Mobility Modeling and Performance Evaluation of Heterogeneous Wireless Networks

Ahmed H. Zahran, *Student Member, IEEE*, Ben Liang, *Senior Member, IEEE*, and Aladdin Saleh, *Senior Member, IEEE*

***Abstract*—The future-generation wireless systems will combine heterogeneous wireless access technologies to provide mobile users with seamless access to a diverse set of applications and services. The heterogeneity in this inter-technology roaming paradigm magnifies the mobility impact on system performance and user perceived service quality, necessitating novel mobility modeling and analysis approaches for performance evaluation. In this paper, we present and compare three mobility models in two-tier integrated heterogeneous wireless systems, the independence model as a naive extension of the traditional cell residence time modeling techniques for homogeneous cellular networks, the basic Coxian model which takes into consideration the correlation between the residence time within different access technologies, and the extended-Coxian model for further improved estimation accuracy. We propose a general stochastic performance analysis framework based on application session models derived from these mobility models, applying it to a 3G-WLAN integrated system as an example. Our numerical and simulation results demonstrate the general superiority of Coxian-based mobility modeling over the independence model. Furthermore, using the proposed modeling and analysis methods, we investigate the impact of different parameters, including WLAN coverage, handoff blocking probability, call holding duration, and mobility pattern, on system performance metrics such as network utilization time, handoff rates, and forced termination probability, for a wide range of user applications.**

***Index Terms*—Heterogeneous wireless networks, mobility modeling, performance analysis, beyond 3G, phase-type distribution, Coxian structure.**

NOTATION

As a general rule, we use bold-face capital letters and lowercase letters to represent matrices and vectors, respectively. Superscripts are used to represent absorption states and/or mobility models.

$t_c \sim PH(\alpha_c, \mathbf{T}_c)$ - cell residence time

$t_{wr} \sim PH(\alpha_{wr}, \mathbf{T}_{wr})$ - WLAN technology residence time

$t_{cr} \sim PH(\alpha_{cr}, \mathbf{T}_{cr})$ - cellular technology residence time

$PH(\alpha_m, \mathbf{T}_m)$ - heterogeneous phase-type cell residence time mobility models

$t_{ch} \sim Exp(\zeta_{ch})$, $t_{wh} \sim Exp(\zeta_{wh})$ - session holding times for cellular network and WLAN respectively

$PH(\alpha_m, \mathbf{T}_S)$ - heterogeneous phase-type session models

$\mathbf{q}^X$ - column vector representing absorbing rates to state $X \in \{Term, SHH, HHFT, VHFT\}$

P_{AB}^X - absorption probability to state X for a session of type $B \in \{n, h\}$ starting in an A-type cell, where $A \in \{C, W\}$

π_{To} - session initial phase distribution

P_n, P_h - probability of new and handoff sessions

P_{wo} - probability of WLAN-cellular technology overlap

P_{co} - probability of unique cellular coverage

B_h and B_v - horizontal and vertical handoff blocking probabilities

$\mathbf{I}$ - identity matrix

$\mathbf{e}$ - all-one column vector

Ahmed H. Zahran and Ben Liang are with the Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ontario, M5S 3G4, Canada. Aladdin Saleh is with Wireless Technology, Bell Canada. E-mail: {zahran,liang}@comm.utoronto.ca, aladdin.saleh@bell.ca. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and Bell Canada through its Bell University Laboratories R&D program.

I. INTRODUCTION

Next-generation wireless networks (NGWN) will feature a high level of heterogeneity due to the service convergence of different pervasive access technologies, such as wireless cellular networks, wireless Local Area Networks (WLAN), and wireless mesh networks. Further contributing to this heterogeneity will be the characteristic diversity of the newly introduced revenue generating applications and services. The 3G-WLAN integrated system is an example of this heterogeneous wireless access paradigm, which has received strong support from industrial [1] and standardization bodies [2], [3] as they currently represent the most pervasive wireless access technologies. In this system, 3G networks provide universal coverage, while WLANs provide large bandwidth resources for the users at cheaper cost wherever available. Thus, the network users will generally enjoy the best of each access technology, while the service providers will save precious wireless resources by offloading part of the 3G traffic to WLANs and take advantage of new potential revenue sources created by novel applications and services. However, the integration of different technologies increases system complexity and complicates the design and performance evaluation. Consequently, developing accurate traffic and mobility models emerges as a crucial requirement for the design of various processes such as location updating and paging, radio resource management, and technical network planning [4].

Generally, mobility models can be categorized into analytical and simulation models. *Analytical models* usually limit user mobility to a specific region of residence within the network while *simulation models* are based on periodic tracking of the user in small time steps. Random way-point [5], [6], random trip [7], and Gauss-Markov [8] models are examples of common generic mobility models in wireless cellular system simulations. Clearly, the simulation models presented in the literature can be also used for NGWN

simulation after introducing heterogeneous network overlays into the simulation setup. However, such models are generally intractable when the network coverage topology is under consideration. Therefore, the goal of this work is to develop new analytical models for heterogeneous systems.

In homogeneous cellular networks, the mobility of a mobile terminal (MT) is uniquely modeled by its cell residence time (CRT), defined as the duration spent by the MT within a cell. This level of granularity is sufficient to describe the MT mobility in homogeneous networks since the exact MT position within the cell is irrelevant to its application traffic pattern or service bandwidth demand. In the literature, the exponential random variable is extensively used to model the CRT [9]–[11] due to its analytical tractability. On the other hand, several works [12], [13] analytically prove, using simple mobility assumptions such as uniformly distributed random variables or bounded random variable variations, that the CRT has other more general distributions. Zonozzi and Dassanayake [12] show that the generalized Gamma distribution is a good fit for CRT, while the channel holding time can be approximated by an exponential distribution. The latter result is also proved by Hong and Rappaport [13]. Similarly, a few papers assume that the CRT follows a general distribution such as the generic phase-type (PH) distribution [14], Erlang distribution [15], [16], Gamma distribution [16], [17], hyper-exponential and hyper-Erlang distributions [15], and Sum of Hyper-Exponential (SOHYP) [18]. Based on such assumptions, these papers analytically derive different performance metrics such as the number of registration area crossings, channel holding time distribution, and forced termination probability [14]–[18].

In NGWN, inter-technology roaming, commonly known as vertical handoff (VHO) [19], affects different system performance metrics such as resource utilization, signaling load, and user perceived QoS, especially when the heterogeneous application characteristics are considered. For example, in the 3G-WLAN integrated model, the bandwidth provided to the MT in these networks may vary by one order of magnitude after any VHO. Combining this fact with the bandwidth greediness of some applications due to their buffering or prefetching capabilities [20], one can infer the large influence of VHO on next-generation session dynamics. Hence, VHO details should be accurately modeled for precise design and performance evaluation of NGWN.

Clearly, the current mobility models employing CRT as a unique MT mobility representation cannot describe VHO details in heterogeneous integrated systems. Hence, the main focus of this work is to develop accurate mobility models and performance analysis methods for next-generation heterogeneous two-tier systems. In a previous work [21], we explore heterogeneous mobility modeling for WLANs located strictly within cell borders. In this work, we relax this constraint, and propose a novel stochastic analysis framework to derive a wide range of performance metrics. To the best of our knowledge, this work is the first extensive study that addresses mobility modeling and performance analysis in an integrated heterogeneous network.

The contribution of this work is three-fold. *First*, we introduce the notion of *technology residence time* (TRT) to model the duration spent by an MT inside a specific access technology. For the 3G-WLAN integrated model, it includes the WLAN residence time and the inter-WLAN residence time. We then describe three MT mobility models using different PH distributions for different TRTs. They include a naive independence model (IM), the Coxian model (CM), and the extended-Coxian model (ECM). IM extends the existing PH representation of the CRT in homogeneous systems, by simply combining the homogeneous CRT model dynamics with the dynamics of different TRTs assuming that they are independent. On contrary, CM and ECM adopt a novel approach where the CRT is modeled as a probabilistic sum of WLAN and inter-WLAN residence times to *physically* represent the MT handoff transitions within a cell. *Second*, we develop new application session models, based on all three mobility models, to derive several salient performance metrics such as network utilization times, handoff rates, and session forced termination probabilities for different applications. Numerical and simulation results show that by accommodating the dependence between the CRT and TRTs, CM and ECM offer significantly improved estimation accuracy over IM, with ECM outperforming CM in modeling highly random MT mobility. *Finally*, we propose a stochastic analysis framework to study the impact of different system parameters, such as WLAN coverage, handoff blocking probability, and user mobility on the modeling accuracy and the system performance based on an extensive set of performance metrics and a wide range of user applications.

The rest of this paper is organized as follows. Section II presents the proposed next-generation mobility models. The corresponding session models are developed in Section III. The proposed network performance evaluation framework is then introduced in Section IV. The simulation and analysis results are presented in Section V; and finally, we conclude in Section VI.

II. Modeling Mobility in Two-Tier Networks

In this section, we present three MT mobility models for any two-tier integrated heterogeneous system, using a 3G-WLAN integrated system as an example for illustration purpose. As shown in Figure 1, in the 3G-WLAN integrated system, WLANs overlap with the 3G cellular coverage forming hotspots with dual coverage and non-hotspot areas with unique 3G cellular coverage. As the MT traverses the overlay cellular cell, it may encounter zero or more hotspots during its CRT, denoted as t_c. Cell to cell transition may occur in a unique coverage area (e.g. T1 in Figure 1) or through a WLAN that overlaps with two different 3G cells (e.g. T2 in Figure 1). Hence, the MT cell residence may initially start in any of the two technologies. Moreover, during any cell visit, the MT may traverse different technologies as in T1 and T2, spending a random amount of time in each visit, or may just traverse one technology, as in T3. All these mobility details are accommodated using TRTs, denoted as t_{wr} and t_{cr} for WLAN and inter-WLAN residence times respectively and can be generally modeled as PH distributions.

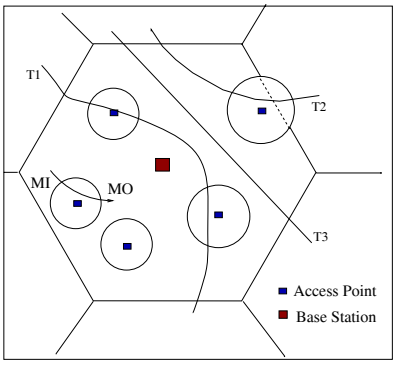


Fig. 1. 3G-WLAN integrated system

A. PH Distribution Overview

Phase-type distributions are highly versatile stochastic models that can be used to approximate the distribution of any non-negative random variables [22] and at the same time enjoy analytical tractability due to their underlying Markovian structure. Generally, a PH random variable is defined as the absorption time of an evanescent finite-state Markov chain to a single absorbent state. This chain is represented by its infinitesimal generator matrix, $\mathbf{Q}$, and an initial state distribution vector υ as follows [23]

$$\mathbf{Q} = \begin{pmatrix} \mathbf{T}_{m\times m} & \mathbf{t}_{m\times 1} \\ \mathbf{0}_{1\times m} & 0 \end{pmatrix}, \tag{1}$$

$$\upsilon = (\alpha_{1\times m}, \gamma_{1\times 1}), \tag{2}$$

where $\mathbf{T}$ corresponds to the chain transient dynamics and $\mathbf{t}$ represents the absorption rate vector. It is worth mentioning that the absorption vector $\mathbf{t}$ can be expressed as $\mathbf{t} = -\mathbf{T}\mathbf{e}$, where $\mathbf{e}$ is an all-one column vector. Hence, the corresponding PH distribution can be defined by $(\alpha, \mathbf{T})$, such that if a random variable X is $PH(\alpha, \mathbf{T})$ of order m, then its probability density function is expressed as

$$f(x) = \alpha \exp(\mathbf{T}x)\mathbf{t}, \;\; x \geq 0\,. \tag{3}$$

Generally, there are two different modeling approaches using PH distributions [24]: fictitious and physical approaches. In the former, PH distributions are used as a versatile, dense, and algorithmically tractable class of distributions defined on the non-negative real numbers; while in the latter, phases or blocks of phases represent real processes or operations that take place in the system. In this work, contrary to the traditional homogeneous models, we use the second approach to develop heterogeneous PH CRT models using different structures of PH distributed TRTs. In the following subsections, we explain the structure of the proposed models and how their parameters are estimated from mobility traces that record actual movement behavior for certain segments of the population [25].

B. Independent PH Mobility Model (IM)

For the purpose of comparison, we first describe a naive modeling approach. IM assumes that the CRT and TRTs are independent. It fits t_c, t_{wr}, and t_{cr} from the collected traces to PH distributions denoted as $PH(\alpha_c, \mathbf{T}_c)$, $PH(\alpha_{wr}, \mathbf{T}_{wr})$, and $PH(\alpha_{cr}, \mathbf{T}_{cr})$ with orders m_c, m_{wr}, and m_{cr} respectively. It models the possible technology alternation within the cell borders by replacing each phase in the CRT PH structure by a stage[1] that represents an alternating renewal process of t_{wr} and t_{cr}. This continuous alternation is terminated by stage exit to another stage or transition into the cell exit state, at the same corresponding transition rates of the original PH distribution for t_c.[2] Clearly, the resultant CRT model has a PH structure with order $m_I = m_c(m_{wr} + m_{cr})$, which is denoted as $PH(\alpha_m^I, \mathbf{T}_m^I)$.

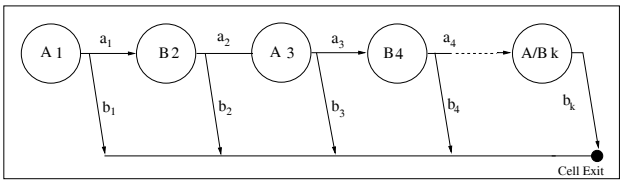


Fig. 2. Coxian and extended-Coxian mobility model

C. Coxian Mobility Model (CM)

The proposed Coxian mobility model structurally accommodates the correlation between CRT and TRTs by expressing the CRT as a probabilistic summation of the TRTs. The Coxian model inherits its structure from the Coxian PH random variable structure [27] shown in Figure 2. In this model, each phase is labeled with a letter and a number. The former represents the access technology, and the latter represents the phase sequence. The technology labels A and B may respectively represent cellular (i.e., inter-WLAN) and WLAN technologies or vice versa depending on the model's initial technology. Hence, inter-phase transitions *physically* represent VHOs that take place within the MT cell residence. Whenever the MT exits phase i, where $i = 1, 2, ..., k-1$, it may exit the cell, i.e. is absorbed into the "cell exit" state, with probability b_i or may be vertically handed off to another technology within the same cell with probability a_i, where $a_i + b_i = 1$. Following the standard definition of the Coxian PH distribution, the duration spent by the MT in any phase i is exponentially distributed with mean $1/\mu_i$. When the MT is in the last phase k, the MT exits the cell with probability $b_k = 1$. Hence, the CRT can be expressed as $PH(\alpha_m^C, T_m^C)$, where

$$\mathbf{T}_m^C = \begin{pmatrix} -\mu_1 & a_1\mu_1 & 0... & ..0 \\ 0 & \mu_2 & a_2\mu_2 & ..0 \\ . & & . & . \\ 0.. & & 0 & \mu_k \end{pmatrix} \tag{4}$$

$$\alpha_m^C = [1\ \mathbf{0}]\,. \tag{5}$$

D. Extended-Coxian Mobility Model (ECM)

The extended-Coxian model generalizes the Coxian model by replacing the exponential TRT phases with PH distributed TRT stages. Hence, each stage will be $PH(\alpha_i, \mathbf{T}_i)$, where i

[1] As a notational remark: a *stage* refers to a group of phases, keeping the *phase* as a notion for exponentially distributed sojourn time states.

[2] Due to space limitation, the interested reader is referred to [26] for further details of IM.

represents the stage index. Clearly, this general model includes the Coxian model when all the stages have exponential residence times. This generalization will be shown later to better accommodate highly random mobility patterns.

The resultant MT mobility model also has a PH structure with cell exit as an absorption state. Hence, the extended-Coxian CRT can be expressed as $PH(\alpha_m^{EC}, \mathbf{T}_m^{EC})$, where

$$\mathbf{T}_m^{EC} = \begin{pmatrix} \mathbf{T}_1 & a_1\mathbf{t}_1\alpha_2 & \mathbf{0}... & ..\mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 & a_2\mathbf{t}_2\alpha_2 & ..\mathbf{0} \\ . & & . & . \\ \mathbf{0}.. & & \mathbf{0} & \mathbf{T}_k \end{pmatrix} \quad (6)$$

$$\alpha_m^{EC} = [\alpha_1 \ \mathbf{0}]\,. \quad (7)$$

where $\mathbf{t}_i = -\mathbf{T}_i\mathbf{e}$.

E. Estimating the Parameters of Mobility Models

The parameters of the proposed models can be estimated from mobility traces that are collected per visited overlay cell, either from practical systems or from simulation. We emphasize that the exact parameter estimation details are non-essential to the proposed mobility models, session models, and stochastic performance analysis framework. We provide the following parameter estimation methods as an example of possible approaches.

The required mobility traces contain the initial technology visited by the MT, WLAN residence times, and inter-WLAN residence times. The CRT measurements are then obtained from the WLAN and inter-WLAN residence time information as the total of these durations per cell. As a first step of the parameter estimation process, the collected traces are clustered into two partitions based on the initial technology. Each cluster is then processed to calculate its corresponding sub-model parameters. These sub-models completely describe the MT mobility in a two-tier integrated system. Hence, we use C-type and W-type to differentiate sub-models with initial cellular and initial WLAN phases respectively.

In this work, the PH distribution parameters are estimated using a PH fitting package such as EMPHT [28]. In this process, the collected data is fitted to different PH distributions according to the measurement coefficient of variation, θ_x, where $\theta_x = \frac{\sigma_x}{\mu_x}$ in which σ_x and μ_x represent the standard deviation and the mean of the corresponding measurements respectively. According to [27], the hyper-exponential distribution can be used to represent any set of measurements with $\theta_x > 1$, the hypo-exponential (generalized Erlang) can be used to represent any set of measurements with $\theta_x < 1$, and the exponential distribution is used to represent measurements with $\theta_x = 1$.

In Coxian models, different parameters are originally estimated according to their corresponding physical events. For example, the exit probabilities are calculated as

$$b_i = \frac{N_c(i)}{\sum_{j=i}^{\infty} N_c(j)}\ ,$$

where $N_c(i)$ denotes the number of cells in which exactly i technology visits take place. The model order k is determined by the number of technology visits within the cells in the collected traces[3]. Additionally, the time statistics of different stages are calculated from the corresponding physical network visit; i.e. the time statistics of the first stage are fitted from the collected measurements of the first visit to technology A within the cell. Similarly, the time statistics of the second stage are calculated from the collected measurements of the first visit to technology B within the cell. We have also examined a uniform time statistics approach in which all stages corresponding to the same technology have the same time statistics. These statistics are fitted from the measurements of different stages collectively; hence, fewer traces are required compared to the exact fitting approach. The uniform approach produces the same level of accuracy as the exact approach while reduces fitting time.

III. Modeling Application Sessions in Heterogeneous Networks

In this section, we present new session models based on the aforementioned mobility models. These models are developed using the application model presented in subsection III-A and a generic session modeling approach presented in subsection III-B. Each session model tracks the session activity, which is affected by user mobility, user-network interaction, and application characteristics. Note that active users request resource allocation from their point of attachment, which may deny this request based on resource availability. Without loss of generality, we assume that users prefer WLAN over 3G due to its larger bandwidth and lower cost. Hence, any active MT will be always handed off to a WLAN whenever it is available.

A. Application Modeling

Generally, applications can be categorized according to different criteria such as bandwidth and delay requirements. In this work, we further categorize the applications as *symmetric* and *asymmetric*. The former preserves the same level of resource utilization independent of the available network resources, while the latter has a greedy nature and can consume as much bandwidth as the network can provide. Conversational applications, such as voice over IP (VoIP) and video conference (V-conf), are examples of the former, while streaming applications with buffering capabilities, such as video on demand (VoD) and radio on demand (RoD), are examples of the latter.

We assume that each service is characterized by exponentially distributed holding times t_{ch} and t_{wh} with rates ζ_{ch} and ζ_{wh} for cellular network and WLAN respectively. Symmetric applications are expected to preserve their holding time and bandwidth requirement in both networks. Hence, both parameters have the same value; i.e. $\zeta_{ch} = \zeta_{wh}$. On the other hand, asymmetric applications have shorter WLAN duration compared to their cellular holding time; i.e. $\zeta_{ch} < \zeta_{wh}$.

The exponential session holding time assumption is common in wireless networks due to their pricing strategies and

[3]Alternatively, other fitting mechanism can be adopted such as fitting the model order to realize a specific confidence of the measurement mean. Our fitting and analytical results show that this approach greatly reduces the model order for highly random mobility, but it decreases the accuracy of the obtained performance metrics. Details are omitted for brevity.

limited power resources. It is well known that charge per time is the most common pricing strategy; hence, users tend to avoid long session durations. Additionally, the high power consumption of active devices is another reason for shortening session durations. Hence, exponential is a good model for user controlled session duration such as conversational applications [29], [30]. In contrast, for streaming applications, several studies [31], [32] show that files transmitted on the Internet feature a large variance in comparison to their sizes. In this case, the hyper-exponential distribution can be used to represent the streaming session duration. Consequently, the session duration, L, is expressed as a probabilistic sum of different exponential distributions, i.e. $f_L(l) = \sum_{i=1}^{k} p_i \zeta_i e^{-\zeta_i l}$, for which a *hyper-metric* z is estimated as a weighted sum of multiple metrics z_i calculated for different values of ζ_i, i.e. $z = \sum_{i=1}^{k} p_i z_i$.

We note that the session holding time model has its limitations. For example, it does not capture packet-level metrics such as packet delay. However, packet-level performance in heterogeneous wireless networks is not the focus of this work and remains an open problem for future research.

B. Combining Application and Mobility Dynamics

Traditionally, different application performance metrics are obtained by considering the minimum of the session holding time and the cell residual time for new calls, or the minimum of the session residual time and the CRT for handoff sessions [9], [10], [15]–[17]. In heterogeneous networks, this approach cannot be generally applied to the proposed models due to the inherent characteristic diversity of different systems. Additionally, this approach limits the performance metrics to the cell level instead of the technology level. In the proposed session models, system heterogeneity is accommodated by shifting the analysis down to the TRT instead of the traditional CRT.

In the heterogeneous session model, the application dynamics are combined with the TRT mobility dynamics by taking the minimum of each PH TRT and the PH session holding time. This minimum operation results in a new PH distribution, i.e. $min(PH(\alpha, \mathbf{T}_{wr}), PH(1, -\zeta_{wh})) = PH(\alpha, \mathbf{T}_{wr} - \zeta_{wh}\mathbf{I})$. The minimum operation is repeated with all cellular and WLAN stages of each mobility model, respectively denoted as C and W stages. The resultant session model will have a PH structure, $PH(\alpha_m, \mathbf{T}_S)$, similar to its corresponding mobility model with modulated stages. It is worth noting that, the absorption rate from any phase i in this session model equals the sum of the corresponding phase holding rate; i.e. either ζ_{ch} or ζ_{wh} depending on the phase technology, and the mobility model absorbing rate, $\mathbf{t}_i$.

In the proposed session model, we additionally define different absorbing states to represent the session status according to application dynamics, user mobility, and user-network interaction. In the case of normal session termination, which may result from session ending at the user's will or due to content transfer completion for the streaming case, the session is absorbed into a $Term$ state. Additionally, the user-network interaction is accounted for at the end of each TRT to represent probabilistic successful and blocked handoffs. Hence, we define SHH and $HHFT$ absorbing states to represent successful and blocked horizontal handoff (HHO) respectively. Similarly, we define $VHFT$ as an absorbing state for sessions blocked during VHO. Note that each successful VHO advances the session Markov chain to a transient cellular technology stage in any of the proposed models.

Hence, the Markovian session process generator matrix, $\mathbf{Q}_S$ can be expressed as

$$\mathbf{Q}_S = \begin{pmatrix} \mathbf{T}_S & \mathbf{q}^{Term} & \mathbf{q}^{SHH} & \mathbf{q}^{HHFT} & \mathbf{q}^{VHFT} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{8}$$

where the vectors $\mathbf{q}^{Term}$, $\mathbf{q}^{SHH}$, $\mathbf{q}^{HHFT}$, and $\mathbf{q}^{VHFT}$ respectively correspond to the aforementioned absorbing states. The following two subsections present detailed derivation of $\mathbf{Q}_S^C$ and $\mathbf{Q}_S^{EC}$ for CM and ECM, respectively. The derivation of $\mathbf{Q}_S^I$ for IM is similar and is omitted due to page limitation.

C. Coxian Session Model

In all session models, the transient dynamics are mainly determined by the minimum of each TRT and the session holding time within the corresponding technology and the probable session blocking during roaming from WLAN to the 3G cellular network. Session blocking reduces the transition rates from WLAN phases to cellular phases to $(1 - B_v)$ of their corresponding values in the mobility models, where B_v represents the VHO blocking probability. The remaining B_v portion of these rates represent the transition rates to $\mathbf{q}^{VHFT}$. Hence, $\mathbf{T}_S^C$ can be expressed as

$$\mathbf{T}_S^C = [q_{rs}] = \begin{cases} -(\mu_r + \zeta_{ch}) & ,\forall\, r = s\,,\, r \in C \\ -(\mu_r + \zeta_{wh}) & ,\forall\, r = s\,,\, r \in W \\ a_r \mu_r & ,\forall\, s = r+1\,,\, r \in C \\ a_r \mu_r (1 - B_v) & ,\forall\, s = r+1\,,\, r \in W \\ 0 & ,\text{otherwise} \end{cases}, \tag{9}$$

and $\mathbf{q}^{C,VHFT}$ is expressed as

$$\mathbf{q}^{C,VHFT} = [q_r] = \begin{cases} a_r \mu_r B_v & ,\forall\, r \in W \\ 0 & ,\text{otherwise} \end{cases},$$

Similarly, due to session blocking possibility at the cell border, the cell exit absorbing state is subdivided into two states $\mathbf{q}^{C,SHH}$ and $\mathbf{q}^{C,HHFT}$. It is worth noting that the HHO that occurs within a WLAN overlapping with the overlay cell border are never blocked because the user does not request network resources instantaneously, while those HHOs that take place while the MT is using the cellular network may be blocked with a HHO blocking probability of B_h or may be successfully handed off to another cell with probability $(1 - B_h)$. Consequently, $\mathbf{q}^{C,SHH}$ and $\mathbf{q}^{C,HHFT}$ can be expressed using the cell exit rates as

$$\mathbf{q}^{C,SHH} = [q_r] = \begin{cases} b_r \mu_r (1 - B_h) & ,\forall\, r \in C \\ b_r \mu_r & ,\forall\, r \in W \\ 0 & ,\text{otherwise} \end{cases},$$

$$\mathbf{q}^{C,HHFT} = [q_r] = \begin{cases} b_r \mu_r B_h & ,\forall\, r \in C \\ 0 & ,\text{otherwise} \end{cases},$$

Additionally, the session normal termination from each phase naturally occurs at the session holding rate that corresponds to the phase technology, i.e., at ζ_{ch} and ζ_{wh} for cellular and WLAN phases respectively. Hence, $\mathbf{q}^{C,Term}$ is expressed as

$$\mathbf{q}^{C,Term} = [q_r] = \begin{cases} \zeta_{ch} & ,\forall\, r \in C \\ \zeta_{wh} & ,\forall\, r \in W \\ 0 & ,\text{otherwise} \end{cases} .$$

D. Extended-Coxian Session Model

Similar to the mobility model, the extended-Coxian session model is a generalized matrix version of the Coxian session model in which stages have matrix representations. Hence, by applying the same rules shown in subsection III-C, the extended-Coxian Markovian session generator matrix can be expressed as

$$\mathbf{T}_S^{EC} = [\mathbf{T}_{ij}] = \begin{cases} \mathbf{T}_i - \zeta_{ch}\mathbf{I} & ,\forall\, i = j\,,\, i \in C \\ \mathbf{T}_i - \zeta_{wh}\mathbf{I} & ,\forall\, i = j\,,\, i \in W \\ a_i \mathbf{t}_i \alpha_j & ,\forall\, j = i+1\,,\, i \in C \\ (1-B_v) a_i \mathbf{t}_i \alpha_j & ,\forall\, j = i+1\,,\, i \in W \\ 0 & ,\text{otherwise} \end{cases} , \tag{10}$$

$$\mathbf{q}^{EC,VHFT} = [\mathbf{q}_i] = \begin{cases} a_i \mathbf{t}_i B_v & ,\forall\, i \in W \\ 0 & ,\text{otherwise} \end{cases} ,$$

$$\mathbf{q}^{EC,SHH} = [\mathbf{q}_i] = \begin{cases} b_i \mathbf{t}_i (1-B_h) & ,\forall\, i \in C \\ b_i \mathbf{t}_i & ,\forall\, i \in W \\ 0 & ,\text{otherwise} \end{cases} ,$$

$$\mathbf{q}^{EC,HHFT} = [\mathbf{q}_i] = \begin{cases} b_i \mathbf{t}_i B_h & ,\forall\, i \in C \\ 0 & ,\text{otherwise} \end{cases} ,$$

$$\mathbf{q}^{EC,Term} = [\mathbf{q}_i] = \begin{cases} \zeta_{ch}\mathbf{e} & ,\forall\, i \in C \\ \zeta_{wh}\mathbf{e} & ,\forall\, i \in W \\ 0 & ,\text{otherwise} \end{cases} .$$

IV. Network Performance Analysis

Generally, the performance modeling and analysis of cellular systems can be conducted at two levels [16]. The first-level modeling uses the amount of wireless resources (e.g., number of radio channels) available in the cell as an input parameter to determine the new call-blocking probability and forced termination probability. The second-level modeling uses the new call-blocking and the forced termination probabilities as input parameters to study the call-completion probability, expected effective call hold times, and expected number of handoffs. Each level of analysis uses some of the output metrics of the other level as input parameters. The integration of both approaches is proposed in [13] and [30] using an iterative rounds of the first and second levels of analysis to accommodate the user-network interaction, e.g. traffic arrival and call admission.

Noting that the channel[4], traffic classes, and admission control concepts can be tailored to the NGWN, the same first-level analytical approaches can still be used in NGWN analysis. On contrary, the traditional second level analytical approaches [14]–[18], which evenly treat the phases in traditional mobility models, are no longer applicable due to the heterogeneity of different phases in the presented mobility models. Hence, we focus in this section on developing a second-level analytical framework for NGWN to calculate several salient performance metrics such as network utilization times, handoff rates, and session termination probabilities. This framework is based on PH distribution properties and Markovian analytical techniques. Additionally, the analysis considers different scenarios evolving from different combinations of mobility sub-models and session types, i.e. new and handoff sessions.

In the analysis, the session type alters the initial phase distribution, π_{To}, defined as the probability distribution of starting the session in a specific phase. For handoff sessions, since the MT activity completely spans the CRT, the initial state distribution is equal to the mobility model initial state distribution, i.e. $\pi_{To} = \alpha_m$. For new calls, the MT activity spans the residual CRT; hence, the initial state distribution equals the initial state distribution of the residual CRT. Since the residual time of a PH distribution $PH(\alpha, \mathbf{T})$ is another phase-type distribution $PH(\beta, \mathbf{T})$, where $\beta = (\alpha \mathbf{T}^{-1}\mathbf{e})^{-1}\alpha\mathbf{T}^{-1}$ [23, Theorem 3.3.1], we have, for new sessions, $\pi_{To} = (\alpha_m \mathbf{T}_m^{-1}\mathbf{e})^{-1}\alpha_m \mathbf{T}_m^{-1}$. Furthermore, the mobility sub-models alter the $\mathbf{T}_m$ matrix depending on the technology where an application session is initiated. Consequently, session generator matrices $\mathbf{Q}_S$ are altered according to the initial technology of the mobility sub-model. It is worth mentioning that the proposed analytical framework represents a novel generic approach that can be applied to any phase-type system representation to obtain the derived performance metrics, including the IM, CM, and ECM mobility and session models.

A. Horizontal Handoff Rate

The HHO rate is defined as the expected number of generated horizontal handoffs from a new session. In an integrated 3G-WLAN network, the HHO rate differs from the homogeneous case due to session dynamic variations resulting from inherent network heterogeneity. This metric is estimated using session absorption probabilities of the Markovian session process. Generally, the absorption probabilities can be estimated using the embedded discrete Markov chain whose probability transition matrix, $\mathbf{W}_S = [w_{ij}]$, can be derived from the session model infinitesimal generator matrix $\mathbf{Q}_S$ [33]:

$$w_{rs} = \begin{cases} \frac{-q_{rs}}{q_{rr}} & ,\forall r \neq s, q_{rr} \neq 0 \\ 0 & ,\forall r = s, q_{rr} \neq 0 \\ 0 & ,\forall r \neq s, q_{rr} = 0 \\ 1 & ,\forall r = s, q_{rr} = 0 \end{cases} . \tag{11}$$

Similar to $\mathbf{Q}_S$, $\mathbf{W}_S$ can be partitioned into transient states and absorbing states. Hence, we have

$$\mathbf{W}_S = \begin{pmatrix} \mathbf{W}_{TT} & \mathbf{W}^{Term} & \mathbf{W}^{SHH} & \mathbf{W}^{HHFT} & \mathbf{W}^{VHFT} \\ \mathbf{0} & \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4 \end{pmatrix} ,$$

where $\mathbf{e}_r$ is an all-zero column vector except at the r^{th} location, which is equal to one.

[4]Channels may be time slots, frequency bands, spreading codes, or combinations of them depending on the multiple-access scheme for the system under consideration.

Let P_{AB}^{X} denote the absorption probability to state X given that a session of type B starts in A-type sub-model, then P_{AB}^{X} can be calculated as [33]

$$P_{AB}^{X} = \pi_{T0,AB}(I - \mathbf{W}_{TT,A})^{-1}\mathbf{W}_{A}^{X}\,, \tag{12}$$

where X can be Term, SHH, HHFT, or VHFT. Using these absorption probabilities, we define the following probabilities.

- P_{hh}: the probability that a handoff session will be further horizontally handed off to a neighbor cell. Hence, $P_{hh} = P_{wo}P_{wh}^{SHH} + P_{co}P_{ch}^{SHH}$, where P_{wo} is the percentage of WLAN coverage to the cellular coverage and P_{co} represents the percentage of unique cellular coverage and equals $1 - P_{wo}$. Note that WLAN is assumed to be randomly located in the cellular coverage.
- P_{hft}: the probability that a handoff session will be terminated in the same cell either due to normal termination or due to forced termination during VHO. Hence, $P_{hft} = P_{wo}(1-P_{wh}^{SHH}-P_{wh}^{HHFT})+P_{co}(1-P_{ch}^{SHH}-P_{ch}^{HHFT})$.
- P_{hs}: the probability that a handoff session will perform exactly one successive HHO. This event takes place either due to a successful handoff to a neighbor cell in which the session terminates or due to forced termination at the cell borders. Hence, $P_{hs} = P_{wo}(P_{wh}^{HHFT} + P_{wh}^{SHH}P_{hft}) + P_{co}(P_{ch}^{HHFT} + P_{ch}^{SHH}P_{hft})$.

Consequently, one can derive the marginal distribution function of the HHO rate, H, assuming the session starts in a WLAN as follows:

$$\begin{aligned} P(H=0|W) &= P_{wn}^{Term} + P_{wn}^{VHFT}\,,\\ P(H=1|W) &= P_{wn}^{HHFT} + P_{wn}^{SHH}P_{hft}\,,\\ P(H=k|W) &= P_{wn}^{SHH}P_{hh}^{k-2}P_{hs}\,, \forall k \geq 2\,. \end{aligned}$$

Then, the expected number of HHOs for a session starting in a WLAN is calculated as

$$E\{H|W\} \;=\; \sum_{k=0}^{\infty} kP(H=k|W)\,.$$

Using the mathematical identity $\sum_{i=0}^{\infty} ic^i = \frac{c}{(1-c)^2},\ |c|<1$, the expected number of HHO can be expressed as

$$E\{H|W\} = P_{wn}^{HHFT} + P_{wn}^{SHH}\left(P_{hft} + P_{hs}\left(\frac{2-P_{hh}}{(1-P_{hh})^2}\right)\right).$$

Similarly, the handoff rate for a session starting in the cellular network is

$$E\{H|C\} = P_{cn}^{HHFT} + P_{cn}^{SHH}\left(P_{hft} + P_{hs}\left(\frac{2-P_{hh}}{(1-P_{hh})^2}\right)\right).$$

The total HHO rate, N_{HH}, is

$$N_{HH} = E\{H|W\}P_{wo} + E\{H|C\}P_{co}\,.$$

B. Session Termination Probabilities

The successful termination probability, $P(ST)$, is defined as the probability that an unblocked session will terminate normally by the user, i.e. will not be forced to terminate during handoff. In our session model, successful termination in a specific cell is represented by the absorption to the *Term* state. Hence, assuming that the session starts in a WLAN, $P(ST|W)$ can be expressed as

$$P(ST|W) = P_{wn}^{Term} + P_{wn}^{SHH}\left(\sum_{k=0}^{\infty} P_{hh}^{k}\right)P_{ht}\,,$$

where $P_{ht} = P_{wo}P_{wh}^{Term} + P_{co}P_{ch}^{Term}$ which represents the probability that a handoff session will terminate within the current cell. Similarly, $P(ST|C)$ can be expressed as

$$P(ST|C) = P_{cn}^{Term} + P_{cn}^{SHH}\left(\sum_{k=0}^{\infty} P_{hh}^{k}\right)P_{ht}\,.$$

Then, the successful termination probability is

$$P(ST) = P(ST|W)P_{wo} + P(ST|C)P_{co}\,.$$

Consequently, the session forced termination probability, $P(FT)$, can be expressed as

$$P(FT) = 1 - P(ST)\,.$$

C. Network Utilization Times

The utilization time of a specific network is defined as the expected time spent by the MT in a certain network before it is handed off to a neighbor cell. For a specific network type, this metric can be calculated as the duration spent by the MT in phases corresponding to the same technology. Hence, in the example integrated 3G-WLAN network, the cellular utilization and WLAN utilization times are calculated as the duration spent in the cellular and WLAN phases respectively before absorption. One way to obtain these metrics is by using [23, Theorem 2.4.3]. This theorem states that $(-\mathbf{T}^{-1})_{rs}$ is the expected total time spent in phase s until absorption given that the initial phase is r. Hence, the expected time spent in different phases until absorption, $\mathbf{L}_T$, can be expressed as

$$\mathbf{L}_T = -\pi_{To}\mathbf{T}_S^{-1}\,. \tag{13}$$

Consequently, the expected cellular network utilization time in the integrated model will be $E\{L_c\} = \sum_{r\in C}\mathbf{L}_T(r)$, and the expected WLANs utilization time will be $E\{L_w\} = \sum_{r\in W}\mathbf{L}_T(r)$. To this end, the estimated values represent conditional metrics estimated for a specific session type and a specific mobility sub-model. The total metric value is then estimated using the total probability theorem over different combinations of mobility sub-models and session types.

The mobility sub-model probability is determined by the initial network probability that depends on the the percentage of WLAN overlapping with the cellular network. Therefore, the probability that the initial network is WLAN equals to P_{wo}, and the probability that the initial network is the cellular network equals P_{co}. On the other hand, the session-type probability depends on the application HHO rate. Denoting λ_n and λ_h as the new and handoff session arrival rates of a specific application respectively, a new session probability can be expressed as $P_n = \frac{\lambda_n}{\lambda_n+\lambda_h}$. Additionally, a handoff session probability is $P_h = 1 - P_n$. In [17], it has been shown that

the handoff arrival rate is $\lambda_h = \lambda_n N_{HH}$. Consequently, P_n and P_h can be expressed as

$$P_n = \frac{1}{1+N_{HH}}\,,\; P_h = \frac{N_{HH}}{1+N_{HH}}\,.$$

Hence, the expected cellular utilization time is

$$\begin{aligned} E\{L_c\} &= P_{wo}P_n E\{L_c|WN\} + P_{co}P_n E\{L_c|CN\} + \\ &\quad P_{wo}P_h E\{L_c|WH\} + P_{co}P_h E\{L_c|CH\}\,. \end{aligned}$$

Similarly, the expected WLAN utilization time can be estimated. Finally, the expected session cell dwelling time, $E\{L_s\} = E\{L_c\} + E\{L_w\}$.

D. Vertical Handoff Rates

There are two types of VHOs, upward and downward handoffs. The former is defined as the transition from a WLAN to the cellular network, and the latter is the reverse case. These two types are also known as move out (MO) and move in (MI) respectively, as shown in Figure 1. The latter taxonomy is due to the fact that WLAN is considered a preferred network to the cellular network due to its higher bandwidth and lower cost. We defined the VHO rate as the expected number of VHOs induced by an active session within a 3G cell. For all proposed mobility and session models, the VHO rate is calculated using Markovian reward models [34]. The expected number of MIs, $E\{N_{MI}\}$, can be expressed as

$$E\{N_{MI}\} = \pi_{To}\Psi^{MI}\,,$$

where Ψ^{MI} is a column vector whose r^{th} element, ψ_r^{MI}, represents the total expected number of MIs induced from an active session given that it starts in phase r. Using the fact that MIs are due to transition from a cellular phase to a WLAN phase, the expected number of MIs induced from a phase $s \in C$, given that the session starts in phase r, can be calculated by assigning a reward that equals the summation of the transition rates from phase s to any phase $l \in W$, i.e. $\rho_s = \sum_{l\in W} q_{sl}$, where q_{sl} is the transition rate from phase s to phase l. Hence, the accumulated reward until absorption equals the product of the assigned phase reward and the duration spent within this phase $(-\mathbf{T}_S^{-1})_{rs}$. Consequently, the total expected number of MIs given that the session starts in phase r can be expressed as $\psi_r^{MI} = \sum_{s\in C}(-\mathbf{T}_S^{-1})_{rs}\rho_s$.

Similarly to the network utilization times, the MI rate N_{MI} is conditionally calculated for different combinations of initial-network and session types. Then, the metric is calculated using the total probability theorem as

$$\begin{aligned} N_{MI} &= P_{wo}P_n E\{MI|WN\} + P_{co}P_n E\{MI|CN\} + \\ &\quad P_{wo}P_h E\{MI|WH\} + P_{co}P_h E\{MI|CH\}\,. \end{aligned}$$

Using a similar reward structure, the MO rate, N_{MO}, can be obtained. Finally, we have the VHO rate $N_{VHO} = N_{MI} + N_{MO}$.

TABLE I
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
P_{wo}	0.3	N	100
B_v	0.01	B_h	0.01
d_H(sec)	5	d_s(sec)	3

TABLE II
APPLICATION PARAMETERS

	VoIP	Vconf	RoD	VoD
$1/\zeta_{ch}$	3	30	60	90
$1/\zeta_{wh}$	3	30	10	15

V. NUMERICAL RESULTS

In addition to the above analysis, we have simulated an integrated heterogeneous system, with square cells for simplicity of illustration. Each cell is composed of N square subdivisions, where WLANs are randomly located with probability P_{wo}. When an MT is handed off to another cell, it experiences a new random WLAN topology. In order to emulate practical MT operation, a handoff area [35] of d_H seconds is assumed between overlay 3G cells. This delay corresponds to the hysteresis introduced in handoff algorithms to decrease the ping-pong impact during horizontal handoff. Additionally, the MT MI is delayed with d_s seconds as a typical delay required for WLAN discovery and handoff signaling [36].

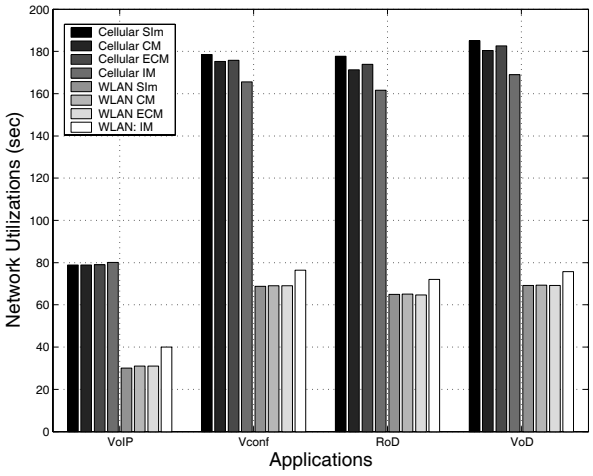
In this work, mobility traces are generated by using the aforementioned network setup. However, it is worth mentioning that the traces can alternatively be collected by direct field measurements with real system implementation using dual interfaced devices. However, to the best of our knowledge, mobility traces with the required level of details are not available in public. On the one hand, cellular traces such as the Stanford University Mobile Activity Traces (SUMATRA) [37] are limited to the zone (cell) level of granularity. On the other hand, traces focusing on WLAN access, such those presented in [38], [39], only concern MT behavior inside WLANs and do not contain any information about the cellular network.

For mobility simulation, we adopt a two-dimensional Gauss-Markov movement model from [8], as it can be easily tuned to represent a wide range of user mobility patterns between the two extreme cases of random-walk and constant velocity fluid-flow. In this model, a MT velocity is assumed to be correlated in time and is modeled by a Gauss-Markov process. In its discrete version, at time n, the MT velocity in each dimension, v_n, is given by

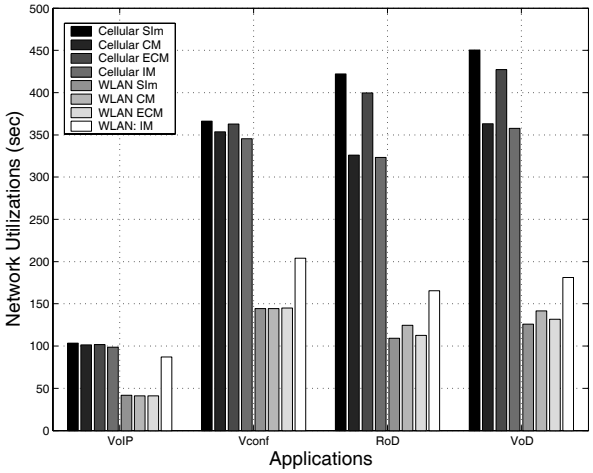
$$v_n = \alpha_v v_{n-1} + (1-\alpha_v)\mu_v + \sqrt{1-\alpha_v^2}\,x_{n-1}\,, \qquad (14)$$

where α_v, $0 \le \alpha_v \le 1$, represents a velocity memory factor, μ_v is the asymptotic mean of v_n, and x_n is an independent and stationary Gaussian process with zero mean and standard deviation σ_v, where σ_v is the asymptotic standard deviation of v_n.

Table I lists the default values of system and simulation parameters. The application parameters used in our simulations are shown in Table II. The chosen applications includes both symmetric-conversational applications, VoIP and Vconf, and asymmetric-streaming applications, RoD and VoD. Note that

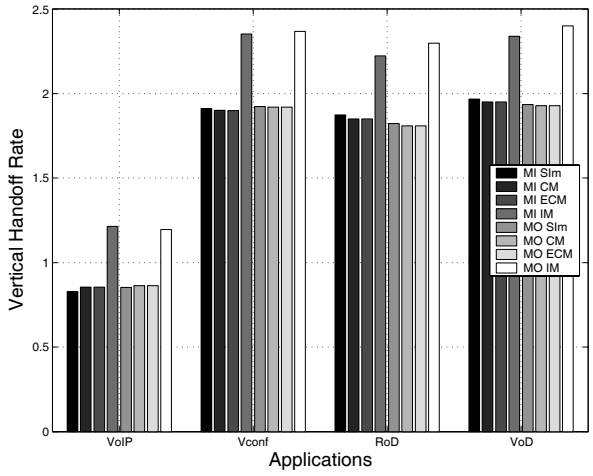


(a) $\alpha_v = 0.9$

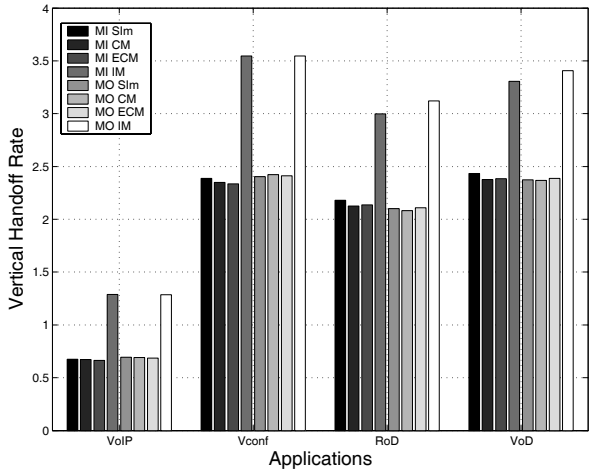


(b) $\alpha_v = 0$

Fig. 3. Network utilization



(a) $\alpha_v = 0.9$



(b) $\alpha_v = 0$

Fig. 4. VHO rate

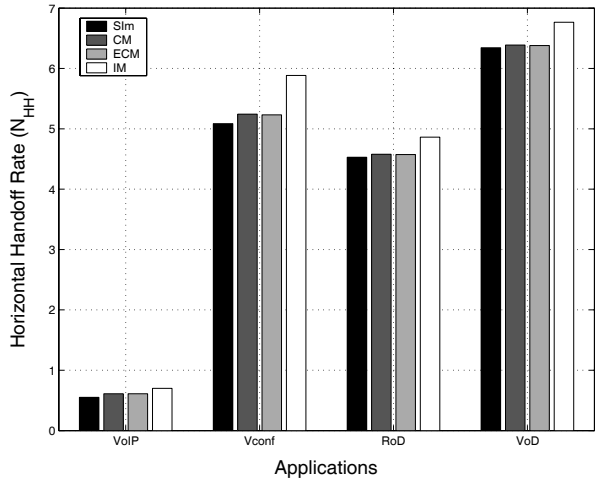
the numbers within Table II represent application durations that combine application data requirements, application symmetry, and the bandwidth offered in each network.

Since system details at the packet level are non-essential to the performance analysis under consideration, both simulation and analysis results have been obtained using Matlab. The following subsections compare the performance of the proposed models and show the impact of different system parameters such as WLAN coverage, MT mobility, and system blocking probability on the derived performance metrics.

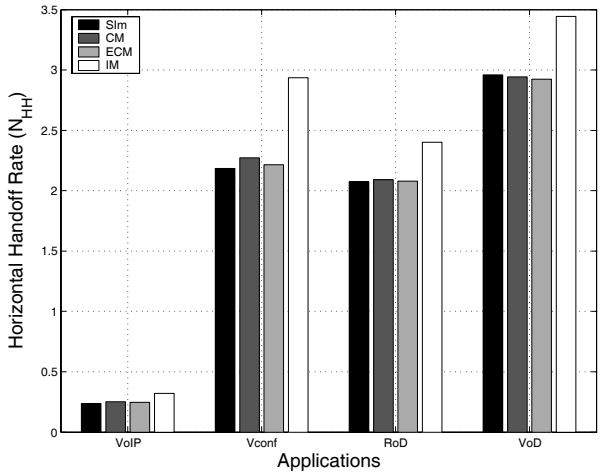
A. Mobility Modeling Accuracy

In this subsection, we compare the accuracy of the proposed models assuming MT mobility parameters to be $\alpha_v = 0.9$ and $\alpha_v = 0$ with $\mu_v = 0$ and $\sigma_v = 2.5$. Figures 3-6 illustrate the cellular and WLAN network utilization times, VHO rate, HHO rate, and forced termination probability respectively for different applications. All figures show that the Coxian-based models provide significantly better match between simulation and analysis results when compared to the independence model. For the Coxian models, and ECM in particular, the discrepancy between simulation and analysis is less than 8%. In comparison, the independence model generally results in much larger mismatch and may lead to 500% discrepancy as in estimating the forced termination probability for conversational applications such as VoIP. It is clear that this mismatch is due to ignoring the dependence between the CRT and the WLAN and inter-WLAN residence times, which leads to an inaccurate estimation of the absorption probabilities and consequently, the performance metrics. On contrary, accommodating the correlation between CRT and TRTs using the Coxian structure results in far better estimates for different metrics.

Additionally, we observe that the difference between CM and ECM is insignificant for large memory values shown for $\alpha_v = 0.9$; however, this difference increases as the motion becomes more random as shown for $\alpha_v = 0$. In general, our results show that the basic Coxian model can be used in most cases, for easy parameter estimation and numerical analysis due to its simpler structure. However, for systems with highly random mobility patterns, the extended-Coxian model provides more accurate results. This observation is further

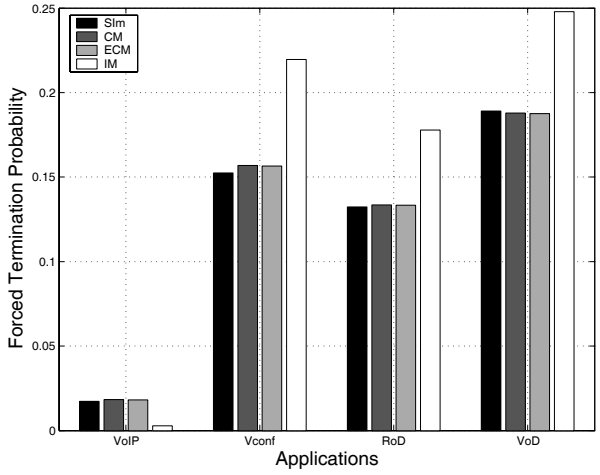


(a) $\alpha_v = 0.9$

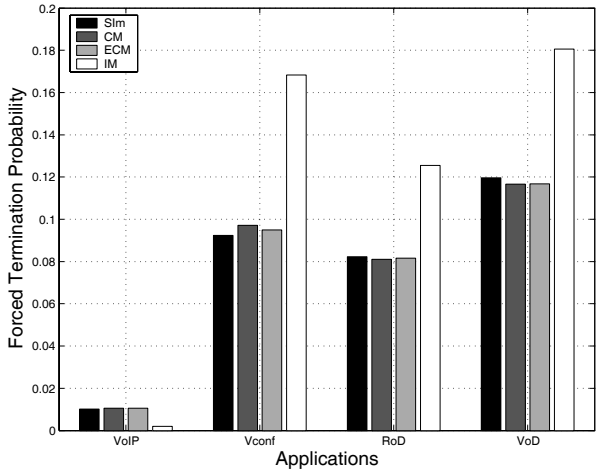


(b) $\alpha_v = 0$

Fig. 5. HHO rate



(a) $\alpha_v = 0.9$



(b) $\alpha_v = 0$

Fig. 6. Forced termination probability

explored in Section V-B4.

B. Network Performance and System Design Guidelines

The integrated system of heterogeneous wireless networks is rich with different system parameters that seriously affect the system performance. In this subsection we investigate the impact of several parameters such as WLAN coverage, user mobility behavior, and other system parameters. This investigation is based on accurate performance analysis using the proposed Coxian-based modeling techniques. In the following, we present analysis results based on CM for most cases, except those involving highly random mobility patterns, in which case results based on ECM are also presented. In all figures, unless stated otherwise, solid lines and dashed lines represent the analysis and simulation results, respectively. Furthermore, all simulation results include 95% confidence intervals.

1) WLAN Coverage: WLAN coverage is one of the most important system parameters in NGWN as it greatly affect system resource utilization as the users migrate between both 3G network and WLANs. Additionally, its impact is even more significant for asymmetric applications due to its noticeable impact on their session dynamics. In this subsection, the MT mobility parameters are set to $\alpha_v = 0.9$, $\mu_v = 0$, and $\sigma_v = 2.5$. Figure 7 shows the network utilization times of different applications versus WLAN coverage. The figures suggest an excellent match between simulation and analysis results with less than 5% discrepancy. Intuitively, the figure shows that the utilization time of a technology is proportional to its coverage. The estimated values enable the estimation of cellular and WLAN traffic load and consequently can be used to determine the required resources for different cells with different WLAN coverage.

Figure 8 plots the VHO rates of different applications versus WLAN coverage. This figure shows an interesting phenomenon in which VHO rate changes its increasing trend after WLAN coverage exceeding 50%. The accurate estimation of VHO rate values is important for determining the required processing capacity of different mobility servers that handles handoff requests, e.g the home agent in Mobile IP. Clearly, WLAN coverage has a great impact on the server capacity as its variation leads to large changes in the number of VHOs performed by the MT. Figure 8 shows that VoD, RoD, and

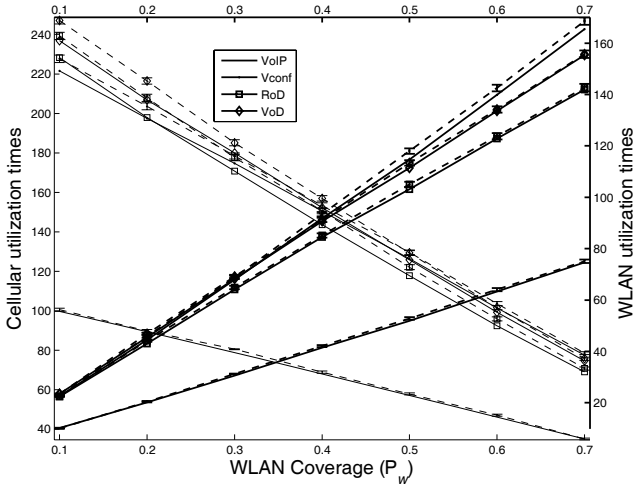


Fig. 7. Network utilization time versus WLAN coverage.

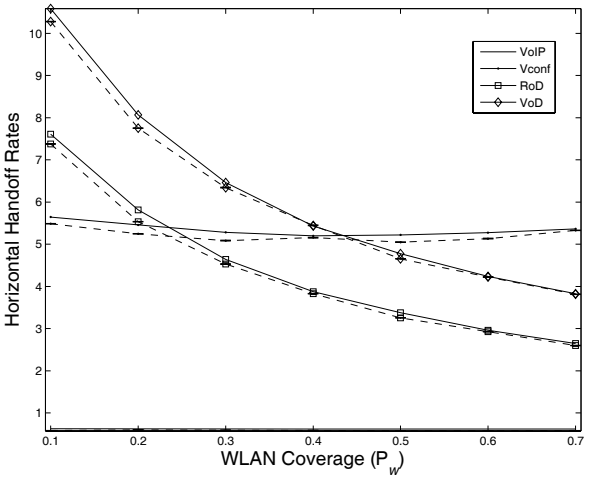


Fig. 9. HHO rate versus WLAN coverage.

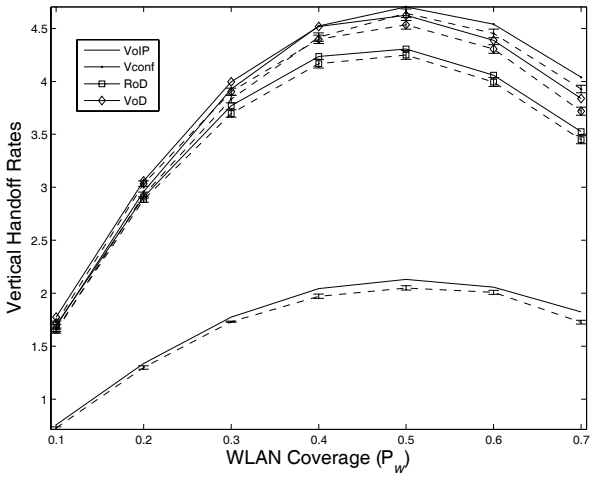


Fig. 8. VHO rate versus WLAN coverage.

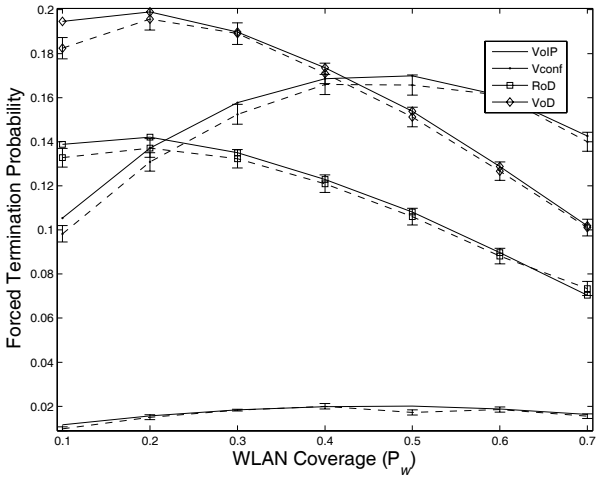


Fig. 10. Forced termination probability versus WLAN coverage.

Vconf perform 1.7 VHOs per session per cell on average for WLAN coverage of 10% and up to approximately 4.5 VHOs per session per cell on average for WLAN coverage of 50%. This VHO rate variation approximately corresponds to 150% increase in the signaling load; hence, WLAN coverage variation should be carefully considered during the design phase.

Figure 9 shows HHO rate variations for different applications versus WLAN coverage. Intuitively, the HHO rates of asymmetric applications decrease as the WLAN coverage increases as they take the advantage of larger bandwidth of WLANs, while symmetric application HHO rate is independent of WLAN coverage. Hence, integrating 3G and WLAN services is generally expected to decrease the HHO signaling load of 3G systems.

Figure 10 shows the session forced termination probability of different applications versus WLAN coverage. The figure suggests a maximum for the forced termination probability of symmetric applications at 50% coverage, after which the Forced termination probability changes its variation trend. The forced termination increasing trend is due to the probable forced termination during repetitive VHO as WLAN coverage increases. However, as the WLAN coverage increases beyond 50%, the probability that the coverage of different WLANs overlap increases and fewer VHOs are performed. Hence, the forced termination probability decreases. On contrary, as asymmetric applications benefit from the higher bandwidth in WLAN, in addition to coverage increase, the turning point is shifted to lower WLAN coverage percentages.

2) Vertical Blocking Probability: The vertical blocking probability, B_v, is a new design parameter beyond traditional cellular systems. Generally, the value of the blocking probabilities are determined by the session management system designer. Figures 11-14 plot the derived metrics versus VHO blocking probability for the same mobility and application parameters as in the previous section and WLAN coverage of 30%. In these figures, we vary B_v between zero and ten times B_h, where zero represents an ideal system with uninterrupted VHO, while with $B_v = B_h$, the system designer choose to treat VHO the same way as HHO from neighbor cell, i.e. the session management system design will be kept unchanged. Similar to the previous results, the CM provides

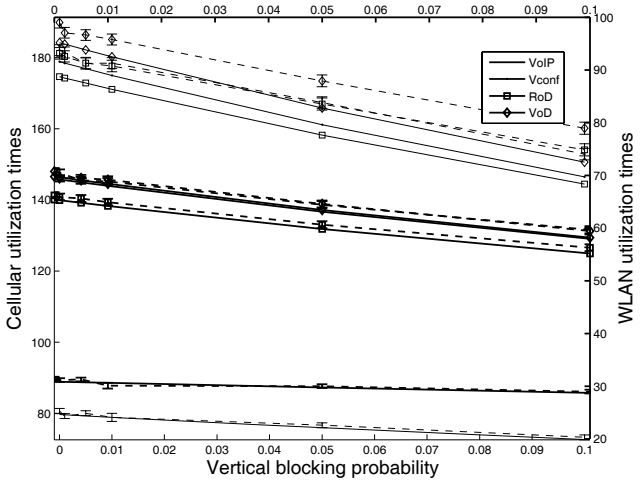


Fig. 11. Network utilization times versus VHO blocking probability.

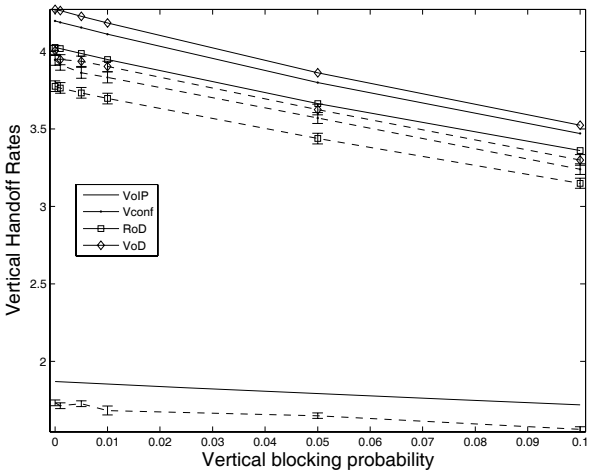


Fig. 12. VHO rate versus VHO blocking probability.

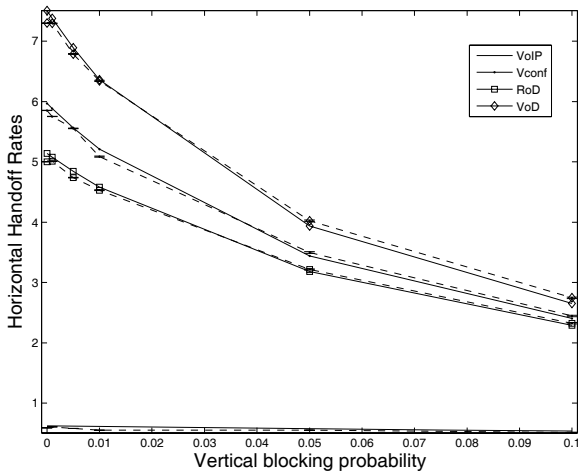


Fig. 13. HHO rate versus VHO blocking probability.

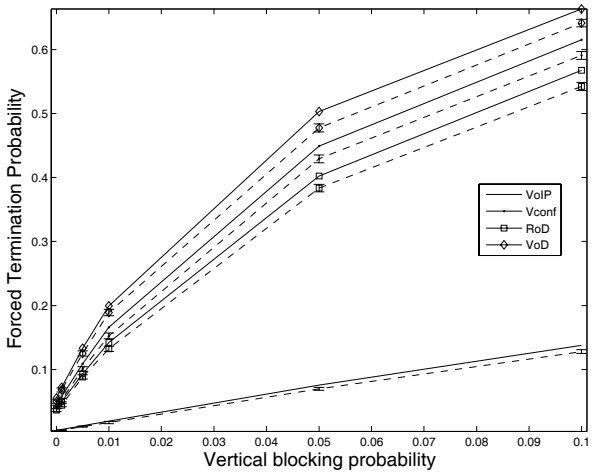


Fig. 14. Forced termination probability versus VHO blocking probability.

an accurate estimate for different metrics. Additionally, the figures clearly show that the impact of B_v variation on session metrics is much more than that on cell metrics. For example, Figure 14 shows that treating VHOs as HHOs result in a noticeable increase in session forced termination probability that approximately equals four times of that in a system with zero VHO blocking. Additionally, the figure shows that the forced termination probability can be effectively decreased by prioritizing VHOs and maintaining B_v at 0.001. This result demonstrates the critical impact of session management system design on NGWN performance. Hence, the system designer should carefully allocate the amount of guard bandwidth to satisfy the application target QoS level.

3) Session Holding Time and Network Gain: In this subsection, we investigate the impact of varying session holding time and the network asymmetry on the derived performance metrics. The latter factor can be used to study the impact of WLAN congestion level or the impact of limiting WLAN-provided service rate according to user service profile assuming that WLAN has infinite resources. Figures 15-18 plot the derived metrics versus cellular holding time and different network gains of 1, 3, 6, and 10, where the network gain is defined as ζ_{wh}/ζ_{ch}, representing the ratio of *utilized* bandwidth (by symmetric or asymmetric applications) between WLAN and the cellular network. These figures show that network utilization and VHO saturate as the application session holding time increase. This saturation is due to the fact that the MT CRT is limited. Hence, one can estimate the maximum expected network utilization and signaling load per cell. On contrary, session based metrics, such as HHO rates and forced termination probability, intuitively increase as session holding time increases.

4) User Mobility Patterns: In this section, we investigate the impact of user mobility patterns on both the performance metrics and the accuracy of CM and ECM. In the simulation and mobility trace collection for the analysis, the mobility change is realized by varying the user memory factor α_v of the Gauss-Markov mobility model, from zero to one corresponding to the complete spectrum between random-walk and fluid-flow mobility patterns respectively.

The mobility pattern variation leads to noticeable changes in both the first and second order statistics of different residence times. Generally, as the motion randomness increases, i.e. the

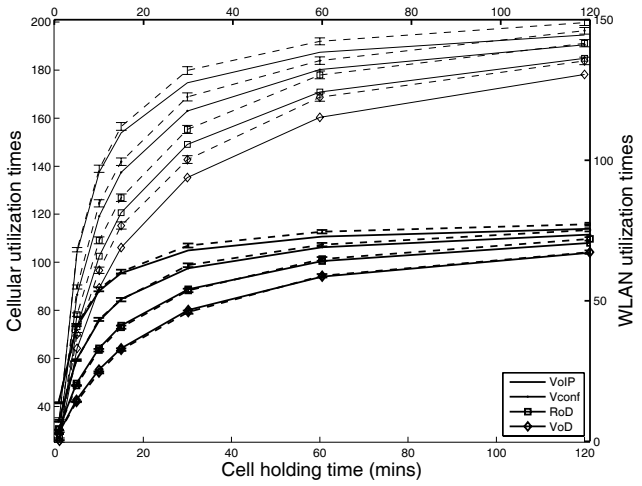


Fig. 15. Network utilization versus call holding time.

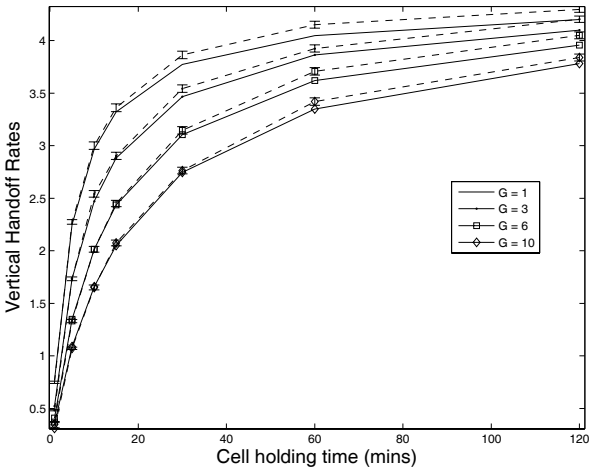


Fig. 16. VHO rate versus call holding time.

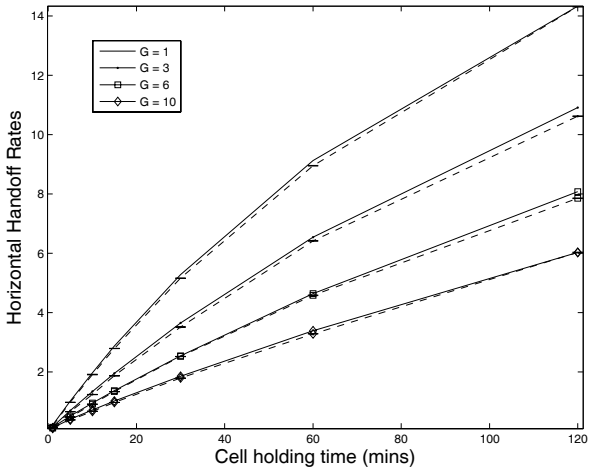


Fig. 17. HHO rate versus call holding time.

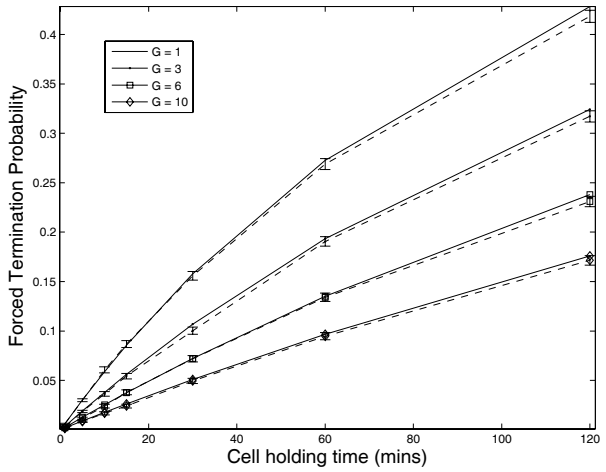


Fig. 18. Forced termination probability versus call holding time.

TABLE III
COXIAN MODEL ORDER

α_v	0	0.25	0.5	0.75	0.9	1
Exact fitting	693	541	455	258	123	15

memory factor decreases, both the mean and the variance of different residence times increase. This increase is due to the inverse relation between motion randomness and the MT locality. Clearly, fluid-flow travelers preserve their direction and speed and consequently, stay for a short duration around the same location, while random walkers continuously change both their direction and their velocity and consequently, stay for a longer duration around the same location. Furthermore, the variance of different time variables increases even more significantly compared to the mean. This fact is depicted in Figure 19.

We observe that increased mobility randomness generally leads to a great increase in the order of the Coxian model as shown in Table III. However, the impact of this increase on the calculation speed of different metrics is limited due to the highly sparse representation of the Coxian distribution. Noting that each stage communicates only with its successor, the matrix fill-in is upper-bounded by $\frac{2}{m} * 100\%$, where m is the model order. This bound also applies to ECM whose order is scaled by the order of its PH stages. Furthermore, it is worth mentioning that two-phase distributions are usually sufficient for accurate data fitting [27], [40].

Figures 20-23 plot the impact of the MT mobility pattern variation on the derived metrics for different applications. An important observation is that both CM and ECM approximately have the same accuracy in all performance metrics except the network utilization times for which ECM provides better estimates as the mobility randomness increases. For example, ECM improves the cellular utilization estimation mismatch from 30% to 6% for $\alpha_v = 0$. On contrary, the handoff rates and forced termination probability estimates of both CM and ECM have similar accuracy level. Hence, we conclude that the exponential assumption is acceptable for transition related metrics such as handoff rates and forced

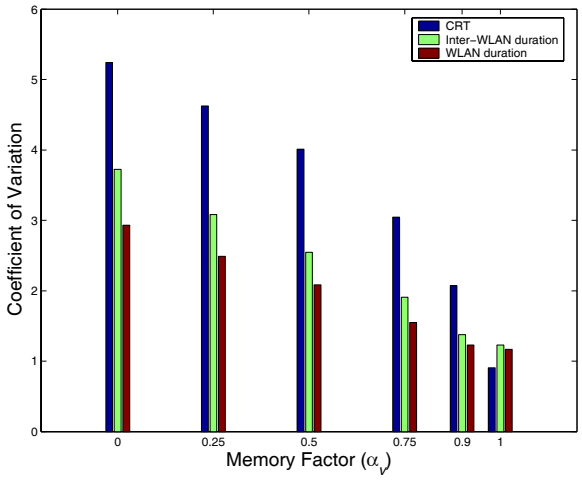


Fig. 19. Coefficient of variation of residence times

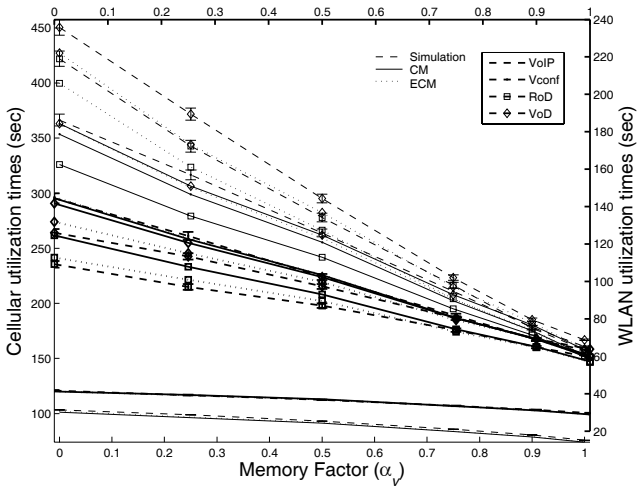


Fig. 20. Network utilization versus mobility randomness

termination probability, while the same assumption leads to inaccurate estimates for time based performance metrics such as utilization times.

Figure 20 shows a significant increase in technology utilization times as the mobility randomness increases, which is a logical consequence for the increase of different residence times. For example, the cellular and WLAN utilization of RoD is approximately doubled as the user mobility pattern changes from fluid-flow to random-walk. Hence, during the system design phase, these results should be considered when different system resources are allocated to different cells with different mobility patterns. Generally, as mobility randomness increases, more resources should be allocated for the cell. For example, cells dominated by random walkers, such as those in downtown locations, should be allocated more resources than other cells where fluid-flow travelers are expected, such as cells that cover highways.

Figure 21 shows a similar increasing trend in VHO signaling for most applications except for VoIP. This exception reflects the impact of the interaction between mobility and application characteristics. Clearly, as TRTs increase, the probability that

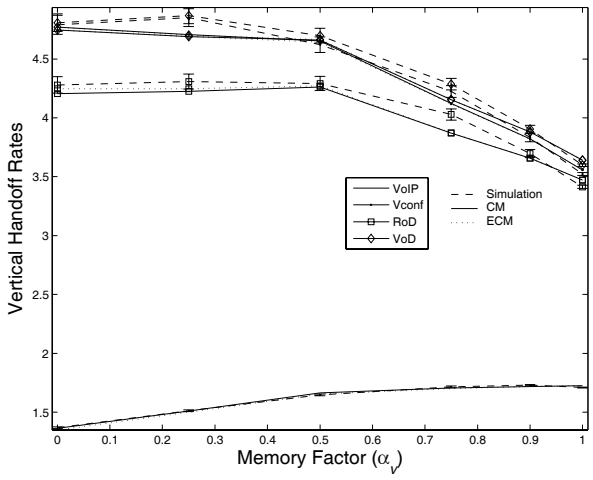


Fig. 21. VHO rate versus mobility randomness

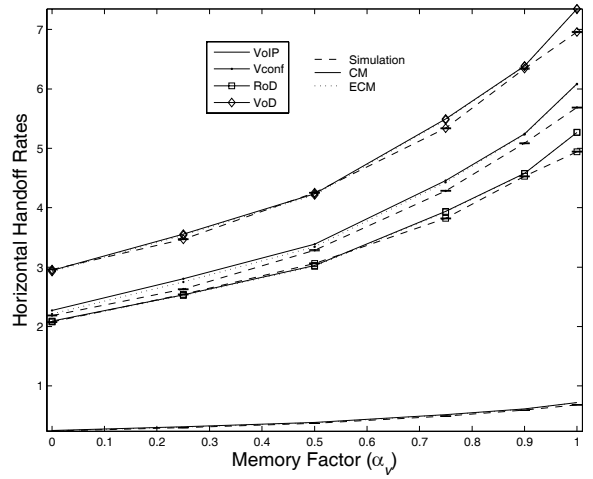


Fig. 22. HHO rate versus mobility randomness

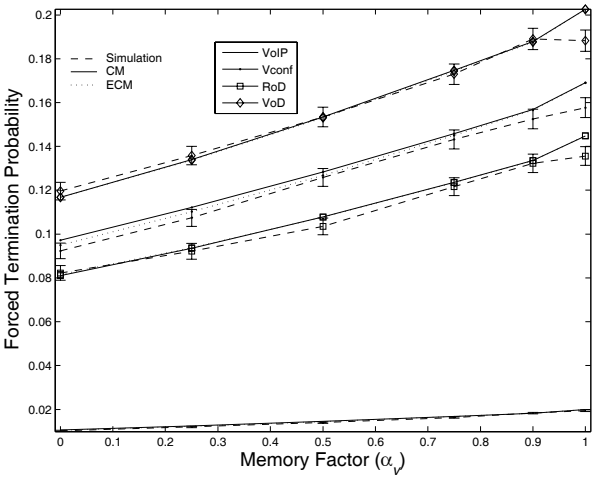


Fig. 23. Forced termination probability versus mobility randomness

a VoIP session ends without performing VHOs increases due to the comparatively shorter session duration. Hence, VoIP VHOs decreases as mobility randomness increases. On contrary, the signaling load of other applications is increased due to the increase of both the session duration and motion randomness. Finally, Figures 22-23 show that both the HHO rate and forced termination probability of different applications increase as mobility randomness decreases, which is a logical consequence for the decreasing residence times.

VI. Conclusion

Network heterogeneity is an intrinsic property of future-generation wireless networks due to the convergence of different access technologies to support diverse applications and services. This imposes design challenges that require novel mobility and analysis models to accommodate the evolving complexities in an integrated wireless system. In this work, we have presented three mobility models, IM, CM, and ECM, their corresponding application session models, and a stochastic analysis framework for performance evaluation in a two-tier heterogeneous wireless network, using the 3G-WLAN integration as an example architecture. We show that the Coxian random variable-based modeling approach accounts for the dependency between cell and WLAN residence times, leading to significant improvement in analysis accuracy. Furthermore, the simpler CM approach provides similar performance results as ECM for a wide range of mobility and traffic patterns, while ECM is more suitable in estimating the network utilization time for systems with highly random mobility patterns. Finally, using the proposed modeling and analysis methods, we have studied the impact of several important parameters on the system performance in terms of different metrics, providing insights and design guidelines for future-generation integrated heterogeneous wireless systems.

References

[1] M. M. Buddhikot, G. Chandranmenon, S. Han, Y. W. Lee, and S. M. L. Salgarelli, "Integration of 802.11 and third generation wireless data networks," in *Proc. of IEEE INFOCOM*, San Francisco, US, Apr. 2003, pp. 503 – 512.

[2] 3GPP2, "Feasibility study on 3GPP systems to wireless local area network (WLAN) interworking," 3GPP TR 22.934, Sep 2003.

[3] ETSI, "Requirements and architectures for interworking between HIPERLAN/3 and 3rd generation cellular systems," ETSI TR 101 957, Tech. Rep., Aug. 2001.

[4] C. Bettstetter, "Mobility modeling in wireless networks: categorization, smooth movement, and border effects," *ACM Mobile Comp. and Commun. Review*, vol. 5, no. 3, pp. 55–66, 2001.

[5] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 2, no. 3, pp. 257–269, July-Sept. 2003.

[6] E. Hyytia, P. Lassila, and J. Virtamo, "Spatial node distribution of the random waypoint mobility model with applications," *IEEE Trans. Mobile Comput.*, vol. 5, no. 6, pp. 680–694, June 2006.

[7] J.-Y. Le Boudec and M. Vojnovic, "Perfect simulation and stationarity of a class of mobility models," in *Proc. IEEE INFOCOM 2005*, vol. 4, 13-17 Mar. 2005, pp. 2743–2754.

[8] B. Liang and Z. J. Haas, "Predictive distance-based mobility management for multi-dimensional PCS networks," *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 718–732, October 2003.

[9] S. Wu, K. Y. M. Won, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks," *IEEE/ACM Trans. Netw.*, pp. 257–271, Apr 2002.

[10] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/dropping," *IEEE/ACM Trans. Netw.*, vol. 2, pp. 166–175, Apr 1994.

[11] L. Yin, B. Li, Z. Zhang, and Y.-B. Lin, "Performance analysis of a dual-threshold reservation (DTR) scheme for voice/data integrated mobile wireless networks," in *Proc. IEEE WCNC*, vol. 1, Chicago, USA, Sept 2000, pp. 258–262.

[12] M. M. Zonoozi and R. Dassanayake, "User mobility modeling and characterization of mobility patterns," *IEEE J. on Select. Areas in Commun.*, vol. 15, no. 7, pp. 1239–1252, Sep. 1997.

[13] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. 35, no. 3, pp. 77 – 92, Aug. 1986.

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

[15] Y. Fang, I. Chlamtac, and Y.-B. Lin, "Modeling PCS networks under general call holding time and cell residence time distributions," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 893–906, 1997.

[16] ——, "Call performance for a PCS network." *IEEE J. on Select. Areas in Commun.*, vol. 15, no. 8, pp. 1568–1581, 1997.

[17] ——, "Channel occupancy times and handoff rate for mobile computing and PCS networks," *IEEE Trans. Comput.*, vol. 47, no. 6, pp. 679–692, 1998.

[18] P. Orlik and S. Rappaport, "A model for teletraffic performance and channel holding time characterization in wireless cellular communications with general session and dwell time distributions," *IEEE J. on Select. Areas in Commun.*, vol. 16, no. 5, pp. 788–803, June 1998.

[19] M. Stemm and R. H. Katz, "Vertical handoffs in wireless overlay networks," *ACM Mobile Networks and Applications*, vol. 3, no. 4, pp. 335 – 350, 1998.

[20] B. Liang, S. Drew, and D. Wang, "Performance of multiuser network-aware prefetching in heterogeneous wireless systems," in press, to appear in *ACM/Springer Wireless Networks*.

[21] A. H. Zahran, B. Liang, and A. Saleh, "Modeling and performance analysis for beyond 3G integrated wireless networks," in *Proc. IEEE International Conference on Communications (ICC)*, vol. 4, June 2006, pp. 1819–1824.

[22] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* The Johns Hopkins University Press, 1981.

[23] G. Latouche and V. Ramaswami, *Introduction to Matrix analytic Methods in Stochastic Modeling*, ser. ASA-SIAM series on Statistics and Applied Probability. SIAM, 1999.

[24] M. W. Fackrell, "Characterization of matrix-exponential distributions," Ph.D. dissertation, Faculty of Engineering, Computer and Mathematical Sciences, University of Adelaide, Nov. 2003.

[25] D. Lam, D. C. Cox, and J. Widom, "Teletraffic modeling for personal communication services," *IEEE Commun. Mag.*, vol. 35, no. 2, pp. 79 – 87, Oct 1997.

[26] A. H. Zahran, "Modeling and design of next-generation heterogeneous wireless networks," Ph.D. dissertation, University of Toronto, 2007.

[27] D. R. Cox, *Renewal Theory*. Methuen and Co., Ltd., 1962.

[28] S. Asmussen, O. Nerman, and M. Olsson, "Fitting phase-type distribution via the EM algorithm," *Scand. J. Statist.*, vol. 23, pp. 419 – 441, 1996.

[29] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling of PCS networks," *IEEE J. on Select. Areas in Commun.*, vol. 17, no. 7, pp. 1062–1071, 1999.

[30] Y.-B. Lin, "Performance modeling for mobile telephone networks," *IEEE Network*, pp. 63–67, November 1997.

[31] A. B. Downey, "Evidence for long-tailed distributions in the internet," in *Proc. IMW '01: Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*. New York, NY, USA: ACM Press, 2001, pp. 229–241.

[32] M. Li, M. Claypool, R. Kinicki, and J. Nichols, "Characteristics of streaming media stored on the web," *ACM Trans. Inter. Tech.*, vol. 5, no. 4, pp. 601–626, 2005.

[33] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. McGraw-Hill, 2002.

[34] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queuing networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, 2nd ed. Wiley, August 1998.

[35] J. Wang, Q.-A. Zeng, and D. P. Agrawal, "Performance analysis of a preemptive and priority reservation handoff scheme for integrated service-based wireless mobile networks." *IEEE Trans. Mobile Comput.*, vol. 2, no. 1, pp. 65–75, 2003.

[36] A. H. Zahran, B. Liang, and A. Saleh, "Application signal threshold adaptation for vertical handoff in heterogeneous wireless networks," *ACM/Spring Mobile Networks and Applications (MONET), Special Issue on Soft Radio Enabled Heterogeneous Networks*, vol. 11, no. 4, pp. 625 – 640, Aug 2006.
[37] Stanford Pleiades Research Group, "Stanford university mobile activity traces (SUMATRA)." [Online]. Available: http://wwwdb.stanford.edu/sumatra
[38] M. McNett and G. M. Voelker, "Access and mobility of wireless pda users," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 9, no. 2, pp. 40–55, 2005.
[39] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," in *Proc. of MobiCom '04*. New York, NY, USA: ACM Press, 2004, pp. 187–201.
[40] S. M. Cox, "Hybrid stochastic models for remaining lifetime prognosis," Ph.D. dissertation, Graduate School of Engineering and Management, Air Force Institute of Technology, Air University, 2004.

Ahmed H. Zahran is a post-doctoral researcher at the computer science department at University College Cork. He received his PhD at the Department of Electrical and Computer Engineering, University of Toronto in 2007. He received his BSc and MSc in Electrical Engineering from Electronics and Electrical Communication Department at Faculty of Engineering, Cairo University in 2000 and 2002 respectively. His research interests span different topics in wireless mobile networking such as network architecture, mobility and resource management, and modeling and performance evaluation. He won the best paper award in IFIP Networking 2005 conference.

Ben Liang received honors simultaneous B.Sc. (valedictorian) and M.Sc. degrees in electrical engineering from Polytechnic University in Brooklyn, New York, in 1997 and the Ph.D. degree in electrical engineering with computer science minor from Cornell University in Ithaca, New York, in 2001. In the 2001 - 2002 academic year, he was a visiting lecturer and post-doctoral research associate at Cornell University. He joined the Department of Electrical and Computer Engineering at the University of Toronto in 2002, where he is now an Associate Professor. His current research interests are in mobile networking and multimedia systems. He won an Intel Foundation Graduate Fellowship in 2000 toward the completion of his Ph.D. dissertation and an Early Researcher Award (ERA) given by the Ontario Ministry of Research and Innovation in 2007. He was a co-author of the Best Paper Award at the IFIP Networking Conference in 2005 and the Runner-up Best Paper Award at the International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks in 2006. He is an associate editor of *Wiley Security and Communication Networks* and serves on the organizational and technical committees of a number of conferences including ACM MobiCom, IEEE INFOCOM, IEEE MASS, and IEEE SECON. He is a senior member of IEEE and a member of ACM and Tau Beta Pi.

Aladdin Saleh earned his Ph.D. degree in Electrical Engineering from London University, England. Since March 1998, Dr. Saleh has been working in the Wireless Technology Department of Bell Canada, the largest service provider of wireless, wire-line, and Internet in Canada. He worked as a senior application architect in the wireless data group working on several projects among them the wireless application protocol (WAP) and the location-based services. Later, he led the work on several key projects in the broadband wireless network access planning group including planning of the IEEE 802.16/ Wimax, the IEEE 802.11/ WiFi, and the integration of these technologies with the 3G cellular network including Mobile IP (MIP) deployment. Dr. Saleh also holds the position of Adjunct Full Professor at the Department of Electrical and Computer Engineering of Waterloo University, Canada since January 2004. He is currently conducting several joint research projects with the University of Waterloo and the University of Toronto on IEEE 802.16-Wimax, MIMO technology, interworking of IEEE 802.11 WLAN and 3G cellular networks, and next generation wireless networks. Prior to joining Bell Canada, Dr. Saleh worked as a faculty member at different universities and was Dean and Chairman of Department for several years. Dr. Saleh is a Fellow of IEE and a Senior Member of IEEE.