$$\boldsymbol{p}_{2M}\mathbf{C}_{2M} = 0; \quad \boldsymbol{p}_i = \boldsymbol{p}_{i+1}\mathbf{D}_{i+1}(\mathbf{C}_i)^{-1}, \quad 0 \leq i \leq 2M-1, \tag{15}$$

where

$$\mathbf{C}_0 = -\mathbf{E}_0, \quad \text{and} \quad \mathbf{C}_i = -\mathbf{E}_i - \mathbf{D}_i(\mathbf{C}_{i-1})^{-1}\mathbf{B}_{i-1}, \quad 1 \leq i \leq 2M. \tag{16}$$

We point out that the recursive solution approaches used in [22], [28], [29] are not applicable here since the block-matrix structures are different. Following the above recursive equations, the stationary probability vector $\boldsymbol{P}$ can be computed numerically. We summarize the algorithm for computing the stationary state probability vector as follows:

- **Step 1**: Compute the matrices $\mathbf{C}_i$ recursively from $i = 0$ to $i = 2M$, by using (16).
- **Step 2**: Compute probability vectors $\boldsymbol{p}_i$ recursively from $i = 2M$ to $i = 0$, by using (15).
- **Step 3**: Normalize the probability vector $\boldsymbol{p}_i$, $0 \leq i \leq 2M$, by using $\boldsymbol{P} \leftarrow \boldsymbol{P}^* = \frac{\boldsymbol{P}}{\boldsymbol{Pe}}$. The obtained vector $\boldsymbol{P}^*$ is the final stationary probability vector.

Remark 1: The proposed performance model can encompass a large class of specific models as special cases, such as a model with MAP arrivals and exponentially distributed service, a model with MMPP arrivals and Erlang-distributed service, a model with Poisson arrivals and hypoexponential service, etc. In the first case, the service time parameters become (cf. [17]): $n = 1$, $\boldsymbol{\alpha}_P = 1$, $\mathbf{T}_P = -\mu_P$, $\mathbf{T_P^0} = \mu_P$, $\boldsymbol{\alpha}_S = 1$, $\mathbf{T}_S = -\mu_S$, $\mathbf{T_S^0} = \mu_S$.

Remark 2: The computational complexity of the model is mainly due to computing the matrix inverse in (15) and (16). Since the complexity of matrix inversion is, in general, $O(i^3)$ for an $i \times i$ matrix [30], the complexity of the recursive algorithm given above is given by

$$O\left(\sum_{i=0}^{M}(i+1)^3m^3n^{3i} + \sum_{i=M+1}^{2M}(2M-i+1)^3m^3n^{3i}\right) = O(M^3m^3n^{6M}). \tag{17}$$

In particular, when $n = 1$, the complexity of the model is $O(M^3m^3)$. In this case, the model can be solved feasibly when M is on the order of a few hundred channels and m is on the order of ten. When $n > 1$, the computational complexity grows exponentially with the number of channels M and only models with moderate values for n and M (e.g., $n = 2$ and M on the order of ten) can be solved in practice using today's computers. The complexity could be reduced by exploiting the block structure of the system generator matrix in the matrix inversion step would have complexity less than $O(i^3)$, but the overall complexity would still be exponential in M when $n > 1$. The only way to reduce this factor would be to simplify the model in some way, thereby trading off model accuracy for reduced computational complexity. Investigation of such approximations is an interesting issue, but beyond the scope of the present paper.

V. Special Cases

The proposed framework can encompass a large class of specific frameworks as special cases, such as modeling a system by considering Poisson arrival process and exponentially distributed service times, the MMPP arrival process and Erlang-distributed service times, MAP arrival process and hypoexponential service times, etc. Here we present two simple special cases: (A) Poisson call arrivals and exponential exponential service; (B) MAP call arrivals and exponential service.

A. Poisson arrivals and exponential service

This is the simplest model class in the proposed generic framework. In this case, we have $m = 1$, $\lambda_P = \lambda_1 q_{11}^P$, $\lambda_S = \lambda_1 q_{11}^S$, $G_0 = -(\lambda_P + \lambda_S)$, $\mathbf{G}_P = \lambda_P$, and $\mathbf{G}_S = \lambda_S$. The MAP is reduced to a Poisson process with rate $\lambda_P + \lambda_S$, which consists of two independent Poisson processes with PT call arrival rate λ_P and ST call arrival rate λ_S. On the other hand, we have $n = 1$, $\boldsymbol{\alpha}_P = 1$, $T_P = -\mu_P$, $\mathbf{T_P^0} = \mu_P$, $\boldsymbol{\alpha}_S = 1$, $T_S = -\mu_S$, $\mathbf{T_S^0} = \mu_S$. The PH-distributions are reduced to exponential distributions with parameters μ_P and μ_S, respectively.

From the previous construction of matrices in Section III, we can simplify B_i, E_i and D_i as follows: If $0 \le i \le M-1$, B_i is an $(i+1) \times (i+2)$ matrix given by

$$b_i = [\lambda_P I_{i+1} \quad \mathbf{0}] + [\mathbf{0} \quad \lambda_S I_{i+1}]. \tag{18}$$

If $M \le i \le 2M-1$, B_i is an $(2M-i+1) \times (2M-i)$ matrix given by

$$B_i = \lambda_P \begin{bmatrix} I_{2M-i} \\ \mathbf{0} \end{bmatrix}. \tag{19}$$

If $1 \le i \le M$, D_i is an $(i+1) \times i$ matrix given by

$$D_i = \begin{bmatrix} \mu_S \text{diag}\{i, i-1, \ldots, 1\} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mu_P \text{diag}\{1, \ldots, i-1, i\}\} \end{bmatrix}, \tag{20}$$

If $M+1 \le i \le 2M$, D_i is a $(2M-i+1) \times (2M-i+2)$ matrix given by

$$D_i = [\mu_P \text{diag}\{i-M, \ldots, M-1, M\}, \mathbf{0}] + [\mathbf{0} \quad \mu_S \text{diag}\{2M-i, 2M-i-1, \ldots, 1\} \quad \mathbf{0}]. \tag{21}$$

If $0 \le i \le M$, E_i is an $(i+1) \times (i+1)$ matrix given by

$$E_i = \text{diag}[E_{i,0}, E_{i,1}, \ldots, E_{i,i}], \tag{22}$$

where $E_{i,j}$, $0 \le j \le i$, is given by

$$E_{i,j} = \begin{cases} -[\lambda_P + \lambda_S + j\mu_P + (i-j)\mu_S], & 0 \le i \le M-1,\ 0 \le j \le i; \\ -[\lambda_P + j\mu_P + (i-j)\mu_S], & i = M,\ 0 \le j \le M-1; \\ -M\mu_P, & i = M,\ j = M. \end{cases} \tag{23}$$

If $M+1 \leq i \leq 2M$, E_i is a $(2M-i+1) \times (2M-i+1)$ matrix given by

$$E_i = \text{diag}[E_{i,i-M}, E_{i,i-M+1}, \ldots, E_{i,i-M+1}], \tag{24}$$

where $E_{i,j}$, $i-M \leq j \leq M$, is given by

$$E_{i,j} = \begin{cases} -[\lambda_P + j\mu_P + (M-j)\mu_S], & M+1 \leq i \leq 2M-1,\ i-M \leq j \leq M-1; \\ -M\mu_P, & M+1 \leq i \leq 2M,\ j = M. \end{cases} \tag{25}$$

By using the proposed recursive computational algorithm, the stationary probability vector of this special case can be easily solved.

B. MAP arrivals and exponential service

In this case, only service times are simplified as $n=1$, $\boldsymbol{\alpha}_P = 1$, $T_P = -\mu_P$, $\mathbf{T_P^0} = \mu_P$, $\boldsymbol{\alpha}_S = 1$, $T_S = -\mu_S$, $\mathbf{T_S^0} = \mu_S$. From the previous construction of matrices in Section III, we can simplify B_i, E_i and D_i and solve the stationary probability vector. We omit the details here, readers with interests are referred to [17].

VI. Performance Measures

After obtaining the stationary state probability vector, we can determine various performance measures of interest.

A. Blocking Probabilities

The ST blocking probability, denoted by B_S, is defined as the probability that all M channels are occupied by either PT or ST sessions and is given by

$$B_S = \sum_{i=M}^{2M} \boldsymbol{p}_i \boldsymbol{e}. \tag{26}$$

The PT blocking probability, denoted by B_P, is defined as the probability that all M channels are occupied by PT sessions and is given by

$$B_P = \sum_{i=M}^{2M} \boldsymbol{p}_i|_{i_P=M} \boldsymbol{e}, \tag{27}$$

where the steady-state probability vector $\boldsymbol{p}_i|_{i_P=M}$ represents the value of $\boldsymbol{p}_i$ given that $i_P = M$.

B. Total Channel Utilization and Carried Traffic

The total channel utilization, denoted by η, is defined as the ratio of the mean number of occupied channels to the total number of channels. We find that

$$\eta = \frac{1}{M} \sum_{i=1}^{2M} \widehat{i_M} \boldsymbol{p}_i \boldsymbol{e}. \tag{28}$$

The carried PT load, N_P, is obtained as

$$N_P = \sum_{i=1}^{M} \sum_{i_P=0}^{i} i_P \boldsymbol{p}_i \boldsymbol{e} + \sum_{i=M+1}^{2M} \sum_{i_P=i-M}^{M} i_P \boldsymbol{p}_i \boldsymbol{e} = \sum_{i=1}^{2M} i_P \boldsymbol{p}_i \boldsymbol{e}. \tag{29}$$

The last equation is obtained using the fact that number of PT calls i_P can only be a number between 0 and i when $1 \leq i \leq M$, and between $i - M$ and M when $M + 1 \leq i \leq 2M$, which are implicitly included in the structure of the steady-state probability vector $\boldsymbol{p}_i$. For example, when $i = 2M$, i_P can only be M; when $i = 2M - 1$, i_P can only be $M - 1$ or M. Similarly, the carried ST load, N_S, is obtained as

$$N_S = \sum_{i=1}^{M} \sum_{i_P=0}^{i} (i - i_P) \boldsymbol{p}_i \boldsymbol{e} + \sum_{i=M+1}^{2M} \sum_{i_P=i-M}^{M} (M - i_P) \boldsymbol{p}_i \boldsymbol{e} = \sum_{i=1}^{2M} [\widehat{i_M} - i_P] \boldsymbol{p}_i \boldsymbol{e}. \tag{30}$$

The total carried traffic is given by $\mathrm{TCT} = N_P + N_S = \sum_{i=1}^{2M} \widehat{i_M} \boldsymbol{p}_i \boldsymbol{e} = M\eta$.

C. Mean Number of Preempted ST Calls and Preemption Ratio

The mean number of preempted ST calls in equilibrium, which can be used to evaluate the performance of the secondary system and design an appropriate range of system parameters is given by $N_Q = \sum_{i=M+1}^{2M} (i - M) \boldsymbol{p}_i \boldsymbol{e}$. As mentioned earlier, when an ST call that vacates its channel cannot find an idle channel, it is moved to the preemption queue. The mean preemption ratio of the ongoing ST calls in equilibrium is defined as the ratio of the mean number of preempted ST calls to the mean total number of ST calls in the system, i.e., $\gamma = \frac{N_Q}{N_{SQ}}$, where

$$N_{SQ} = N_S + N_Q = \sum_{i=1}^{M} \sum_{i_P=0}^{i} (i - i_P) \boldsymbol{p}_i \boldsymbol{e} + \sum_{i=M+1}^{2M} \sum_{i_P=i-M}^{M} (i - i_P) \boldsymbol{p}_i \boldsymbol{e} = \sum_{i=1}^{2M} (i - i_P) \boldsymbol{p}_i \boldsymbol{e}. \tag{31}$$

D. Perceived and Actual Waiting Times

As mentioned in Section II, preempted ST calls remaining in the preemption queue either reconnect back in first-come first-served (FCFS) order as channels become available or leave the system when their maximum waiting times expire. Consider a given preempted ST call, referred to as the *test call*. The probability that l preempted ST calls are waiting in the preemption queue, upon the arrival of the test call, is given by $\boldsymbol{p}_i|_{i=M+l} \boldsymbol{e}$, $0 \leq l \leq M - 1$. Note that the condition $0 \leq l \leq M - 1$ implies $M \leq i \leq 2M - 1$. The test call reconnects to the system only if $l + 1$ calls leave the system, either releasing a channel or a position in the queue, before its maximum waiting time expires.

Let φ_l, $0 \leq l \leq M-1$, denote the time interval, in steady-state, between a transition to a state $(i, i_p, u, \boldsymbol{S}_{i_P}, \boldsymbol{S}_{i_S}, \boldsymbol{S}_{i_Q}(l+1))$ until a transition to a new state $(i - 1, i'_p, u', \boldsymbol{S}'_{i_P}, \boldsymbol{S}'_{i_S}, \boldsymbol{S}_{i_Q}(l))$, due to either the service completion of an ongoing PT or ST call, or the departure of a queued ST call due to expiry of its maximum waiting time, where $\boldsymbol{S}_{i_Q}(l) \triangleq (S_1^Q, S_2^Q, ..., S_l^Q)$ with $1 \leq S_k^Q \leq n, 1 \leq k \leq l$. If a PT call leaves, then $i'_p = i_p - 1$, while the elements in $\boldsymbol{S}_{i_S}$ remain the same; if an ongoing ST call leaves, then the elements in $\boldsymbol{S}_{i_S}$ are reduced by 1 while $i'_p = i_p$; if

a queued ST call leaves, then both i_P and the elements in $\boldsymbol{S}_{i_S}$ remain the same. When a PT or ST call leaves the system, the head-of-line ST call in the preemption queue reconnects to the system and the remaining queued ST calls advance by one position in the queue. Similarly, the dropping of a queued ST call leads to the advancement, by one position, of each of the remaining queued ST calls that were behind it.

We define a new performance metric, *perceived waiting time of the test call with queue size l*, defined by

$$\Psi_{\mathrm{per}}(l) \triangleq \varphi_0 + \varphi_1 + \cdots + \varphi_l,$$

to represent the queueing behavior of the preempted ST calls. Recall that the maximum waiting time of a ST call in the preemption queue $\Psi_{\max}$ is assumed to follow a PH distribution with dimension r and is represented by PH$(\boldsymbol{\theta}, \boldsymbol{\Theta})$. To compute the perceived waiting time of preempted ST calls $\Psi_{\mathrm{per}}(l)$, the residual channel holding times of the ongoing PT and ST calls must be determined. By [21, theorem 2.2.3], the residual channel holding times of the ongoing PT and ST calls, denoted by Y_P and Y_S, follow PH-distributions with dimension n, represented by PH$(\boldsymbol{\beta}_P, \mathbf{T}_P)$ and PH$(\boldsymbol{\beta}_S, \mathbf{T}_S)$, respectively, where $\boldsymbol{\beta}_P = \boldsymbol{\beta}_P(\mathbf{T}_P + \mathbf{T}_\mathbf{P}^\mathbf{0}\boldsymbol{\alpha}_P)$ and $\boldsymbol{\beta}_P\boldsymbol{e} = 1$ and also $\boldsymbol{\beta}_S = \boldsymbol{\beta}_S(\mathbf{T}_S + \mathbf{T}_\mathbf{S}^\mathbf{0}\boldsymbol{\alpha}_S)$ and $\boldsymbol{\beta}_S\boldsymbol{e} = 1$. The following theorem characterizes the perceived waiting time distribution (a proof is provided Appendix A).

Theorem 1: The perceived waiting time of the test call given that the preemption queue is of size l, $\Psi_{\mathrm{per}}(l)$, follows a PH distribution of order $\frac{n^M(r^{l+1}-1)}{r-1}$, represented as PH$(\boldsymbol{\omega}_{\mathrm{per}}(l), \boldsymbol{\Omega}_{\mathrm{per}}(l))$, where M is the number of channels shared by the PT and ST calls, and

$$\boldsymbol{\omega}_{\mathrm{per}}(l) = \left[\boldsymbol{a}_0, (1-\boldsymbol{a}_0\boldsymbol{e})\boldsymbol{a}_1, (1-\boldsymbol{a}_0\boldsymbol{e})(1-\boldsymbol{a}_1\boldsymbol{e})\boldsymbol{a}_2, \cdots, \prod_{u=0}^{l-1}(1-\boldsymbol{a}_u\boldsymbol{e})\boldsymbol{a}_l\right], \tag{32}$$

$$\boldsymbol{\Omega}_{\mathrm{per}}(l) = \begin{bmatrix} \mathbf{A}_{00} & \mathbf{A}_{01} & \mathbf{A}_{02} & \mathbf{A}_{03} & \cdots & \mathbf{A}_{0,l-1} & \mathbf{A}_{0l} \\ & \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} & \cdots & \mathbf{A}_{1,l-1} & \mathbf{A}_{1l} \\ & & \mathbf{A}_{22} & \mathbf{A}_{23} & \cdots & \mathbf{A}_{2,l-1} & \mathbf{A}_{2l} \\ & & & & \ddots & \vdots & \vdots \\ & & & & & \mathbf{A}_{l-1,l-1} & \mathbf{A}_{l-1,l} \\ & & & & & & \mathbf{A}_{l,l} \end{bmatrix}, \tag{33}$$

where

$$\boldsymbol{a}_0 = [\underbrace{\boldsymbol{\beta}_P \otimes \cdots \otimes \boldsymbol{\beta}_P}_{i_P} \otimes \underbrace{\boldsymbol{\beta}_S \otimes \cdots \otimes \boldsymbol{\beta}_S}_{M-i_P}], \tag{34}$$

$$\boldsymbol{a}_u = [\underbrace{\boldsymbol{\beta}_P \otimes \cdots \otimes \boldsymbol{\beta}_P}_{i_P} \otimes \underbrace{\boldsymbol{\beta}_S \otimes \cdots \otimes \boldsymbol{\beta}_S}_{M-i_P} \otimes \underbrace{\boldsymbol{\theta} \otimes \cdots \otimes \boldsymbol{\theta}}_{u}], \quad 1 \le u \le l, \tag{35}$$

$$\boldsymbol{A}_{00} = \underbrace{\mathbf{T}_P \oplus \cdots \oplus \mathbf{T}_P}_{i_P} \oplus \underbrace{\mathbf{T}_S \oplus \cdots \oplus \mathbf{T}_S}_{M-i_P}, \tag{36}$$

$$\boldsymbol{A}_{uu} = \underbrace{\mathbf{T}_P \oplus \cdots \oplus \mathbf{T}_P}_{i_P} \oplus \underbrace{\mathbf{T}_S \oplus \cdots \oplus \mathbf{T}_S}_{M-i_P} \oplus \underbrace{\boldsymbol{\Theta} \oplus \cdots \oplus \boldsymbol{\Theta}}_{u}, \quad 1 \le u \le l, \tag{37}$$

$$\boldsymbol{A}_{uv} = -\sum_{k=u}^{v-1} (\boldsymbol{A}_{uk} \cdot \boldsymbol{e}) \boldsymbol{a}_v, \quad 0 \le u \le l-1,\ u+1 \le v \le l, \tag{38}$$

and i_P is the number of PT calls in service, and u is the number of preempted ST calls in the preemption queue. Note that matrices $\boldsymbol{A}_{uv}$ can be obtained recursively from $\boldsymbol{A}_{u,u+1}$ to $\boldsymbol{A}_{u,l}$.

The actual waiting time of the test call is the minimum of its maximum waiting time and its perceived waiting time, i.e.,

$$\Psi_{\rm act}(l) = \min\{\Psi_{\max}, \Psi_{\rm per}(l)\}. \tag{39}$$

Using [21, theorem 2.2.9], we can obtain the following result.

Corollary 1: The actual waiting time of the test call, $\Psi_{\rm act}(l)$, follows a PH distribution with dimension $\frac{rn^M(r^{l+1}-1)}{r-1}$, represented by PH($\boldsymbol{\omega}_{\rm act}(l)$,$\boldsymbol{\Omega}_{\rm act}(l)$), where $\boldsymbol{\omega}_{\rm act}(l)$ and $\boldsymbol{\Omega}_{\rm act}(l)$ are given by

$$\boldsymbol{\omega}_{\rm act}(l) = [\boldsymbol{\theta} \otimes \boldsymbol{\omega}_{\rm per}(l)], \text{ and } \boldsymbol{\Omega}_{\rm act}(l) = \boldsymbol{\Theta} \oplus \boldsymbol{\Omega}_{\rm per}(l). \tag{40}$$

The complementary distribution function of $\Psi_{\rm act}(l)$ can be obtained as

$$P_r(\Psi_{\rm act}(l) > t) = 1 - P(\Psi_{\rm act}(l) \le t) = \boldsymbol{\omega}_{\rm act}(l) \exp(\boldsymbol{\Omega}_{\rm act}(l)t)\boldsymbol{e}. \tag{41}$$

Since the number of queued ST calls seen by the test call lies between 0 and $M-1$, the complementary distribution function of the mean actual waiting time of the preempted ST calls, can be calculated as

$$P\{\Psi_{\rm act} > t\} = \sum_{l=0}^{M-1} P(\Psi_{\rm act}(l) > t) \cdot \boldsymbol{p}_i|_{i=M+l}\boldsymbol{e} = \sum_{l=0}^{M-1} \boldsymbol{\omega}_{\rm act}(l) \exp(\boldsymbol{\Omega}_{\rm act}(l)t)\boldsymbol{e} \cdot \boldsymbol{p}_i|_{i=M+l}\boldsymbol{e}. \tag{42}$$

From (42), the non-central moments, $\overline{\Psi_{\rm act}^k}$, can be obtained as

$$\overline{\Psi_{\rm act}^k} = (-1)^k k! \sum_{l=0}^{M-1} \boldsymbol{\omega}_{\rm act}(l)(\boldsymbol{\Omega}_{\rm act}(l))^{-k}\boldsymbol{e} \cdot \boldsymbol{p}_i|_{i=M+l}\boldsymbol{e}, \quad k \ge 1. \tag{43}$$

When $k=1$, we obtain the mean actual waiting time of the preempted ST calls, $\overline{\Psi_{\rm act}}$.

Remark 3: Given a constraint on the *channel utilization* η, the results derived above can be used to design an optimal ST arrival rate that minimizes the *ST blocking probability* B_S. Alternatively, one could fix the value of B_S below a predefined threshold and maximize *carried ST load* N_S due to a tradeoff between B_S and N_S (see Figs. 1 and 3 in Section VII).

VII. Numerical Results

In this section, we present both numerical and simulation results in terms of the obtained performance measures under the following parameter settings[5]: $M = 5$, $m = 2$, $n = 2$, $r = 2$. The MAP parameters are set as:

$$\mathbf{G}_P = \begin{bmatrix} 0.2\lambda_1 & 0.2\lambda_1 \\ 0.15\lambda_2 & 0.25\lambda_2 \end{bmatrix}, \ \mathbf{G}_S = \begin{bmatrix} 0.25\lambda_1 & 0.3\lambda_1 \\ 0.2\lambda_2 & 0.35\lambda_2 \end{bmatrix}, \ \mathbf{G}_0 = \begin{bmatrix} -\lambda_1 & 0.05\lambda_1 \\ 0.05\lambda_2 & -\lambda_2 \end{bmatrix},$$

where we set $\lambda_1 = \lambda_2 = \lambda$ for simplicity. The parameters of the various PH distributions are set as follows: $\boldsymbol{\alpha}_P = [0.5 \ \ 0.5]$, $\boldsymbol{\alpha}_S = [0.5 \ \ 0.5]$, $\boldsymbol{\theta} = [0.5 \ \ 0.5]$,

$$\mathbf{T}_P = \begin{bmatrix} -5 & 3 \\ 4 & -6 \end{bmatrix}, \ \mathbf{T}_S = \begin{bmatrix} -6 & 4 \\ 5 & -7 \end{bmatrix}, \ \boldsymbol{\Theta} = \begin{bmatrix} -3 & 2 \\ 2 & -4 \end{bmatrix}.$$

The proposed system model was simulated in MATLAB. Each simulated data point was averaged over 1,000 trials and the associated 95% confidence intervals were computed.

Fig. 1 shows the impact of the call arrival rates λ_P and λ_S (through the MAP parameter λ) on the PT and ST call blocking probabilities B_P and B_S. As expected, when the MAP parameter λ is increased, λ_P and λ_S increase linearly, leading to an increase in both B_P and B_S. For high call arrival rates, the performance of the secondary system deteriorates due to the lack of available channels. Note that the analytical results are validated by the simulation results. In Fig. 2, we compare the channel utilization of the OSS system with the single primary system[6]. We observe that the OSS system has a much higher channel utilization than the single primary system. We also observe that the channel utilization of the OSS system η increases as the call arrival rate is increased through the MAP parameter λ. This is intuitive, since the larger the call arrival rate, the higher the channel utilization.

In Fig. 3, we observe the relationship between different types of carried traffic and the traffic arrival rate. The carried PT and ST loads both increase as the arrival rate parameter λ increases. Note the total carried traffic corresponds to the sum of the carried PT and ST carried loads. In Fig. 4, we observe that the number of preempted ST calls in the preemption queue increases when the MAP parameter λ increases. This is because when λ increases, the PT call arrival rate λ_P increases linearly, leading to more ongoing ST calls being preempted. Fig. 4 validates the result derived in Section VI-C, i.e., the total number of ST calls is equal to the sum of the ST calls in service and the preempted ST calls.

Fig. 5 shows the mean preemption ratio of the ongoing ST calls γ as a function of the call arrival rate through λ. As λ increases, the mean preemption ratio γ increases. As more PT calls enter the system, fewer channels

[5]Here we choose small values of M, m, n, and r to keep the computational complexity small, while retaining the salient characteristics of MAP and PH distributions.

[6]The single primary system is obtained by suppressing the ST call arrivals, i.e., converting the contribution of q_{ij}^S, $1 \leq i, j \leq m$ to that of q_{ij}^0, $j \neq i$, $1 \leq i, j \leq m$, and forcing q_{ij}^S to 0.

are available for ST calls. Thus, a vacated ST call from its current channel has a smaller chance of obtaining an idle channel to continue its call, leading to a greater chance of generating a preempted ST call. Fig. 6 shows the complementary distributions of different waiting times (under $i_P = 3, l = 2$). The relationship of the actual waiting time of the test call $\Psi_{\rm act}(l)$ with respect to the maximum waiting time $\Psi_{\rm max}$ and the perceived waiting time $\Psi_{\rm per}(l)$ agrees with intuition.

VIII. Conclusions

We formulated a general performance model for an opportunistic spectrum sharing (OSS) wireless system. The OSS system consists of two types of users: primary and secondary. Primary calls have preemptive priority over secondary calls; a preempted secondary call attempts to resume service on an available channel, or waits in a preemption queue in the event that all channels are occupied. Arrivals of calls or sessions from both types of users are modeled by a Markovian arrival process (MAP), while channel holding times are modeled by phase-type distributions. Using matrix-analytic methods, computational algorithms are developed and performance metrics of interest are derived

The OSS system model encompasses a large class models as special cases and is useful for performance evaluation and design of future OSS systems. Although the model assumes that secondary users can perfectly detect the presence of primary users in a channel, the impact of unreliable spectrum sensing [27] could, in principle, be incorporated into the model. When the service distribution is nonexponential, the computational complexity of the model grows exponentially in the number of channels. To evaluate such scenarios, it would be worthwhile to investigate computationally efficient approximations to the general model proposed here.

Appendix

A. Proof of Theorem 1

When the *test call* arrives to find l, $0 \leq l \leq M-1$, queued ST calls, without loss of generality, the system state is assumed to consist of a total of i calls, among which are i_P PT calls and $M - i_P$ ST calls in service and l queued ST calls in the preemption queue. ¿From Section VI-D, the time variable φ_l can be derived from any of the following three events: (1) service completion of an ongoing PT call; (2) service completion of an ongoing ST call; and (3) dropping of a queued ST call. The event that occurs first triggers the transition to a new system state. For each event, any call that completes or drops first triggers a state transition. Since the maximum waiting time of a queued ST call and the residual channel holding times of the ongoing PT and ST calls all follow PH-distributions with representations PH$(\boldsymbol{\theta}, \boldsymbol{\Theta})$, PH$(\boldsymbol{\beta}_P, \mathbf{T}_P)$ and PH$(\boldsymbol{\beta}_S, \mathbf{T}_S)$, respectively, by applying [21, theorem 2.2.9], it can

be easily shown that φ_l follows a PH distribution PH$(\boldsymbol{a}_l, \boldsymbol{A}_{ll})$ with

$$\boldsymbol{a}_l = [\underbrace{\boldsymbol{\beta}_P \otimes \cdots \otimes \boldsymbol{\beta}_P}_{i_P} \otimes \underbrace{\boldsymbol{\beta}_S \otimes \cdots \otimes \boldsymbol{\beta}_S}_{M-i_P} \otimes \underbrace{\boldsymbol{\theta} \otimes \cdots \otimes \boldsymbol{\theta}}_{l}], \quad 0 \le l \le M-1, \tag{44}$$

$$A_{ll} = \underbrace{\mathbf{T}_P \oplus \cdots \oplus \mathbf{T}_P}_{i_P} \oplus \underbrace{\mathbf{T}_S \oplus \cdots \oplus \mathbf{T}_S}_{M-i_P} \oplus \underbrace{\boldsymbol{\Theta} \oplus \cdots \oplus \boldsymbol{\Theta}}_{l}, \quad 0 \le l \le M-1. \tag{45}$$

The dimension of the PH variable φ_l can be calculated as

$$n^{i_P} \cdot n^{M-i_P} \cdot r^l = n^M r^l. \tag{46}$$

Based on [24, theorem 2.6.1], the *perceived waiting time* of the test call, $\Psi_{\rm per}(l) \triangleq \varphi_0 + \varphi_1 + \cdots + \varphi_l$, follows the PH distribution with representation PH$(\boldsymbol{\omega}_{\rm per}(l), \boldsymbol{\Omega}_{\rm per}(l))$, where

$$\boldsymbol{\omega}_{\rm per}(l) = [\boldsymbol{\xi}_0, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_l], \tag{47}$$

and $\boldsymbol{\Omega}_{\rm per}(l)$ has the form given in (33). By repeatedly applying [24, theorem 2.6.1], we obtain

$$\begin{aligned}
\boldsymbol{\xi}_0 &= \boldsymbol{a}_0, \quad \boldsymbol{\xi}_1 = (1 - \boldsymbol{\xi}_0 \boldsymbol{e})\boldsymbol{a}_1 = (1 - \boldsymbol{a}_0 \boldsymbol{e})\boldsymbol{a}_1, \\
\boldsymbol{\xi}_2 &= (1 - \boldsymbol{\xi}_0 \boldsymbol{e} - \boldsymbol{\xi}_1 \boldsymbol{e})\boldsymbol{a}_2 = (1 - \boldsymbol{a}_0 \boldsymbol{e})(1 - \boldsymbol{a}_1 \boldsymbol{e})\boldsymbol{a}_2, \quad \cdots\cdots \\
\boldsymbol{\xi}_l &= (1 - \boldsymbol{\xi}_0 \boldsymbol{e} - \boldsymbol{\xi}_1 \boldsymbol{e} - \cdots - \boldsymbol{\xi}_{l-1} \boldsymbol{e})\boldsymbol{a}_l = \prod_{u=0}^{l-1} (1 - \boldsymbol{a}_u \boldsymbol{e})\boldsymbol{a}_l;
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{A}_{u,u+1} &= -\boldsymbol{A}_{uu} \boldsymbol{e} \boldsymbol{a}_{u+1}, \quad \boldsymbol{A}_{u,u+2} = -(\boldsymbol{A}_{uu} \boldsymbol{e} + \boldsymbol{A}_{u,u+1} \boldsymbol{e})\boldsymbol{a}_{u+2}, \quad \cdots\cdots \\
\boldsymbol{A}_{ul} &= -(\boldsymbol{A}_{uu} \boldsymbol{e} + \boldsymbol{A}_{u,u+1} \boldsymbol{e} + \cdots + \boldsymbol{A}_{u,l-1} \boldsymbol{e})\boldsymbol{a}_l = \sum_{k=u}^{l-1} (\boldsymbol{A}_{uk} \boldsymbol{e})\boldsymbol{a}_l.
\end{aligned}$$

From equations (47) and (46), we can calculate the dimension of the PH variable $\Psi_{\rm per}(l)$ as

$$n^M r^0 + n^M r^1 + n^M r^2 + \cdots + n^M r^l = \frac{n^M (r^{l+1} - 1)}{r - 1}. \tag{48}$$

REFERENCES

[1] M. McHenry, "Frequency agile spectrum access technologies," in *Proc. FCC Workshop on Cognitive Radio*, May 2003.

[2] G. Staple and K. Werbach, "The end of spectrum scarcity," *IEEE Spectrum*, vol. 41, pp. 48–52, March 2004.

[3] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 201–220, Feb. 2005.

[4] A. E. Leu, M. McHenry, and B. L. Mark, "Modeling and analysis of interference in listen-before-talk spectrum access schemes," *Int. J. Network Mgmt*, vol. 16, pp. 131–147, July 2006.

[5] S. Tang and B. L. Mark, "Performance analysis of a wireless network with opportunistic spectrum sharing," in *Proc. IEEE Globecom'07*, pp. 4636–4640, Nov. 2007.

[6] X. Liu and S. Shankar N., "Sensing-based opportunistic channel access," *Mobile Networks and Applications*, vol. 11, pp. 577–591, Aug. 2006.

[7] P. Pawelczak, R. V. Prasad, and R. Hekmat, "Waterfilling may not good neighbors make," in *Proc. IEEE ICC'07*, June 2007.

[8] L. Le and E. Hossain, "QoS-aware spectrum sharing in cognitive wireless networks," in *Proc. of IEEE Globecom'07*, pp. 3563–3567, Nov. 2007.

[9] S. Tang and B. L. Mark, "An adaptive spectrum detection mechanism for cognitive radio networks in dynamic traffic environments," in *Proc. IEEE Globecom'08*, (New Orleans, LA), Nov. 2008.

[10] A. Ghasemi and E. S. Sousa, "Opportunistic spectrum access in fading channels through collaborative sensing," *Journal of Communications*, vol. 2, pp. 71–82, March 2007.

[11] M. Rajaratnam and F. Takawira, "Nonclassical traffic modeling and performance analysis of cellular mobile networks with and without channel reservation," *IEEE Trans. Veh. Technol.*, vol. 49, no. 3, pp. 817–834, 2000.

[12] Y. Fang and I. Chlamtac, "A new mobility model and its application in the channel holding time characterization in PCS networks," in *Proc. IEEE INFOCOM'99*, pp. 20–27, Mar. 1999.

[13] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Netw.*, vol. 3, pp. 226–244, June 1995.

[14] L. Muscariello, M. Mellia, M. Meo, M. Marsan, and R. Cigno, "An MMPP-based hierarchical model of Internet traffic," in *Proc. of IEEE ICC'04*, pp. 2143–2147, June 2004.

[15] A. T. Andersen and B. F. Nielsen, "A Markovian approach for modeling packet traffic with long-range dependence," *IEEE J. Selected Areas in Comm.*, vol. 16, no. 5, pp. 719–732, 1998.

[16] A. Horvath and M. Telek, "Markovian modeling of real data traffic: Heuristic phase type and MAP fitting of heavy tailed and fractal like samples," in *Performance Evaluation of Complex Systems: Techniques and Tools, IFIP Performance'02, LNCS Tutorial Series*, vol. 2459, pp. 405–434, Springer-Verlag, 2002.

[17] S. Tang and B. L. Mark, "Modeling an opportunistic spectrum sharing system with correlated arrival process," in *Proc. IEEE WCNC'08*, pp. 3297–3302, 2008.

[18] H. Heffes and D. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Selected Areas in Comm.*, vol. 4, no. 6, pp. 856–868, 1986.

[19] W. Fischer and K. S. Meier-Hellstern, "The Markov-Modulated Poisson Process (MMPP) cookbook," *Performance Evaluation*, vol. 18, no. 2, pp. 867–885, 1993.

[20] L. Breuer and D. Baum, *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Springer, Dec. 2005.

[21] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University Press: Baltimore, MD, 1981.

[22] A. S. Alfa and W. Li, "A homogeneous PCS network with Markov call arrival process and phase type cell residence time," *Wireless Networks*, vol. 8, no. 6, pp. 597–605, 2002.

[23] A. Jayasuriya, D. Green, and J. Asenstorfer, "Modelling service time distribution in cellular networks using phase-type service distributions," in *Proc. of IEEE ICC'01*, pp. 440–444, June 2001.

[24] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA/SIAM Series on Statistics and Applied Probability, 1999.

[25] L. B. Le, E. Hossain, and A. S. Alfa, "Delay statistics and throughput performance for multi-rate wireless networks under multiuser diversity," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 3234–3243, Nov. 2006.

[26] L. B. Le, E. Hossain, and M. Zorzi, "Queueing analysis for GBN and SR ARQ protocols under dynamic radio link adaptation with non-zero feedback delay," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 3418–3428, Sep. 2007.

[27] S. Tang and B. L. Mark, "Modeling and analysis of opportunistic spectrum sharing with unreliable spectrum sensing," *IEEE Trans. Wireless Commun.*, 2009 (to appear).

[28] D. P. Gaver, P. A. Jacobs, and G. Latouche, "Finite birth-and death models in randomly changing environments," *Adv. Applied Prob.*, vol. 16, pp. 715–731, 1984.

[29] S. Tang and W. Li, "An adaptive bandwidth allocation scheme with preemptive priority for integrated voice/data mobile networks," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 2874–2886, Oct. 2006.

[30] P. Mishra, "Order-recursive Gaussian elimination," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 32, pp. 396–400, Jan. 1996.

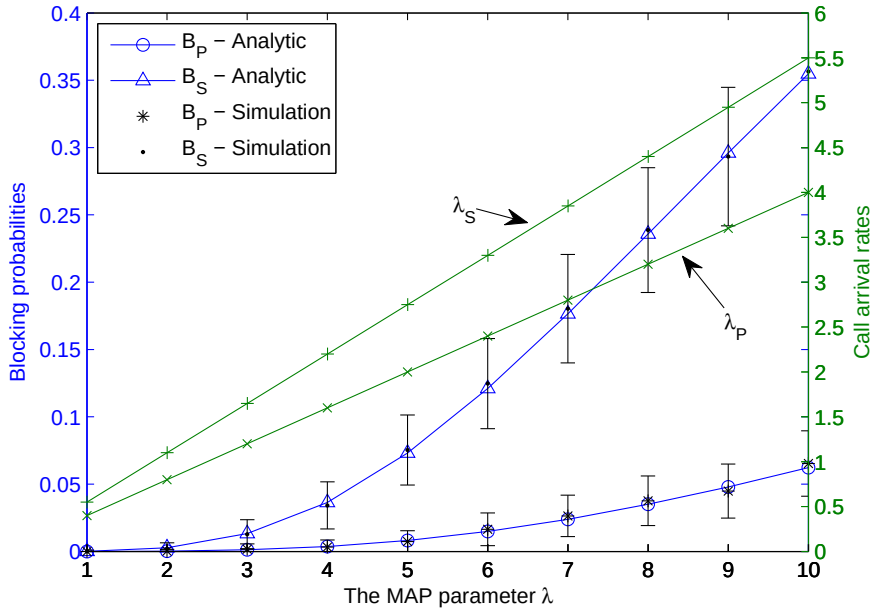

Fig. 1. PT and ST call blocking probabilities vs. call arrival rates.

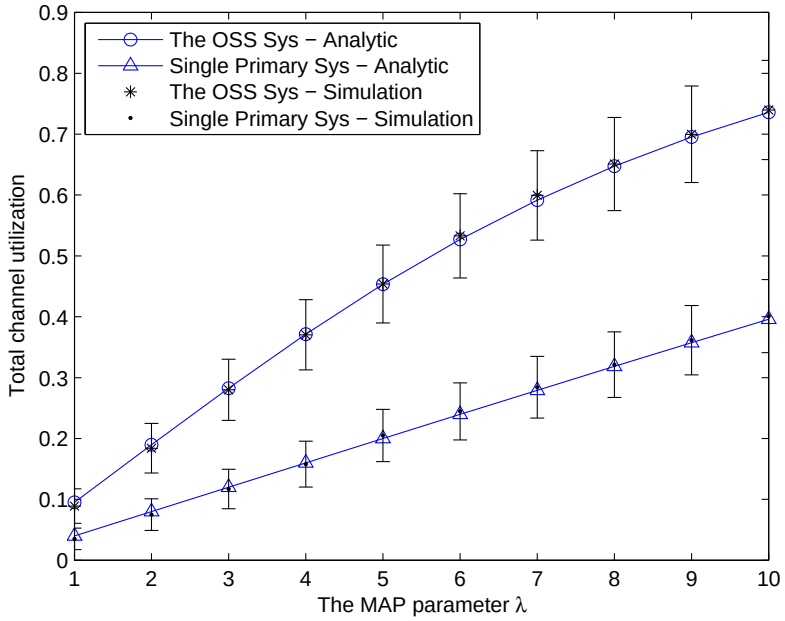

Fig. 2. Total channel utilization vs. call arrival rate (via parameter λ).

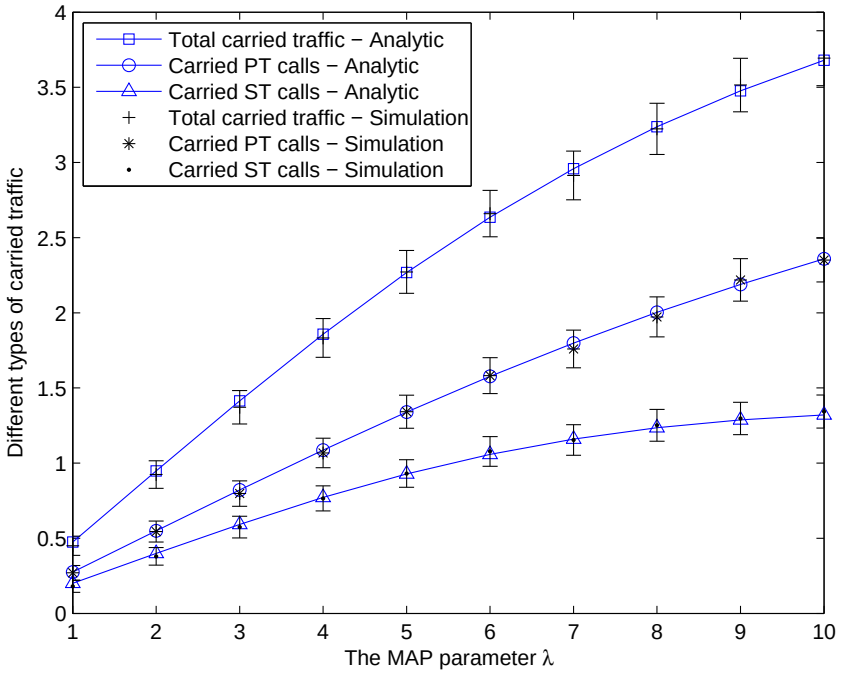

Fig. 3. Carried traffic vs. call arrival rate (via parameter λ).

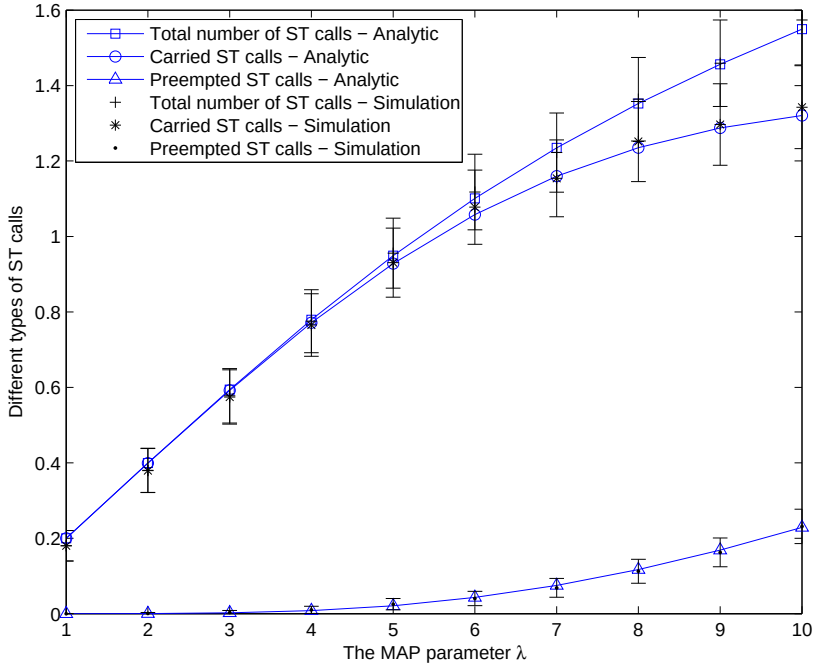

Fig. 4. Number of different types of ST calls vs. call arrival rate (via parameter λ).

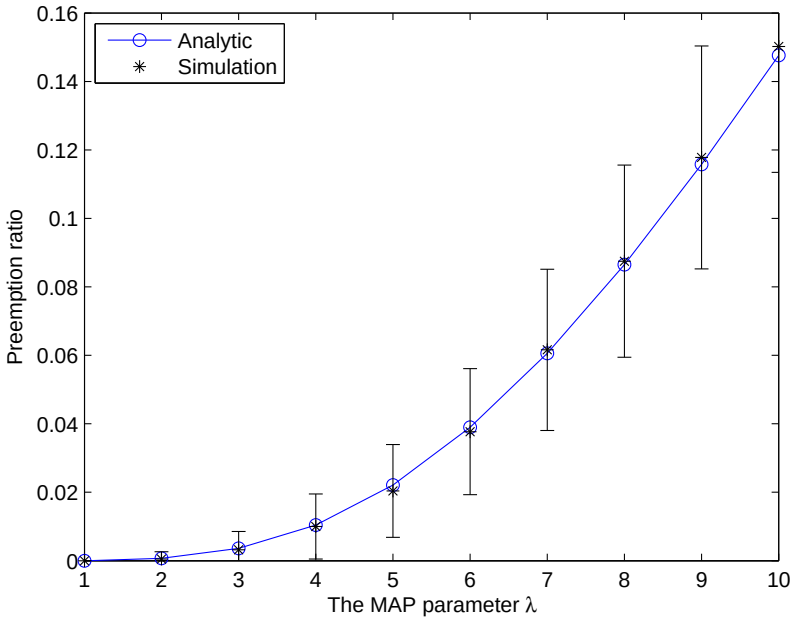

Fig. 5. Preemption ratio vs. call arrival rate (via parameter λ).

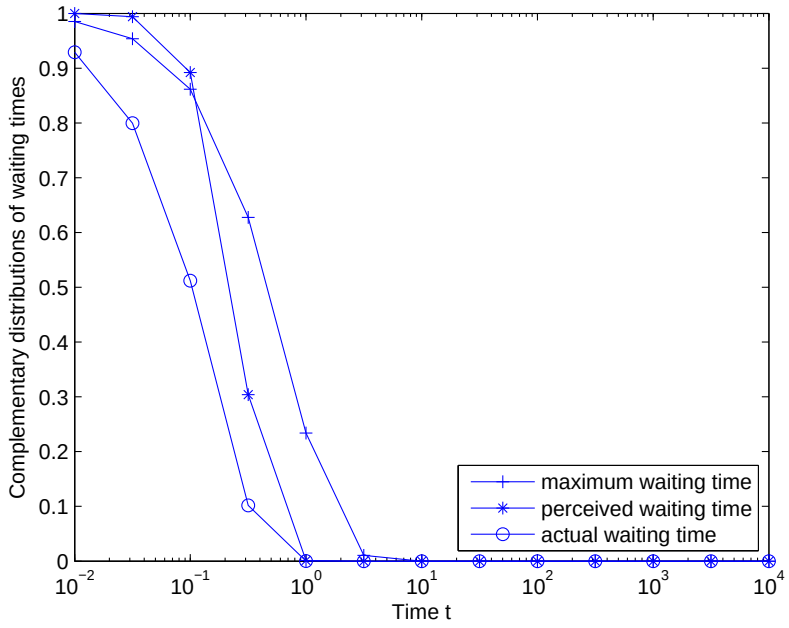

Fig. 6. Complementary distributions of different waiting times (under $i_P = 3, l = 2$).