An Integrated Mobility and Traffic Model for Resource Allocation in Wireless Networks

Hisashi Kobayashi
Dept. of Electrical Eng.
Princeton University
Princeton, NJ 08544
hisashi@ee.princeton.edu

Shun-Zheng Yu
Dept. of Electrical Eng.
Princeton University
Princeton, NJ 08544
syu@ee.princeton.edu

Brian L. Mark
Elect. and Comp. Eng. Dept.
George Mason Univerity
Fairfax, VA 22030
bmark@gmu.edu

ABSTRACT
In a wireless communications network, the movement of mobile users presents significant technical challenges to providing efficient access to the wired broadband network. In this paper, we construct a new analytical/numerical model that characterizes mobile user behavior and the resultant traffic patterns. The model is based on a semi-Markov process representation of mobile user behavior in a general state-space. Using a new algorithm for parameter estimation of a general Hidden Semi-Markov Model (HSMM), we develop an efficient procedure for dynamically tracking the parameters of the model from incomplete data. We then apply our integrated model to obtain estimates of the computational and bandwidth resources required at the wireless/wired network interface to provide high performance wireless Internet access and quality-of-service to mobile users. Finally, we develop a threshold-based admission control scheme in the wireless network based on the velocity information that can be extracted from our model.

Keywords
wireless networks, mobility, traffic modeling, resource allocation, admission control

1. INTRODUCTION
In a wireless communications network, the movement of mobile users presents significant technical challenges to providing efficient wireless access to the Internet. For an individual mobile user, the point of contact to the wired network changes with time. It is therefore imperative to be able to track and to take into account dynamic mobile behavior when allocating resources to traffic at the interface between the wireless and wired networks.

Construction of mobility patterns for analysis and simulation has attracted considerable attention in recent years (see e.g., [2, 13, 12]). Mobility models find application in geolo-

cation, the measurement of location information for mobile users. For example, mobility models can be used to compute how frequently geolocation of the mobile should be done. Given the cost of geolocation, which consists of the signaling delay and overhead for each geolocation transaction, we may wish to compute the probability of failure in reaching all mobiles that are in a target area.

Chen [2] proposes a cellular-based location tracking system which utilizes the estimated distance between the mobile and the referenced base station, together with sector information and employs a Kalman filter for location estimation. Maass [13] develops a location information server based on directory data models and services. Liu and Maguire [12] propose a mobility management based on two algorithms: one algorithm for detecting and storing the regular itinerary patterns of the user and the second algorithm for predicting the next state of movement of the user. Other references on dynamic location tracking include [9, 4, 11, 5, 19]. These works focus on modeling mobile location at the physical level in order to reduce location updating and paging signaling cost.

Several works have modeled mobile behavior as a random walk or Brownian motion [20] on two-dimensional or three-dimensional (to model mobility in a multi-story building) grids. Such models can be used to drive simulation models of the wireless network. The street map of a city or the blueprint of a building can be used to provide input for the degree of freedom of realistic mobility patterns. In [10], a stochastic model for mobility called the Markovian highway Poisson arrival location model (PALM) is introduced and developed rigorously. This model uses a pair of coupled partial differential equations or ordinary differential equations to describe the evolution of the system.

In this paper we introduce a new integrated model of mobility and traffic that differs from existing work in two key aspects: 1) The model allows us to exploit recent results in the theory of queueing and loss networks [8] to reduce significantly the amount of information that needs to be tracked and stored; 2) The tracking model can be implemented in real-time using a computationally efficient parameter estimation algorithm that has been invented recently [24]. Our model is based on an underlying semi-Markov chain. A new method for estimating the parameters of an arbitrary hidden semi-Markov model (HSMM), makes it feasible to char-

acterize the macroscopic mobility and traffic behavior in the wireless network.

We apply the new model to the important problem of efficient resource allocation in wireless networks. First, we show how the mobility information obtained from our model can be used in an adaptive admission control scheme that improves the blocking probability of in-progress calls. Second, we show how traffic information obtained from the model can be used to estimate the amount of bandwidth that should be reserved for wireless traffic at the wireless/wired network interface.

The remainder of the paper is organized as follows. Section 2 develops our integrated model of user mobility and traffic in the wireless network. Section 3 discusses a novel algorithm for estimating the parameters of the model dynamically. Sections 4 and 5 discuss applications of the mobility and traffic model, respectively, to resource allocation at the network interface and to admission control in the wireless network. Section 6 discusses a numerical example that illustrates the mobility/traffic state estimation. Finally, Section 7 concludes the paper.

2. MOBILITY AND TRAFFIC MODEL

2.1 Abstract Mobility State Space

We define the state of a mobile user in terms of a vector $(x_1, \cdots, x_n)$, where the ith component, x_i, represents a value from a finite *attribute* space $\mathcal{A}_i$. The attribute spaces represent properties of the mobile user such as location, moving direction, speed, etc. The set of possible states for a mobile user is an n-dimensional vector space given by

$$\mathcal{S} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n,$$

where $\times$ denotes the Cartesian product. The abstract space $\mathcal{S}$ can be made as rich as desired by including the appropriate attributes as components in the state vector. The dynamic motion of a user, as defined by its time-varying attribute values, can then be described by its trajectory in this space.

We enumerate all possible states in $\mathcal{S}$ and label them as $1, \cdots, M$ such that the state space $\mathcal{S}$ can more simply be represented as follows:

$$\mathcal{S} = \{1, \cdots, M\}.$$

We introduce two *inactive* states in addition to the set of *active* states $\mathcal{S}$: the *source* state s and the *destination* state d. A user enters the system by assuming the state s. A user exits the system by assuming the state d. Thus, the user can assume states in the augmented state-space $\tilde{\mathcal{S}} = \mathcal{S} \cup \{s, d\}$. The state transitions of a user are characterized by a Markov chain with transition probability matrix $A = [a_{nm} : n, m \in \tilde{\mathcal{S}}]$.

No transitions occur from states $j \in \mathcal{S}$ to the source state, i.e., $a_{js} = 0$. From any such state j, the user next assumes the destination state d with probability a_{jd}. No transitions are allowed from the destination state. Hence, the state d is considered to be the *absorbing* state of the Markov chain. Further, no transitions occur from state s to state d, i.e., $a_{sd} = 0$. $\tilde{\mathcal{S}} = \mathcal{S} \cup \{s, d\}$. The transition probability matrix thus has the following form:

$$\tilde{A} = \begin{array}{c} d \\ s \\ 1 \\ \vdots \\ M \end{array} \left[\begin{array}{cccccc} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & a_{s1} & \cdots & a_{s,M} \\ a_{1,d} & 0 & a_{11} & \cdots & a_{1,M} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{M,d} & 0 & a_{M,1} & \cdots & a_{M,M} \end{array} \right].$$

We allow the dwell time of a user in state $m \in \mathcal{S}$ to be generally distributed with mean $\bar{d}_m$. Hence, the state process of a user is, in general, a semi-Markov chain. The aggregate behavior of the system of mobile users can be represented by the vector process

$$\mathbf{N}(t) = (N_1(t), \cdots, N_M(t)), \tag{1}$$

where $N_m(t)$ represents the number of mobile users in state m at time t. Given the assumptions above, $\mathbf{N}(t)$ is also a semi-Markov chain. We further make the assumption that users arrive to the system in state s according to a Poisson process. In general, the average arrival rate, $\Lambda(N)$ may be a function of the current system population:

$$N = ||\mathbf{N}|| = \sum_{m \in \mathcal{S}} N_m. \tag{2}$$

Observe that the above system is equivalent to an open queueing network with M infinite-server stations corresponding to the states in $\mathcal{S}$. Clearly, the source and destination stations of the queueing network correspond to s and d, respectively. Results from the theory of queueing and loss networks [8] show that the steady state distribution of $\mathbf{N}(t)$ is insensitive to the distributions of the dwell times at each station. Furthermore, the steady-state distribution is given by a simple product form solution:

$$P[\mathbf{n}] = P\{\mathbf{N}(t) = \mathbf{n}\} = P[0]\Lambda(\mathbf{n}) \prod_{s \in \mathcal{S}} \frac{(e_m \bar{d}_m)^{n_m}}{n_m!}, \tag{3}$$

where

$$P[0] = \left[\sum_{\mathbf{n} \geq 0} \Lambda(\mathbf{n}) \prod_{s \in \mathcal{S}} \frac{(e_m \bar{d}_m)^{n_m}}{n_m!} \right]^{-1}, \tag{4}$$

and the values e_m satisfy the following equations:

$$e_m = a_{sm} + \sum_{j \in \mathcal{S}} e_j a_{jm}, \quad m \in \mathcal{S}. \tag{5}$$

The value e_m can be interpreted as the average number of visits that a user makes to state m during its sojourn in the system.

Our proposed abstract mobility state space model differs from other proposed mobility models (cf. [20, 10] in that it leads to a simple parametric representation of the mobile behavior that can be related to a general queueing network with multi-class users in which each service center is infinite server (IS) with multiple types. This representation allows us to capitalize on recents results in queueing and loss network theory [8] which show that the steady-state distribution is surprisingly robust to all state time distributions and state transition behaviors. This result in turn implies that to obtain the state distribution of mobile users, we need

only have two sets of parameters: the mean dwell time, $\bar{d}_m$, in state m and the expected number of visits, e_m, the user makes to state m in its lifetime per user class. Thus, only $2M$ pieces of numeric data per user class provide sufficient statistics of the user mobility, as far as the steady-state distribution and related performance measures are concerned. This data can be estimated by means of a new parameter estimation algorithm to be discussed in Section 3.

We can augment the basic mobility model by introducing state-dependent information. Let $\mathcal{J} = \{1, \cdots, J\}$ represent a set of user requirements. We shall suppose that a mobile user in state m requires data of type j (e.g., web content of a certain type) from the network with probability $c_m(j)$. Alternatively, the requirement j could represent the network resources (e.g., bandwidth) that a user requires to transmit or receive a certain type of real-time stream (e.g., real-time video).

Dynamic information on user traffic can be integrated into the basic mobility model via an appropriate specification of the mobility attributes and/or the user requirements. Thus, we can incorporate both mobility and traffic information in a single integrated model. The generality of the model allows it to be applied in a variety of ways to enhance network performance. As we discuss in Sections 4 and 5, the model can be applied to improve resource allocation at the wireless/wired network interface and in the wireless network itself.

2.2 Practical Model Realizations

Any practical realization of the general model should balance the desire for model accuracy with considerations of computational complexity. Suppose that we are primarily interested in tracking the mobility of a user within a certain geographic region. Geolocation measurement accuracy may be as high as 20 m or 100 m, but for the purposes of mobility modeling, the resolution need not be that high. For smooth handoff, it is usually sufficient to consider a small number of ranges of speed. Similarly, a handful of direction attributes should be sufficient for most applications. As an example, a geographic area might be subdivided into about one hundred important locations. The location space of mobile user locations could then be represented as follows:

$$\mathcal{L} = \{A_1, A_2, \cdots, A_{100}\}$$

We may specify the feasible directions of movement as follows:

$$\mathcal{D} = \{\text{north}, \text{south}, \text{east}, \text{west}\}.$$

The speed ranges of interest are given as follows:

$$\mathcal{V} = \{\text{stationary}, \text{walking}, \text{city driving}, \text{highway driving}\}.$$

Then the system state-space for this example would be given by

$$\mathcal{S} = \mathcal{L} \times \mathcal{D} \times \mathcal{V}.$$

The total number of states, M, for this example will be on the order of one thousand. We note, however, that transitions among the states is limited and we may assume that from a given state transitions can occur to on the order of

ten neighboring states. For example, suppose that the geographic area of interest is represented as a ten by ten grid of the 100 squares $A_1, A_2, \cdots, A_{100}$ in the set $\mathcal{L}$. Each square, A_j, has at most four neighbors. If we consider the location and direction attributes together, i.e., the Cartesian product $\mathcal{L} \times \mathcal{D}$, we observe that each (location, direction) pair has exactly one neighbor.

Such considerations imply that the transition probability matrix will be highly sparse in practical applications. As will be discussed below, our model tracking algorithm has complexity on the order of the number of matrix elements, which should be significantly less than the worst-case of M^2. This makes our general model amenable to practical implementation.

3. DYNAMIC STATE TRACKING
3.1 Hidden Semi-Markov Model
The general mobility model was discussed in the context of a continuous-time parameter t. In practice, tracking of the system parameters must be based on measured observations sampled at discrete time instances. Therefore, we shall represent the user dynamics by a discrete-time semi-Markov chain, where the t is now discrete, taking values in $\{0, 1, 2, \cdots\}$. Furthermore, the system states cannot, in general, be observed directly, i.e., the states are *hidden*. Hence, an appropriate model for the system is a discrete-time *Hidden Semi-Markov Model* (HSMM).

As in the continuous-time model, the evolution of the user state is characterized by a state transition probability matrix denoted by

$$\mathbf{A} = [a_{ij} : i, j \in \tilde{\mathcal{S}}]. \tag{6}$$

We shall assume that the mobile user dwell time in a given state is a random variables taking values in the set $\{1, \cdots, D\}$, with probability distribution function denoted by $p_m(d)$, $d = 1, \cdots, D$. We introduce the $M \times D$ matrix

$$\mathbf{P} = [p_m(d) : m \in \tilde{\mathcal{S}}, \ d = 1, \cdots, D]. \tag{7}$$

As discussed earlier, we characterize the user requirements in terms of a finite set $\mathcal{J} = \{1, \cdots, J\}$ and a requirements probability distribution matrix:

$$\mathbf{C} = [c_m(j) : m \in \tilde{\mathcal{S}}, \ j \in \mathcal{J}]. \tag{8}$$

The matrices $\mathbf{A}$, $\mathbf{P}$ and $\mathbf{C}$ constitute an analytical discrete-time semi-Markov model that captures the dynamic mobility and requirements of a given user.

In order to track user mobility, the parameters of the semi-Markov model must be estimated based on observations of the user state. This leads to a Hidden Semi-Markov Model (HSMM) described as follows. Let $S_t \in \{1, \cdots, M\}$ denote the state of the user at time t is the discrete time parameter, i.e., t takes values in $\{0, 1, 2, \cdots\}$. We denote the sequence of states from time a to time b as $S_a^b = \{S_a, S_{a+1}, \cdots, S_b\}$. Let $\pi = [\pi_m]$, $m = 1, \cdots, M$, be the initial state probability distribution vector, where π_m denotes the probability that the initial state of the user is state m.

Let o_t denote the value of an observation of the user state at time t. We assume that there are K distinct state observation values, $1, \cdots, K$. The sequence of observations from

time a to time b is denoted by o_a^b. Note that the observation value o_t is generally different from the true state S_t, due to geolocation and estimation errors. We define the following observation probability distribution matrix:

$$\mathbf{B} = [b_m(k) : m \in \tilde{\mathcal{S}}, \ k = 1, \cdots, K], \qquad (9)$$

where $b_m(k)$ denote the probability that the observed value at an arbitrary time t is $o_t = k$, given that the actual user state is $S_t = m$. The observation of the user requirements at time t is denoted by $q_t \in \mathcal{J}$. The corresponding requirement observation sequence from time a to b is denoted by q_a^b. The 5-tuple $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{P}, \pi)$ provides a complete specification the discrete Hidden Semi-Markov Model for the system.

To track the state of a mobile user, we apply the forward-backward and re-estimation algorithms for HSMM parameter estimation to be discussed in Section 3. The main steps of the tracking algorithm are summarized as follows:

1. Apply the *HSMM re-estimation algorithm* to obtain initial estimates $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \hat{\mathbf{P}}, \hat{\pi})$, of the HSMM model parameters by using training data.

2. Apply the *HSMM forward-backward estimation algorithm* to predict at time t the next requirement, q_{t+1}, of the mobile user, based on the geolocation and requirement observation sequences o_1^t and q_1^t, respectively. Find the maximum likelihood state sequence, s_1^T, for given observation sequences q_1^T and o_1^T, where T is the active period of the mobile user.

3. Obtain refined estimates, $(\hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k, \hat{\mathbf{C}}_k, \hat{\mathbf{P}}_k, \hat{\pi}_k)$, by applying the HSMM re-estimation algorithm to the given observation sequences.

Figure 1 illustrates the dynamic mobility tracking model. The mobile user generates the "true" state sequence S_1^T. The observation sequence o_1^T is obtained from geo-location measurement and tracking. A server attached to the wired network records the user requirements, producing the sequence q_1^T. The sequences o_1^T and q_1^T are inputs to the HSMM parameter estimation algorithm. Finally, the HSMM parameter estimation algorithm produces estimates, ($\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$, $\hat{\mathbf{P}}$, $\hat{\pi}$), of the model parameters and an estimate, $\hat{S}_1^T$, of the user state sequence. In addition, a prediction, $\hat{q}_{t+1}$, of the next user requirement, is produced as an output. This information can be used to anticipate future Internet document requests from the user. Thus, the mobility model can be used to enhance the performance of prefetch caching algorithms [23, 22].

3.2 Estimation from Insufficient Data

Estimation of the mobility model parameters must in general be made based on missing data. Due to physical constraints, geolocation measurement and/or transmission of geolocation data may not take place frequently enough to allow precise tracking of the user's state at all times. The task of estimation from insufficient data involves two important aspects: (a) estimation and prediction of the users' moving behaviors and requirements; (b) re-estimation of the model parameters based on missing data.
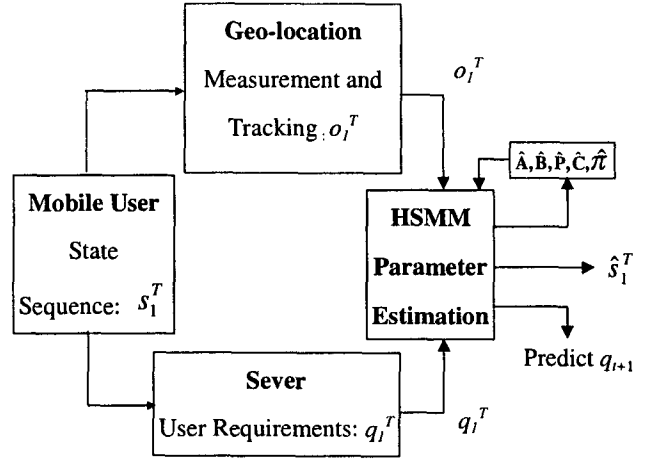

Figure 1: Dynamic mobility/traffic state tracking model.

There are two different cases for missing data problem. The first case is when we know that a state occurs but have no observation. In this case, we can assume that there is a complete observation sequence mixed with an independent random erasure process. Hence this case can be modeled as a discrete hidden Markov model (HMM) with an erasure process. The second case is that we do not know when a state transition occurs because of missed observations. In this case, we do not know how many state transitions occur during the interval of missing observations. Therefore, we should explicitly consider the state duration so that we can estimate the maximum likelihood state sequence including the missed period. This case should be modeled as a hidden semi-Markov model (HSMM) with missing data, where the state duration has some general probability distribution.

The key issues in dealing with such an HSMM are: (a) finding an efficient algorithm for estimating the state sequence and for re-estimating the model parameters based on missing data; (b) proving that the proposed algorithm provides the best estimates, i.e., maximum likelihood estimates. The well-studied HMM can be viewed as a special case of the HSMM. Similarly, an HSMM with complete observation data can be treated as a special case of an HSMM with partial observation data.

An HSMM is more general than an HMM since the latter model that assumes either a constant or a geometrically distributed dwell time (cf. [17]). Although the statistical literature addresses estimation procedures for missing data, a computationally feasible algorithm has not previously been reported for an HSMM with erasures. The well known Baum-Welch algorithm [21] applies only to the HMM.

The main elements of the HSMM parameter estimation algorithm. A detailed development of the algorithm and its validation by simulation are reported in [24]. Recall that the HSMM is specified by a 5-tuple $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{P}, \pi)$. The observation interval is assumed to be segmented into T subintervals indexed by $1, 2, \cdots, T$. Observations may not necessarily be available in each of the T subintervals. We de-

note the set of observation time instents by $\mathcal{G} = \{t_1 = 1, t_2, t_2, \cdots, t_n = T\}$.

3.2.1 Forward-Backward algorithm

In [24], a forward-backward algorithm has been devised to estimate an HSMM from observations with erasures. The algorithm has a computational complexity proportional to D, where D is the maximum value of the dwell time for all states. The more general forward-backward algorithm reduces to the Baum-Welch algorithm when $D = 1$. We note that the algorithm offers a significant improvement over an earlier algorithm by Ferguson (1980) [3] which has computational complexity proportional to D^2.

We define the *forward variables* (cf. [3]) as follows:

$$\alpha_t(m) = P[o_1^t, \text{ state } m \text{ sojourn ends at } t], \quad t \geq 1$$
$$\alpha_t^*(m) = P[o_1^t, \text{ state } m \text{ sojourn begins at } t+1], \quad t \geq 1.$$

The *backward variables* are defined by:

$$\beta_t(m) = P[o_t^T | \text{sojourn in state } m \text{ begins at } t], \quad t \leq T,$$
$$\beta_t^*(m) = P[o_t^T | \text{sojourn in state } m \text{ ends at } t-1], \quad t \leq T.$$

The forward variables are then computed inductively for $t = 1, 2, \cdots, T$ [24]. Similarly, the backward variables are computed inductively for $t = T, T-1, \cdots, 1$. After computing the forward and backward variables, the maximum a posterior (MAP) state estimate can be found. Define:

$$\gamma_t(m) = P[o_1^T; s_t = m]. \tag{10}$$

Then the MAP estimate of s_t is given by

$$\hat{s}_t = \arg \max_{1 \leq m \leq M} \frac{\gamma_t(m)}{P[o_1^T]}, \quad t = T, T-1, \cdots, 1. \tag{11}$$

3.2.2 Re-estimation algorithm

A simple iterative procedure for re-estimating the HSMM parameters is reported in [24]. By applying the well-known EM (Expectation/Maximization) algorithm [21], it can be shown that this iterative procedure is increasing in likelihood. The overall computational complexity of the re-estimation algorithm is essentially proportional to T. Thus, the parameters for the HSMM model can be estimated efficiently within the framework of dynamic mobility model tracking illustrated in Figure 1.

4. RESOURCE ALLOCATION

The information obtained from the mobility and traffic model can be used to characterize the traffic streams arriving from mobile users at the wireless/wired interface. Traffic characterization is a necessary step in determining the amount of network resource that should be allocated for each user in order to meet their requirements on quality-of-service. We consider consider two main types of user traffic: 1) user requests for data (e.g., web content) from the wired network; 2) real-time or non-real-time data transmission from the user to the wired network. The network interface should allocate sufficient computational resources to process user requests with a low probability of losing requests. The network interface should also allocate sufficient bandwidth and buffer resources to provide QoS for transmissions from the mobile user. Using the mobility and traffic model, we shall obtain

estimates on the amount of network resource that should be allocated in both cases.

4.1 Overall state transition rate

Let us examine how often the user state transitions occur in the HSMM model. Define the vector, $\mathbf{a}$, of state transition probabilities from the source state s to the states in $\mathcal{S}$:

$$\mathbf{a} = (a_{sj} : j \in \mathcal{S}). \tag{12}$$

and the submatrix, $\mathbf{A}_s$, of the overall transition probability matrix $\mathbf{A}$, which characterizes the state transitions within the set of active states $\mathcal{S}$:

$$\mathbf{A}_a = [a_{mn} : m, n \in \mathcal{S}]. \tag{13}$$

It is convenient at this point to introduce a special *inactive* state, denoted 0, which subsumes the roles of the states s and d in a single state. The state 0 may be considered to consist of two substates s and d. The associated state process is an absorbing Markov chain. with fundamental matrix given by [7]:

$$\mathbf{F} = \left(\mathbf{I} - \begin{bmatrix} 0 & \mathbf{a} \\ 0 & \mathbf{A}_a \end{bmatrix}\right)^{-1} = \begin{bmatrix} 0 & \mathbf{q}_0 \\ 0 & \mathbf{Q} \end{bmatrix}, \tag{14}$$

where $\mathbf{I}$ denotes the identity matrix, 0 denotes a column vector of zeros and $\mathbf{q}_0$ is defined by:

$$\mathbf{q}_0 = \mathbf{a}\mathbf{Q}. \tag{15}$$

The element q_{mn} of $\mathbf{Q}$, $(m, n \in \mathcal{S})$, is the expected number of visits to state n that a user makes starting from state m until the user is finally absorbed into the destination substate d, and the element q_{0n} of vector bq_0, $(n \in \mathcal{S})$ is the expected number of visits to state n that the user makes during its active period starting from the source substate s until reaching the destination substate d. When the user reaches the destination substate d, it immediately transits to the source substate s. The dwell time of a user in state 0 is denoted by $\bar{d}_0$ and has a general distribution.

The total expected number of state transitions that the user makes during its active period is given by

$$S = \sum_{n \in \mathcal{S}} q_{0n}, \tag{16}$$

and the expected total active time of a mobile is:

$$T = \sum_{n \in \mathcal{S}} q_{0n} \bar{d}_n. \tag{17}$$

The total expected state transition rate of a mobile user is given by $\lambda = S/T$, which provides a measure the amount of system resources required for storing and transferring mobility tracking information.

4.2 User request rate

Let $\bar{N}_m$ denote the expected number of users in state m in equilibrium $(m = 0, 1, \cdots, M)$. The mean departure rate from state m is given by

$$\gamma_m = \bar{N}_m / \bar{d}_m = \sum_{j=0}^{M} \bar{N}_j a_{jm} / \bar{d}_j, = \bar{N}_0 / \bar{d}_0 q_{0m} \tag{18}$$

43

where $\gamma_0 = \bar{N}_0/\bar{d}_0$ is the total rate at which mobile users transit from state 0 to an active state, i.e., the total rate of entry into the system. The state process, $N(t) = (N_1(t), \cdots, N_M(t))$, corresponds to the vector process $N(t)$ defined in Eq.(1).

The state transition rate, i.e., the expected rate that the active mobile users request state (or geolocation)-dependent content when they transit from one state to another, is given by

$$R_s = \sum_{m \in S} N_m/\bar{d}_m, \tag{19}$$

where the state transitions from state 0 to active states are included, but the transitions from active states to the state 0 are not included.

We assume that the mean rate at which a mobile user requests geolocation-independent content while it stays in state m is R_m. Then the overall request rate for geolocation-independent content is

$$R_r = \sum_{m=1}^{M} R_m N_m. \tag{20}$$

Therefore, the total request rate is

$$R = R_s + R_r = \sum_{m=1}^{M} (\bar{d}_m^{-1} + R_m)N_m. \tag{21}$$

Define

$$R(t) = \sum_{m=1}^{M} (\bar{d}_m^{-1} + R_m)N_m(t), \tag{22}$$

where $N(t)$ is now interpreted as a multiple-state Markov modulated rate process (MMRP). Since we allow the dwell times to have a general distribution, $N(t)$ is actually a *semi-Markov* modulated rate process, which extends the model studied in [18]. Let $X(t)$ be an M-dimensional diffusion process approximating $N(t)$. Under a set of reasonable assumptions, $X(t)$ can be expressed as an M-dimensional Ornstein-Uhlenbeck (O-U) process. Hence, the process $R(t)$ can be approximated by a Gaussian process

$$\tilde{R}(t) = \sum_{m=1}^{M} (\bar{d}_m^{-1} + R_m)X_m(t), \tag{23}$$

and the total number of active users, $K(t)$, is approximated by

$$\tilde{K}(t) = \sum_{m=1}^{M} X_m(t), \tag{24}$$

which is also a Gaussian process.

Suppose that the wired/wireless gateway can process at most C requests per second. Then the change of the queue length $Q(t)$ can be represented by the stochastic differential equation:

$$\frac{dQ}{dt} = \begin{cases} \tilde{R}(t) - C, & \text{if } \tilde{R} > C \text{ or } Q(t) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

As discussed in [18], the asymptotic complementary queue length distribution can be approximated by

$$P\{Q > x\} \approx \left(\frac{e^{-\theta^2/2}}{\theta\sqrt{2\pi}} \right) \exp\left(-\frac{2\sigma_{\tilde{R}}\theta_{\tilde{R}}}{a_{\tilde{R}}}x \right), \tag{26}$$

where $\sigma_{\tilde{R}}$ is the standard deviation of the rate process $\tilde{R}(t)$ in equilibrium. The parameter θ is given by:

$$\theta_{\tilde{R}} = \frac{C - \mu_{\tilde{R}}}{\sigma_{\tilde{R}}}, \tag{27}$$

where $\mu_{\tilde{R}}$ is the mean of the rate process in equilibrium. The constants $\mu_{\tilde{R}}$, $\sigma_{\tilde{R}}$ and $a_{\tilde{R}}$ (in Eq. (26)) can easily be computed from the transition probabilities and rate values governing the rate process $\tilde{R}(t)$. Let B denote the number of requests that can be held in a buffer. Requests are lost whenever the length. $Q(t)$, of the request queue, exceeds B. In this case, the steady-state loss probability can be approximated by $P\{X > B\}$.

4.3 User Traffic

Information about the user's traffic patterns can be extracted from the mobility model and used to characterize the aggregate traffic stream arriving from the wireless network at the network interface. By developing an appropriate traffic model, we can obtain an approximation for the required bandwidth to satisfy the QoS requirements at the interface to the wired network (see [14, 15]). The required bandwidth can then be used to make admission control decisions and to set the parameters of the scheduler at the network interface.

In constructing the mobility model, we can specify an attribute specifying the current transmission rate of a mobile. The associated attribute space, $\mathcal{R}$, consists of a number of discrete rates:

$$\mathcal{R} = \{r_0 = 0, r_1, \cdots, r_R\}. \tag{28}$$

We can then define a function $r(m)$ that maps a user state $m \in S$ to the corresponding transmission rate. Let $\mathcal{B} = \{1, 2, \cdots, B\}$ denote the set of base stations in the network. We define $\mathcal{B}_j \subseteq S$ to be the subset of states in which the mobile user is connected to base station $j \in \mathcal{B}$.

With these preliminaries, we can express the aggregate mobile user traffic stream arriving at base station $j \in \mathcal{B}$ as follows:

$$R_j(t) = \sum_{m \in \mathcal{B}_j} N_m(t)r(m). \tag{29}$$

We observe that $R_j(t)$ has the form of a multiple-state Markov modulated rate process (MMRP). The process $R_j(t)$ differs from the standard MMRP in that the sojourn times in each state may be generally distributed. Such a process can be approximated using an Ornstein-Uhlenbeck diffusion process.

Suppose that the quality-of-service (QoS) requirement at base station j is that the packet loss rate should be less than ϵ_j. The diffusion process approximation leads to a simple form for the required bandwidth at base station j [18]:

$$C_j = \mu_j + \theta_j\sigma_j, \tag{30}$$

44

where μ_j and σ_j are, respectively, the mean and variance of $R_j(t)$ in steady-state. The parameters μ_j and σ_j can be computed from the parameters of MMRP $R(t)$. The parameter θ_j can be computed in two ways. If the network interface has only a small number of buffers available to the wireless traffic at the network interface, then the multiplexing system can be modeled as a loss system. In this case, the expression for θ_j is as follows [18]:

$$\theta_j = 1.8 - 0.46 \log_{10}\left(\frac{\mu_j\sqrt{2\pi}}{\sigma_j}\epsilon_j\right).$$

Alternatively, if the number, B_j, of available buffers at the network interface is sufficiently large, then the network interface can be modeled as a multiplexer with an infinite buffer. In this case, the packet loss probability can be approximated by the probability that the queue length $Q_j(t)$ exceeds B_j. In this case, the diffusion approximation leads to a required bandwidth of the form (30), but with θ_j given as follows [18]:

$$\theta_j = \sqrt{\psi_j - 2\ln\left(\sqrt{\psi_j} - \frac{2\sigma_j}{a_j}B_j\right)} - \frac{2\sigma_j}{a_j}B_j, \quad (31)$$

$$\psi_j = -2\ln(\sqrt{2\pi}\epsilon_j) + \frac{4\sigma_j^2}{a_j^2}B_j^2. \quad (32)$$

The parameter a_j can be computed from the parameters of the MMRP $R(t)$ (see [18]). In general, C_j tends to be a conservative estimate of the bandwidth that should be set aside at base station j for real-time mobile traffic. The estimate of required bandwidth could be further refined using traffic measurements at the base station (cf. [15]).

5. ADMISSION CONTROL

In the wireless network, the service area is divided into cells in order to distribute the allocation of network resources among multiple base stations. Nonadjacent cells share frequency channels to make efficient use of the limited spectrum allocated for mobile communication services. When a mobile user attempts a new call in a given cell, one of the available channels associated with the cell is allocated to it. If no channels are available, the call is blocked. After a call is established within a given cell, the mobile user may move to an adjacent cell while the call is in progress. In this case, the call must be handed off to the neighboring cell in order to provide uninterrupted service to the mobile user. If no channels are available in the new cell, the handoff attempt is blocked.

A major issue in resource management for wireless networks is to develop efficient schemes for channel allocation that maximize channel utilization subject to the satisfaction of quality-of-service (QoS) requirements. The typical QoS metrics include new call blocking probability, handoff failure probability, and handoff delay. Various channel allocation schemes have been proposed and analyzed in the literature (see e.g., [6, 16]). A relatively simple scheme for admission control is related to trunk reservation, whereby a pool of guard channels in the cell is reserved for the handoff calls. Asawa [1] formulates the admission of new calls in a cellular network as a dynamic programming problem and proves that the optimal admission policy is of threshold

type. With g guard channels, new calls are blocked if the number of free channels is less than g, while handoff calls are accepted whenever there is an idle channel available. Both blocked handoff calls and blocked new calls may be queued, generally with priority given to the handoff calls. Such a scheme gives preferential treatment to the handoff calls by penalizing the new calls. A variation of this scheme admits new calls with a certain positive probability when the number of free channels is less than g.

From the mobility model, we may classify mobile users according to their average travel velocity. We shall consider a simple classification of users into two velocity types: slow-moving users and fast-moving users. We may then devise an admission control policy, based on velocity, to improve, in particular, the handoff blocking probability of the slow-moving users. Consider the following velocity-based admission control policy: Suppose that there are g channels in a given cell. A handoff call (type 1 or 2) is admitted provided there is at least one channel available. A new type 1 call is admitted if and only if the number of available channels exceeds G_1. A new type 2 call is admitted if and only if the number of available channels exceeds G_2, where $0 < g_2 < g_1 < g$. Using Markov decision theory, one can establish the optimality of such a policy with respect to minimizing the expected discounted cost due to rejection of new call requests and handoff calls over the set of admissible policies (cf. [1]).

In the following, we provide an analysis of the velocity-based admission control policy. First, we introduce some basic notation to characterize the system. Let the new call arrival rate to a given cell be denoted by Γ and let the average handoff request rate be denoted by Γ_h. The average rate at which new calls are admitted the cell is given by $\Gamma_a = \Gamma(1 - P_b)$, where P_b is the probability of blocking for new calls. Similarly, the average rate at which handoff calls are admitted is given by $\Gamma_{ha} = \Gamma_h(1 - P_{bh})$, where P_{bh} is the probability of blocking for handoff calls. We introduce several random variables associated with the system:

- T_H : the channel holding time in a cell.

- T_M : the connection holding time.

- T_n : the time period from the origination of a new call to the time it crosses the cell boundary and requires a handoff.

- T_h : the time period from the admission of a handoff call to the time when it requires another handoff.

- T_{Hn} : the channel holding time for a new call in a cell.

- T_{Hh} : the channel holding time for a (successful) handoff call in a cell.

Let P_N be the probability that a new call (which is not blocked) will require at least one handoff before completion and let P_H be the probability that a call which has already been handed off successfully will require another handoff before completion. These probabilities can be expressed as follows:

$$P_N = P\{T_M > T_n\} \text{ and } P_H = P\{T_M > T_h\}. \quad (33)$$

45

Noting that $T_{Hn} = \min(T_M, T_n)$, the channel holding time distribution for new calls can be calculated as follows:

$$F_{T_{Hn}}(t) = F_{T_M}(t) + F_{T_n}(t)(1 - F_{T_M}(t)). \qquad (34)$$

The channel holding time distribution for handoff calls can be computed similarly. We now apply this general model to the type-based admission control policy, with some additional assumptions. For concreteness and ease of exposition, we shall assume a "Manhattan Street" cell pattern which is a rectangular grid, wherein each cell has four neighbors.

We shall assume uniform velocity distributions for each type of mobile user. Type 1 mobiles travel at a velocity V_1, uniformly distributed between 0 and v, while type 2 mobiles travel at a velocity V_2, uniformly distributed between v_a and v_b. We assume further that handoffs occur only between a given cell and its four neighbors and that the mobile maintains a constant speed and direction throughout its holding time. The latter assumption becomes more accurate as the cell size decreases. The overall call request rate to a cell is the sum of the arrival rates for new and handoff calls of both types:

$$\Gamma = \Gamma_{1,n} + \Gamma_{2,n} + \Gamma_{1,h} + \Gamma_{2,h} \qquad (35)$$

where $\Gamma_{j,n}$ and $\Gamma_{j,h}$ denote the type j arrival rates for new calls and handoff calls, respectively. We can express the handoff call arrival rates in terms of the new call arrival rates and the probabilities, P_b, P_{bh}, P_N, and P_H, as follows:

$$P_{j,h} = \Gamma_{j,n}(1 - P_{j,b})/(1 - P_{j,H}(1 - P_{j,bh})), \qquad (36)$$

where the additional subscript j refers to the mobile type. The distribution of the cell residing time for a new type j call can be computed as:

$$F_{T_{j,n}}(t) = \int\int_{s < vt} f_{j,S,V}(s,v)\,ds\,dv, \qquad (37)$$

where $f_{j,S,V}(s,v)$ denotes the joint probability density of position S and velocity V for type j mobiles. With our assumptions on the mobile, this probability takes a simple form and $F_{T_{j,n}}(t)$ can be computed easily. Similarly, the distribution of cell residing time for a handoff type j call, $F_{T_{j,h}}(t)$, can be obtained. The quantities $P_{j,N}$ and $P_{j,H}$ can then be expressed using the relations in Eq. (33).

The system can be formulated as a $G(N)/G/g(0)$ loss station, which has a product-form solution (see [8]). Using computational algorithms discussed in [8], one can obtain the equilibrium state probabilities p_i, where the state i denotes the number of occupied channels in the cell. The handoff call blocking probability (same for both types) can then be obtained as $P_{bh} = p_N$. Finally, the new call blocking probability for type j users can be computed in terms of the state probability as follows:

$$P_{j,b} = \sum_{i=g-g_j}^{g} p_i. \qquad (38)$$

Using the above analysis, the guard channel thresholds, g_j, can be set to achieve the desired tradeoff in call blocking probability for the various call types. In practice, the parameters of the mobility model will change (slowly) with time. Therefore, the admission thresholds should be dynamically
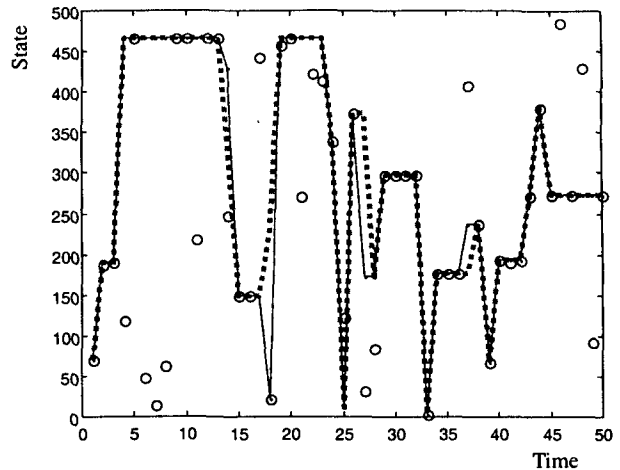

Figure 2: Mobility state tracking example.

adjusted to reflect the current mobility parameters. We are currently investigating this extension of the basic threshold-based admission control scheme discussed above.

6. NUMERICAL EXAMPLE

Figure 2 illustrates the mobility tracking algorithm for a wireless network covering a radius of several hundred meters. The attributes for the abstract mobility state space are location (subarea), direction and speed. The model consists of 500 active states, resulting in a 500×500 transition probability matrix A. We assume that from each active state, a user can transit to on the order of ten states in neighboring subareas. As a result, the matrix A is sparse. We have assumed that the initial state probability distribution is uniform and that the state dwell time distribution is geometric. The state emission probability distribution is given as follows. The probability that the observed state is $m \in S$ given that the true state is m is set to 0.67. The probability that the observed state is $n \in S$, where $n \neq m$ is uniform with a total probability of 0.33.

In Figure 2, the observed state values are shown as open circles. The true state sequence is shown as a dashed line while the estimated state sequence is shown as a solid line. The observation interval T is 50 minutes. In this example, the average observation error is 42%, whereas the average estimation error is 8%. The figure illustrates the ability of the algorithm to track the user's state in spite of the observation errors.

7. CONCLUSION

In this paper, we have introduced a new integrated model of mobility and traffic for the wireless network. The model is very general and can capture user mobility, traffic and requirements. We have developed an algorithm for tracking the parameters of a model based on the new parameter estimation algorithm for the general Hidden Semi-Markov Model reported in [24]. We have applied the model to resource allocation at the wireless/wired network interface and admission control within the wireless network.

We obtained an approximate expression for the computational resource required to process user requests (e.g., Internet web pages) from the wireless network. Using the integrated model, we obtained an approximation for the bandwidth requirement at a base station to support real-time traffic streams from the mobile network. Finally, we demonstrated how the information obtained from tracking the mobile user's travel velocity could be used to improve the blocking performance of handoff calls via a threshold-based admission control scheme.

We are currently gaining computational experience with the dynamic mobility tracking algorithm under various mobility and traffic scenarios. We plan to incorporate traffic measurements to refine estimating the bandwidth required for wireless traffic at the network interface. Finally, we plan to extend the velocity-based admission control described above to adjust the admission thresholds adaptively in accordance with state transition information obtained from the mobility model.

8. REFERENCES

[1] M. Asawa. Optimal admissions in cellular networks with handoffs. In *Proc. IEEE ICC '96*, June 1996.

[2] P. C. Chen. A cellular based mobile location tracking system. In *Proc. IEEE VTC'99*, pages 1979–1983, 1999.

[3] J. D. Ferguson. Variable duration models for speech. In *Symp. on the Application of Hidden Markov Models to Text and Speech*, pages 143–179, Oct. 1980.

[4] D. Gu and S. S. Rappaport. A dynamic location tracking strategy for mobile communication systems. In *Proc. VTC'98*, pages 259–263, 1998.

[5] M. Hellebrandt and R. Mathar. Location tracking of mobiles in cellular radio networks. *IEEE Trans. on Vehicular Technology*, 48(5):1558–1562, Sept. 1999.

[6] 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. Tech.*, 35(3), Aug. 1986.

[7] D. L. Isaacson and R. W. Madsen. *Markov Chains: Theory and Applications*. John Wiley & Sons, Inc., 1976.

[8] H. Kobayashi and B. L. Mark. Product-Form Loss Networks. In J. H. Dshalalow, editor, *Frontiers in Queueing: Models and Applications in Science and Engineering*, pages 147–195. CRC Press, 1997.

[9] U. Leonhardt and J. Magee. Multi-sensor location tracking. In *Proc. 4th ACM/IEEE Int. Conf. on Mobile Computing and Networking*, pages 203–214, 1998.

[10] K. K. Leung, W. A. Massey, and W. Whitt. Traffic models for wireless communication networks. *IEEE J. Select. Areas in Comm.*, 12(8):1353–1364, Oct. 1994.

[11] Y. Lin and P. Lin. Performance modeling of location tracking systems. *Mobile Computing and Communications Review*, 2(3):24–27, 1998.

[12] G. Y. Liu and G. Q. Maguire. A predictive mobility management scheme for supporting wireless mobile computing. *Walkstation Project Technical Report, 1995-02-0. (available online: http://www.it.kth.se/labs/ccs/WS/papers)*, 1995.

[13] H. Maass. Location-aware mobile applications based on directory services. *Mobile Networks and Applications*, (3):157–173, 1998.

[14] B. L. Mark and G. Ramamurthy. Real-time Estimation and Dynamic Renegotiation of UPC Parameters for Arbitrary Traffic Sources in ATM Networks. *IEEE/ACM Trans. on Networking*, 6(6):811–827, 1998.

[15] B. L. Mark and G. Ramamurthy. Real-time Traffic Characterization for Quality-of-Service Control in ATM Networks. *IEICE Trans. on Comm.*, E81-B(5):832–839, July 1998.

[16] S. Nanda. Teletraffic models for urban and suburban microcells: cell sizes and handoff rates. *IEEE Trans. Veh. Tech.*, 42(4), Nov. 1993.

[17] P. V. Orlik and S. S. Rappaport. A model for teletraffic performance and channel holding time characterization in wireless cellular communication. *IEEE J. Select. Areas. in Comm.*, 16(5):788–803, 1998.

[18] Q. Ren and H. Kobayashi. Diffusion process approximations of a statistical multiplexer with markov modulated bursty traffic sources. *IEEE Jour. of Select. Areas in Commun.*, 16(5):679–691, 1998.

[19] C. Rose and R. Yates. Location uncertainty in mobile networks: a theoretical framework. *IEEE Communications Magazine*, 35(2), Feb. 1997.

[20] S. Tekinay. Modeling and analysis of cellular networks with highly mobile heterogeneous sources. In *Ph.D. dissertation*. School of Information Technology and Engineering, George Mason University, 1994.

[21] W. Turin. *Digital Transmission Systems*. McGraw Hill, 1998.

[22] S.-Z. Yu and H. Kobayashi. An optimal prefetch cache scheme. *submitted for publication*, November 1999.

[23] S.-Z. Yu and H. Kobayashi. A prefetch cache scheme for location dependent services. *submitted for publication*, October 1999.

[24] S.-Z. Yu and H. Kobayashi. A Forward-Backward Algorithm for Hidden Semi-Markov Model and its Implementation. *submitted to IEEE Trans. on Signal Processing*, January 2000.