A Distributed Dynamic Regional Location Management Scheme for Mobile IP

Jiang Xie and Ian F. Akyildiz
Broadband and Wireless Networking Laboratory
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332
Email: {jxie, ian}@ece.gatech.edu

***Abstract*— Mobile IP is a simple and scalable global mobility solution. However, it may cause excessive signaling traffic and long signaling delay. Mobile IP regional registration is proposed to reduce the number of location updates to the home network, and reduce the signaling delay. This paper introduces a novel distributed and dynamic regional location management for Mobile IP where the signaling burden is evenly distributed and the regional network boundary is dynamically adjusted according to the up-to-date mobility and traffic load for each terminal. In our distributed system, each user has its own optimized system configuration, which results in the minimal signaling traffic. In order to find the signaling cost function, a new discrete analytical model is developed which captures the mobility and packet arrival pattern of a mobile terminal. This model does not impose any restrictions on the shape and the geographic location of subnets in the Internet. Given the average total location update and packet delivery cost, an iterative algorithm is then used to determine the optimal regional network size. Analytical results show that our distributed dynamic scheme outperforms the IETF Mobile IP regional registration scheme for various scenarios in terms of reducing the overall signaling cost. Through our approach, the system robustness is also enhanced.**

I. Introduction

The growth of the Internet and the success of mobile wireless networks lead to an increasing demand for mobile wireless access to Internet applications. Mobile IP is a mobility-enabling protocol for the global Internet. Standards for Mobile IP have been developed by the Internet Engineering Task Force (IETF) and outlined in Request for Comments (RFCs) 2002-2006 [1] [2].

Mobile IP enables terminals to maintain all ongoing communications while moving from one subnet to another. It is a simple and scalable global mobility solution. However, it is not a good solution for highly mobile users [3]. When a mobile node (MN) moves among subnets, its location must be updated. Mobile IP requires that an MN sends a location update to its home agent (HA) whenever it moves from one subnet to another, even though the MN does not communicate with others while moving. The signaling cost associated with location updates may become very significant as the number of MNs increases [4]. Moreover, if the distance between the visited network and the home network of the MN is large, the signaling delay for the location registration is long.

Mobile IP regional registration aims to reduce the number of signaling messages to the home network, and reduce the signaling delay. The detailed protocol specification can be found in [5] and the general model of operation is illustrated in Fig. 1. Regional registration is a solution for performing registrations locally in a regional network. When an MN moves from one regional network to another one, it performs a home registration with its HA. During the home registration, the HA registers the care-of address of the MN, which is actually a publicly routable address of another mobility agent called gateway foreign agent (GFA). When an MN changes foreign agent (FA) within the same regional network, it does not need to register with its HA. Instead, it performs a regional registration to the GFA to update its FA care-of address. During the communications, when packets are sent to the MN by a correspondent node (CN), they are addressed to the HA of the MN first. The HA intercepts these packets and encapsulates them inside packets that are addressed to the care-of address of the MN. These packets are tunneled through the network until they reach the registered GFA of the MN. The GFA checks its visitor list and forwards the packets to the corresponding FA in the visiting subnet of the MN. The FA further relays the packets to the MN.

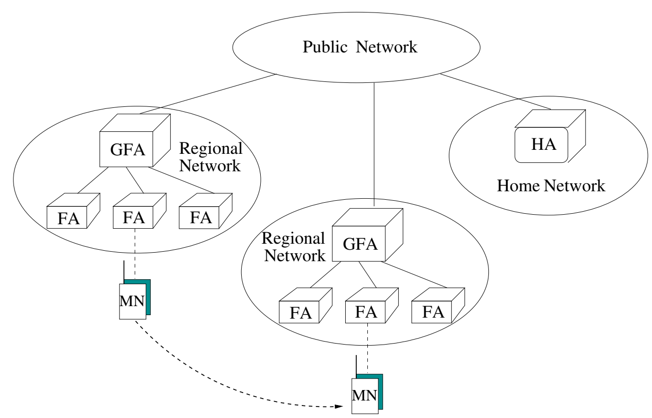

Fig. 1. The IETF Mobile IP regional registration.

However, because of the centralized system architecture, i.e., a centralized GFA manages all the traffic within a regional network, Mobile IP regional registration is more sensitive to the failure of GFAs. The failure of a GFA will prevent packets routed to all the users in the regional network [6]. Another issue that draws our attention is how many FAs should be beneath a GFA within a regional network. The number of FAs under a GFA is very critical for the system performance. A small number of FAs will lead to excessive location updates to the home network and consequently cannot provide the full benefit of regional registration. A large number of FAs will also degrade the overall performance since it will generate a high traffic load on

This work is supported by National Science Foundation (NSF) under Grant No. CCR-99-88532

GFAs, which results in a high cost of packet delivery [4].

To improve the system performance, we propose a distributed GFA management scheme where each FA can function either as an FA or a GFA. Whether an agent should act as an FA or a GFA depends on the user mobility. Thus, the traffic load in a regional network is evenly distributed to each FA. Through this approach, the system robustness is enhanced. We also propose a dynamic scheme which is able to adjust the number of FAs under a GFA for each MN according to the user-variant and time-variant user parameters. In this dynamic system, there is no fixed regional network boundary for each MN. An MN decides when to perform a home location update according to its changing mobility and packet arrival pattern.

In order to minimize the signaling traffic, it is desirable to find the optimal number of FAs beneath a GFA in a regional network. This optimal number is user-variant and time-variant. A method for calculating the optimal location area (LA) size in personal communication systems (PCS) to reach the minimal costs for location update and terminal paging is introduced in [7]. However, there are some differences between the analysis of location management schemes for Mobile IP and those in PCS. First, the cellular network is geographic-oriented. Most researchers adopted structured cell configurations for evaluations [8]. For example, mesh or hexagonal cell configurations are often used in two-dimensional models [9]-[11]. But Internet is more spatial-oriented. We cannot use any geometric shape to accurately abstract a subnet, which increases the difficulty for analysis. Second, in PCS, the geographic distance between two cells is used for analysis [12]. However, the distance between two end points in Internet has nothing to do with the geographic location of these two points. Their distance is usually counted by the number of hops packets travel. This type of distance is called "*virtual*" distance. Third, when an incoming call arrives, the cellular network locates the terminal by simultaneously paging all cells within an LA. Whereas in Mobile IP, HAs and GFAs know the corresponding FA of each MN. But because of the triangular routing, packet delivery introduces extra processing and transmission costs. So there is packet delivery cost instead of paging cost for Mobile IP.

In this paper, we also introduce a new mathematical model to compute the optimal number of FAs under a GFA such that the total signaling traffic for location update and packet delivery consumes the minimal network resource. This model does not impose any restrictions on the shape and the geography of system topology. It is a general model which is applicable for all types of subnets. The distance unit in our model is the number of hops packets travel. Based on this model, we obtain the average location update and packet delivery costs. We use an iterative method to determine the optimal number of FAs under a GFA. We then incorporate this optimal value to our distributed dynamic scheme to further enhance the system performance.

This paper is organized as follows. In Section II, the distributed dynamic regional location management scheme is explained and the protocol for operating the scheme is given. Then, in Section III, the mobility model for deriving the total location update and packet delivery cost is introduced. After that, in Section IV, an algorithm for obtaining the optimal number of FAs beneath a GFA is provided. In Section V, analytical results are presented, followed by the conclusions in Section VI.

II. Distributed And Dynamic Regional Location Management

In this section, we introduce our distributed dynamic regional location management. We also present the operation protocols of our distributed dynamic scheme. In the following discussion, we suppose that the regional registration protocol supports one level of foreign agent hierarchy beneath the GFA.

A. *Overview of Distributed Dynamic Scheme*

We propose a new distributed system architecture. In this system, each FA can function either as an FA or a GFA. Whether an agent should act as an FA or a GFA depends on the user mobility. When an MN enters a regional network, the first FA of the subnet the MN visits will function as a GFA of this regional network. If an agent acts as a GFA, it needs to maintain a visitor list and keep entries in the list updated according to the regional registration requests sent from other FAs within the regional network. The GFA also relays all the home registration requests to the HA. Other agents in the regional network act as the general foreign agents. Of course there should be some authentication setup between mobility agents to guarantee the security of message delivery.

We also propose a dynamic location management mechanism. In this scheme, the number of FAs under a GFA is not fixed but optimized for each MN to minimize the total signaling traffic. The optimal number is obtained based on the incoming packet arrival rate and mobility characteristics of each user. Since the mobility and the packet arrival rate of each user are different and they may also not be constant from time to time, the optimal number of FAs is different for each user and it is adjustable from time to time. Thus, the dynamic system is able to perform optimally for all users.

Therefore, in our distributed and dynamic system, each user has different network configuration with others: different mobility agents act as the GFA for each user and different size of a regional network in terms of the number of FAs. The system architecture of our new scheme is shown in Fig. 2. The advantages of this distributed dynamic system are: the traffic load for all the users in a regional network is distributed to each mobility agent; the system robustness is enhanced since the failure of a GFA will only effect the packets routing to MNs managed by the failing GFA; and each MN has its own optimized system configuration from time to time.

B. *Operations of Distributed Dynamic Scheme*

Now, we describe how an MN determines the dynamically adjusted boundaries of regional networks.

Each MN keeps a buffer for storing IP addresses of mobility agents. An MN records the address of the GFA into its buffer when it enters a new regional network and then performs a home registration through the new GFA. After the home registration, the optimal number of FAs for a regional network is computed based on the up-to-date parameters of the MN. The algorithm for deriving the optimal value k_{opt} will be described

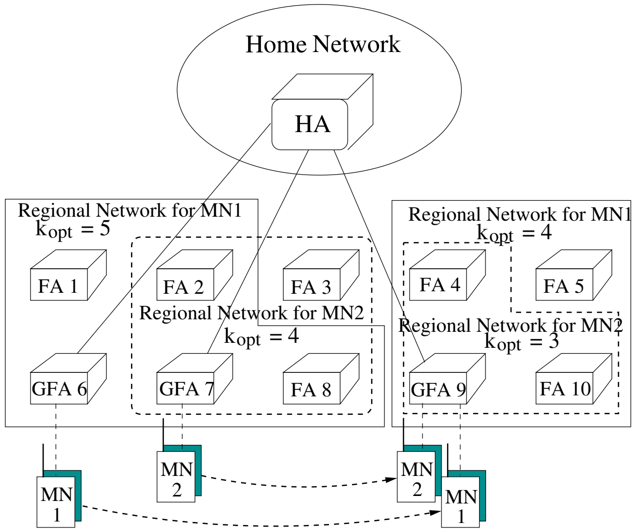

Fig. 2. The distributed dynamic Mobile IP regional registration.

in the next section. This optimal value k_{opt} is set for the buffer length threshold of the MN. If the MN detects that it enters a new subnet, it does a regional registration by sending a regional registration request to the recorded IP address of the GFA, i.e., the first FA it met in the regional network. The MN then compares the IP address of the FA in the new subnet with the addresses recorded in its buffer. If the address of the current FA has not been recorded in the buffer, then the MN records it. Otherwise, ignores it. If the total number of addresses in the buffer as well as the address of the current FA exceeds the threshold, it means the MN is in a new regional network. The MN deletes all the addresses in its buffer, saves the new one, and requests a home registration. Thus, there is no strict regional network boundary for each MN. An MN may move back and forth between two subnets and it may also visit a subnet more than once. The zigzag effect will not lead to excessive home location registrations since the MN knows that it has moved out of a regional network only after it has visited k_{opt} *different* subnets.

The protocol descriptions of the distributed dynamic regional location registration for MNs are shown in Fig. 3.

```
if (MN enters a new subnet)
  compare the address of the new FA to the addresses in buffer;
  if (the new address ≠ any address in buffer)
    if (# of addresses in buffer + the new address > kopt)
      delete all the addresses in buffer;
      record the new FA address in buffer;
      mark the new FA address as the new GFA address;
      perform a regional registration to the new GFA;
      perform a home registration through the new GFA;
      compute the new kopt;
    else
      record the new FA address in buffer;
      perform a regional registration to the GFA;
    end
  else
    perform a regional registration to the GFA;
  end
end
```

Fig. 3. Protocols of the distributed dynamic system for MNs.

C. Comparison

Note that "*distributed* system architecture" and "*dynamic* regional network" are independent. "*distributed*" means that GFAs of different users are distributed among FAs, and "*dynamic*" means changing regional network size k_{opt} from time to time. Consequently, there are four possible combinations:

- *Centralized* system architecture and *fixed* regional network
- *Centralized* system architecture and *dynamic* regional network
- *Distributed* system architecture and *fixed* regional network
- *Distributed* system architecture and *dynamic* regional network

Centralized fixed scheme is the IETF Mobile IP regional registration, which is shown in Fig. 1; centralized dynamic scheme is difficult for implementation, since each FA is required to know the whole network configuration in order to be aware of when to send registration requests to which GFA; distributed fixed scheme is shown in Fig. 4; and distributed dynamic scheme is our proposed scheme, which is shown in Fig. 2. Note that for distributed fixed scheme, the regional network size k_{opt} may be either the same for all users or user-variant. We will compare our distributed dynamic scheme to the centralized fixed scheme and the distributed fixed scheme in the following sections.

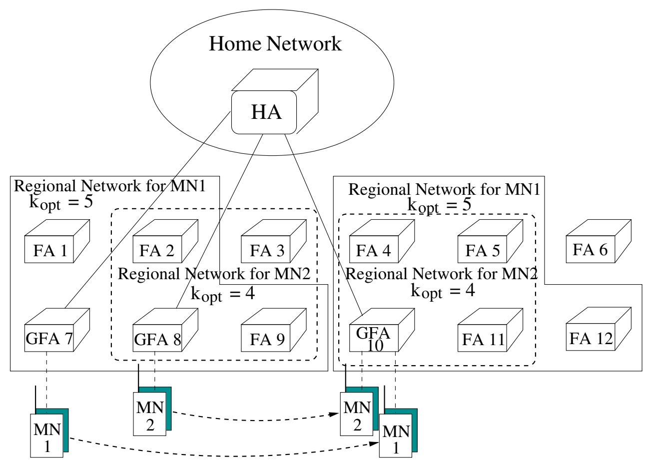

Fig. 4. The distributed fixed Mobile IP regional registration.

III. Signaling Cost Function

In this section, we derive the cost function of location update and packet delivery. Our previous work [13] introduced the cost function of centralized fixed scheme. Here, we analyze the case for distributed dynamic scheme. The total signaling cost in location update and packet delivery is considered as the performance metric. We do not take the periodic binding updates that an MN sends to its HA or FA to refresh their cache into account.

A. Location Update Cost

Similar to [14], we define the following parameters for location update in the rest of this paper:

C_{hg} The transmission cost of location update between the HA and the GFA.

C_{gf} The transmission cost of location update between the GFA and the FA.

C_{fm} The transmission cost of location update over the wireless link between the FA and the MN.

a_h The location update processing cost at the HA.

a_g The location update processing cost at the GFA.

a_f The location update processing cost at the FA.

According to the signaling message flows for location registration in [13], the home registration cost and the regional registration cost for each location update can be calculated as:

$$C_{Uh} = 2a_f + 2a_g + a_h + 2C_{hg} + 2C_{gf} + 2C_{fm} \quad (1)$$
$$C_{Ur} = 2a_f + a_g + 2C_{gf} + 2C_{fm} \quad (2)$$

Let l_{hg} be the average distance between the HA and the GFA in terms of the number of hops, and l_{gf} be the average distance between the GFA and the FA. We assume the transmission cost is proportional to the distance between the source and the destination mobility agents and the proportionality constant is δ_U. Thus C_{hg} and C_{gf} can be expressed as $C_{hg} = l_{hg}\delta_U$ and $C_{gf} = l_{gf}\delta_U$. Since usually the transmission cost of the wireless link is generally higher than that of the wired link, we suppose that the transmission cost over the wireless link is ρ times higher than the unit distance wireline transmission cost. The transmission cost between the FA and the MN can be written as $C_{fm} = \rho\delta_U$. Then the home registration and regional registration costs can be expressed as:

$$C_{Uh} = 2a_f + 2a_g + a_h + 2(l_{hg} + l_{gf} + \rho)\delta_U \quad (3)$$
$$C_{Ur} = 2a_f + a_g + 2(l_{gf} + \rho)\delta_U \quad (4)$$

Note that for distributed GFA architecture, the first FA of the subnet the MN visits acts as a GFA. When the MN resides in the subnet of the GFA, the regional registration cost is different from the one when the MN is in the subnet not serviced by the GFA. Define this special regional registration as $\widetilde{C}_{Ur}$. Then,

$$\widetilde{C}_{Ur} = a_g + 2C_{fm} = a_g + 2\rho\delta_U \quad (5)$$

Assume each MN may move randomly between N subnets and there are k subnets within a regional network. We model the movements of an MN as a discrete system. In the model, an MN may visit a subnet more than once and it may also move back and forth between two subnets. We call the action each MN moving out of a subnet "*a movement*". Define a random variable M so that each MN moves out of a regional network at movement M. At movement 1, an MN may reside in either subnet $1,\ 2,\ \cdots$ or N. At movement 2, the MN may move to any of the other $N-1$ subnets. We assume the MN will move out to the other $N-1$ subnets with equal probability $\frac{1}{N-1}$.

For centralized fixed scheme, the probability of moving out of a regional network, i.e., the probability of performing a home registration at movement m is:

$$P^m_{h\_cf} = \frac{N-k}{N-1} \cdot \left(\frac{k-1}{N-1}\right)^{m-2}, \quad \text{where } 2 \le m < \infty \quad (6)$$

It can be shown that the expectation of M is:

$$E[M]_{cf} = \sum_{m=2}^{\infty} m P^m_{h\_cf} = 1 + \frac{N-1}{N-k} \quad (7)$$

Assume within a regional network, the average time an MN stays in each subnet before making a movement is T_f. Therefore, the average location update cost for centralized fixed scheme is:

$$C_{LU\_cf} = \frac{E[M]_{cf} C_{Ur} + C_{Uh}}{E[M]_{cf} T_f} \quad (8)$$

For distributed GFA system architecture, the MN will move out of a regional network only after it has visited all the k subnets. Previous researchers used either Markovian model [15] or random walk model [10] [11] [16] for probability analysis. However, the movement of MNs for distributed scheme is not a Markov process because the decision of whether an MN can move out of a regional network depends on its mobility history, i.e., whether an MN is in another regional network depends on whether it has visited different k subnets. This increases the difficulty of analysis.

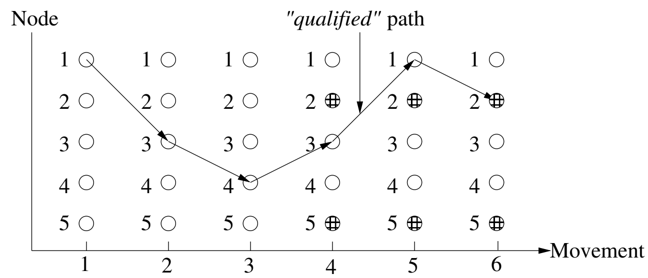

Fig. 5. Discrete system mobility model of an MN.

We define the paths by which the MN has visited different k subnets "*qualified*" paths. If an MN moves out of a regional network at movement m, where m is an arbitrary integer larger than k, the path by which the MN has gone through from movement 1 to movement $m-1$ must consist k *and only* k different nodes. Fig. 5 shows an example of our discrete system in which $N = 5$ and $k = 3$. In the figure, each node represents a subnet. As shown in the figure, at movement 3, the MN has visited subnet 1, 3, and 4. Therefore, subnet 2 and 5 belong to another regional network for this MN after this moment. If the MN moves out of its regional network to subnet 2 at movement 6, the subnets it visited at movement 4 and 5 are among subnet 1, 3, and 4.

Therefore, the expectation of the moment at which an MN moves out of a regional network for distributed scheme is equal to the expectation of the moment at which an MN has visited *different* k subnets plus the expectation of the time period that an MN moves within specific k subnets. The latter one is exactly the $E[M]_{cf}$ for centralized fixed scheme. Define $E[M]_{1\to 2}$ the expectation of the number of movements it takes an MN moving from its first subnet to its second *new* subnet, i.e., an MN has visited 2 different nodes. Then

$$E[M]_{1\to 2} = 1 \quad (9)$$

Similarly, when an MN has visited two different subnets, define $E[M]_{2\to 3}$ the expectation of the number of movements it takes an MN moving to its third *new* subnet. Then

$$E[M]_{2\to 3} = \sum_{n=1}^{\infty} n \cdot \left(\frac{1}{N-1}\right)^{n-1} \frac{N-2}{N-1} = \frac{N-1}{N-2} \quad (10)$$

and the expectation of the number of movements it takes an MN moving from its $(k-1)$th subnet to its kth *new* subnet is:

$$\begin{aligned} E[M]_{k-1\to k} &= \sum_{n=1}^{\infty} n \cdot \left(\frac{k-2}{N-1}\right)^{n-1} \frac{N-k+1}{N-1} \\ &= \frac{N-1}{N-k+1} \end{aligned} \quad (11)$$

Then the expectations of the moment at which an MN moves out of a regional network for distributed fixed scheme and distributed dynamic scheme are:

$$\begin{aligned} &E[M]_{df} = E[M]_{dd} \\ &= E[M]_{1\to 2} + E[M]_{2\to 3} + \cdots + E[M]_{k-1\to k} + E[M]_{cf} \\ &= 1 + \tfrac{N-1}{N-2} + \cdots + \tfrac{N-1}{N-k+1} + \tfrac{N-1}{N-k} + 1 \\ &= 1 + (N-1)\textstyle\sum_{i=1}^{k} \frac{1}{N-i} \end{aligned} \quad (12)$$

Note that the expectation of the moment at which an MN moves out of a regional network for the distributed system is always larger than that for the centralized system. As a result, the number of home registrations per unit time is reduced. The upper bound of the total location update costs per unit time for distributed fixed scheme and distributed dynamic scheme are:

$$C_{LU\_df} \le \frac{\widetilde{C}_{Ur} + (E[M]_{df} - 1)C_{Ur} + C_{Uh}}{E[M]_{df} T_f} \quad (13)$$

$$C_{LU\_dd} \le \frac{\widetilde{C}_{Ur} + (E[M]_{dd} - 1)C_{Ur} + C_{Uh}}{E[M]_{dd} T_f} \quad (14)$$

Based on (4)-(14), we may get the average location update cost. Note that our method does not impose any restrictions on the shape and the geographic location of subnets. It is a general model which is applicable to arbitrary subnets.

B. Packet Delivery Cost

Because of the triangular routing, there are extra costs for packet delivery under Mobile IP regional registration. The packet delivery cost includes the transmission and processing cost to route a tunneled packet from the HA to the registered GFA and further forward to the current serving FA of an MN. Assume

- T_{hg} The transmission cost of packet delivery between the HA and the GFA.
- T_{gf} The transmission cost of packet delivery between the GFA and the FA.
- v_h The packet delivery processing cost at the HA.
- v_g The packet delivery processing cost at the GFA.

The cost for packet delivery procedure can be expressed as:

$$C_{PD} = v_h + v_g + T_{hg} + T_{gf} \quad (15)$$

We assume the transmission cost of delivering data packets is proportional to the distance between the sending and the receiving mobility agents with the proportionality constant δ_D. Then $T_{hg} = l_{hg}\delta_D$ and $T_{gf} = l_{gf}\delta_D$.

The processing cost at GFAs includes decapsulation of the tunneled IP packets from the HA, checking its visitor list to see whether it has an entry for the destination MN, re-encapsulation of the IP packets, and management of routing packets to the FAs. The load on a GFA for processing and routing packets to each FA depends on k, the number of FAs under a GFA. If k is large, the complexity of the visitor list lookup and IP routing lookup in the GFA is high, and the system performance is degraded. In addition, since the total bandwidth of the network is limited, if the traffic to a GFA is heavy, the transmission delay and the number of retransmissions cannot be bounded. These factors will result in a high processing cost at the GFAs. Assume on average there are ω MNs in a subnet. As stated in [13], for centralized system architecture, a GFA serves for all the MNs moving within a regional network, and the total number of MNs in a regional network is ωk on average. Therefore, the complexity of the GFA visitor list lookup is proportional to ωk. On the other hand, for distributed system architecture, different MNs choose different FAs as their GFAs. A GFA only serves the MNs which first enter the subnet managed by this GFA in a regional network. The packet processing load of a GFA in the distributed system is much lower than that in the centralized system because the traffic is allocated evenly among all the FAs in a regional network. Therefore, the complexity of the GFA visitor list lookup for distributed system is proportional only to ω. Since IP routing table lookup is based on the *longest prefix matching* and most implementations use the traditional *Patricia trie* [17], the complexity of IP address lookup is proportional to the logarithm of the length of the routing table k [18]. We define the packet processing cost functions at the GFA for centralized system and distributed system as:

$$v_{g\_cf} = \zeta k \cdot \lambda_a \left(\alpha\omega k + \beta log(k)\right) \quad (16)$$

$$v_{g\_df} = v_{g\_dd} = \zeta k \cdot \lambda_a \left(\alpha\omega + \beta log(k)\right) \quad (17)$$

where λ_a is the packet arrival rate for each MN, α and β are weighting factors of visitor list and routing table lookups, and ζ is a constant which captures the bandwidth allocation cost at the GFA. The larger the ζ is, the more negative effects an MN experiences from not enough network bandwidth available.

The processing cost function at the HA can be defined as: $v_h = \eta\lambda_a$, where η is a packet delivery processing cost constant at the HA. Then the total packet delivery costs per unit time for the three schemes are:

$$C_{PD\_cf} = \eta\lambda_a + \zeta k \cdot \lambda_a \left(\alpha\omega k + \beta log(k)\right) + (l_{hg} + l_{gf})\delta_D \quad (18)$$

$$\begin{aligned} &C_{PD\_df} = C_{PD\_dd} \\ &= \eta\lambda_a + \zeta k \cdot \lambda_a \left(\alpha\omega + \beta log(k)\right) + (l_{hg} + l_{gf})\delta_D \end{aligned} \quad (19)$$

C. Total Signaling Cost

Based on the above analysis, we may get the overall signaling cost function as:

$$C_{TOT\_(\cdot)}(k, \lambda_a, T_f) = C_{LU\_(\cdot)} + C_{PD\_(\cdot)} \quad (20)$$

where $C_{TOT\_(\cdot)}$, $C_{LU\_(\cdot)}$, and $C_{PD\_(\cdot)}$ represent the total signaling cost, location update cost, and packet delivery cost for the three different schemes.

IV. Optimal Value

The optimal number of FAs beneath a GFA, k_{opt}, is defined as the value of k that minimizes the cost function derived in Section III. Because k can only be an integer, the cost function is not a continuous function of k. Therefore, it is not appropriate to take derivatives with respect to k of the cost function to get the minimum. We use an iterative algorithm. Note that iterative algorithm may result in a local minimum. Solutions to solving the local minimum problem were discussed in [12]. Similar to the algorithm proposed in [7], we define the cost difference function between the system with number k and the system with number $k-1 (k \geq 2)$, i.e.,

$$\Delta_{cf}(k, \widetilde{\lambda}_a, \widetilde{T}_f) = C_{TOT\_cf}(k, \widetilde{\lambda}_a, \widetilde{T}_f) - C_{TOT\_cf}(k-1, \widetilde{\lambda}_a, \widetilde{T}_f) \quad (21)$$

$$\Delta_{df}(k, \bar{\lambda}_a, \bar{T}_f) = C_{TOT\_df}(k, \bar{\lambda}_a, \bar{T}_f) - C_{TOT\_df}(k-1, \bar{\lambda}_a, \bar{T}_f) \quad (22)$$

$$\Delta_{dd}(k, \lambda_a, T_f) = C_{TOT\_dd}(k, \lambda_a, T_f) - C_{TOT\_dd}(k-1, \lambda_a, T_f) \quad (23)$$

where $\widetilde{\lambda}_a$ and $\widetilde{T}_f$ are the average packet arrival rate and average subnet residence time for all MNs; $\bar{\lambda}_a$ and $\bar{T}_f$ are the average packet arrival rate and average subnet residence time for each MN. Given $\Delta_{(\cdot)}$, the algorithm to find the optimal value of k is defined as follows:

$$k_{opt\_cf}(\widetilde{\lambda}_a, \widetilde{T}_f) = \begin{cases} 1, \text{ if } \Delta_{cf}(2, \widetilde{\lambda}_a, \widetilde{T}_f) > 0 \\ max\{k : \Delta_{cf}(k, \widetilde{\lambda}_a, \widetilde{T}_f) \leq 0\}, \text{ otherwise.} \end{cases} \quad (24)$$

$$k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f) = \begin{cases} 1, \text{ if } \Delta_{df}(2, \bar{\lambda}_a, \bar{T}_f) > 0 \\ max\{k : \Delta_{df}(k, \bar{\lambda}_a, \bar{T}_f) \leq 0\}, \text{ otherwise.} \end{cases} \quad (25)$$

$$k_{opt\_dd}(\lambda_a, T_f) = \begin{cases} 1, \text{ if } \Delta_{dd}(2, \lambda_a, T_f) > 0 \\ max\{k : \Delta_{dd}(k, \lambda_a, T_f) \leq 0\}, \text{ otherwise.} \end{cases} \quad (26)$$

Note that the optimal value of the centralized fixed scheme is the same for all the MNs and is fixed all the time; the optimal value of the distributed fixed scheme is fixed all the time, but each user may have different optimal value; and the optimal value of our proposed distributed dynamic scheme is adapted to each MN and it depends on the up-to-date packet arrival rate and the mobility.

The algorithm for estimating packet arrival rate can be found in [7]. Each MN may use a timer to count the time it spent in each subnet and the average value within a regional network, T_f, is calculated before computing the k_{opt}. T_f can also be estimated if the probability density function (pdf) of the MN residence time in each subnet within a regional network is known. For example, if the pdf of the MN residence time $f_r(t)$ is of Gamma distribution which has Laplace transform $F_r(s) = \left(\frac{\mu\gamma}{s+\mu\gamma}\right)^{\gamma}$ with mean value $\frac{1}{\mu}$, variance V, and $\gamma = \frac{1}{V\mu^2}$. Then $T_f = \frac{1}{\mu}$. Our algorithm also need to know the number of hops between the HA and the GFA, l_{hg}, and the number of hops between the GFA and the FA, l_{gf}. If each MN has dedicated paths for transmitting signaling messages from FAs to GFAs and HAs, the number of hops between mobility agents (HA, GFA and FA), l_{hg} and l_{gf}, are fixed numbers. If not, signaling packets may take different paths each time according to the traffic load and routing algorithms at each mobility agent. Thus, l_{hg} and l_{gf} vary within a certain range. An MN may use the *time-to-live* (TTL) field in IP packet headers to get the number of hops packets travel [19]. Then the average value may be used for optimal number computation.

V. Analytical Results

In this section, we demonstrate the performance improvement of our distributed dynamic scheme to the centralized fixed scheme, i.e., the IETF Mobile IP regional registration [5]. Since these two schemes are not comparable, first we show the cost saving of distributed fixed scheme to the centralized fixed scheme. Next we demonstrate the advantages of our proposed distributed dynamic scheme over the distributed fixed scheme.

Table I lists some of the parameters used in our performance analysis. Since the total number of subnets that MNs may access through wireless channels is limited, we assume $N = 30$. For our numerical evaluation, we assume l_{hg} and l_{gf} are fixed numbers. Since the TTL field in IP header is usually initialized by the sender to 32 or 64 [19], i.e., the upper limit on the number of hops through which a packet can pass is 32 or 64, we assume $l_{hg} = 25$ and $l_{gf} = 10$.

A. Centralized Fixed Scheme vs. Distributed Fixed Scheme

First, we compare the performance of centralized fixed scheme and distributed fixed scheme. Similar to the analysis in PCS, we define the call-to-mobility ratio (CMR) as the ratio of the packet arrival rate to the mobility rate, i.e., $CMR = \lambda_a T_f$. Since the cost functions of the two schemes derived in Section III are different, we focus on compare the total signaling cost of centralized fixed scheme $C_{TOT\_cf}(k_{opt\_cf}(\widetilde{\lambda}_a, \widetilde{T}_f), \lambda_a, T_f)$ with that of the distributed fixed scheme $C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f), \lambda_a, T_f)$ when the average values of residence time in each subnet and packet arrival rate of all the MNs are the same, i.e., $\widetilde{T}_f = \bar{T}_f$ and $\widetilde{\lambda}_a = \bar{\lambda}_a$.

Fig. 6 plots the optimal k as a function of CMR for centralized fixed scheme and distributed fixed scheme. Note that for the two systems, the optimal regional network size k_{opt} is a designed value. It is computed before the communications based on the average values of user parameters. As shown in the figure, the optimal regional network size decreases as CMR increases for both centralized and distributed systems. When the CMR is low, the mobility rate is high compared to the packet arrival rate and the cost for location update dominates. Systems with larger regional networks may reduce the number of home registrations and provide the benefit of regional registration. When the CMR is high, the packet delivery cost dominates and the saving in packet delivery becomes significant. The saving can be attributed to the smaller regional network size. Note that the optimal regional network size of distributed

TABLE I

PERFORMANCE ANALYSIS PARAMETERS

Pkt Process Cost			Distance Cost Unit		Wireless Multiple	# of MNs/subnet	Weight		Pkt Process Const.	
a_h	a_g	a_f	δ_U	δ_D	ρ	ω	α	β	ζ	η
25.0	15.0	10.0	0.1	0.05	10	15	0.3	0.7	0.01	10.0

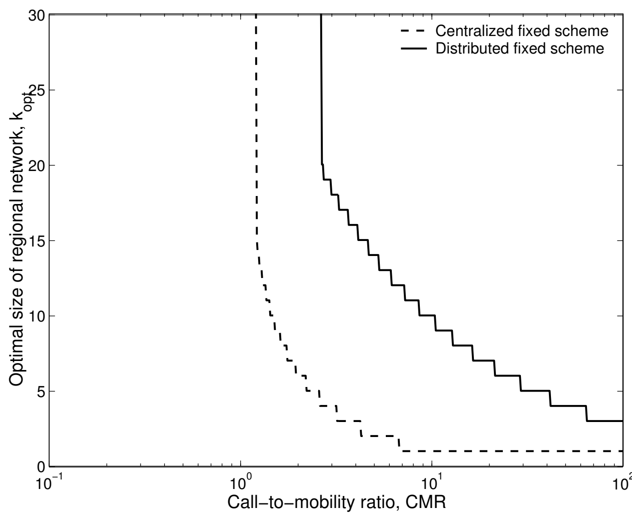

Fig. 6. Optimal regional network size for centralized and distributed systems.

system is always larger than or equal to that of the centralized system. This means for the same CMR, the distributed system has larger regional network size and consequently performs less home registrations compared with the centralized system.

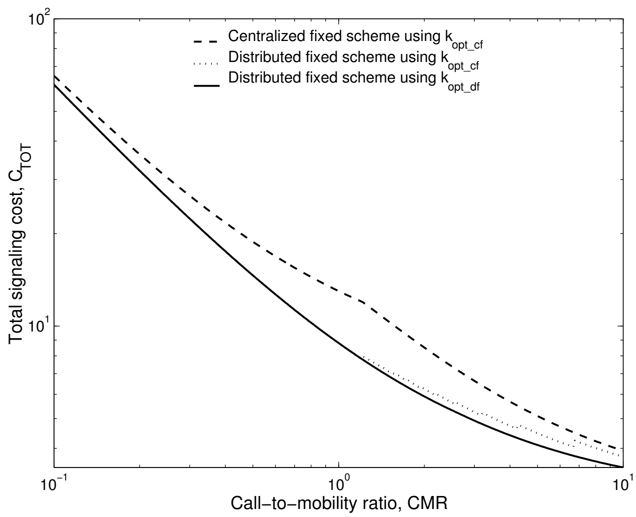

Fig. 7. Comparison of total signaling cost for fixed schemes.

Fig. 7 shows the total signaling cost as a function of CMR for the two schemes. The dashed line in the figure is the signaling cost of centralized fixed scheme when regional network size is k_{opt_cf}. The dotted line is the signaling cost of distributed fixed scheme with k_{opt_cf} as the regional network size. Note that k_{opt_cf} is the optimal value for centralized fixed scheme, in the sense that the minimal cost can be reached. But k_{opt_cf} is not the optimal value for distributed scheme. The solid line in the figure is the signaling cost of distributed fixed scheme under k_{opt_df}. Fig. 7 indicates that even under non-optimal regional network size, the distributed scheme always performs better than the centralized Mobile IP regional registration scheme. And the distributed scheme with optimal regional network size can further improve the performance. Up to 36% signaling cost can be saved when using distributed system architecture.

B. Distributed Fixed Scheme vs. Distributed Dynamic Scheme

Next, we compare the total signaling cost of distributed fixed scheme $C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f), \lambda_a, T_f)$ with that of our proposed distributed dynamic scheme $C_{TOT\_dd}(k_{opt\_dd}(\lambda_a, T_f), \lambda_a, T_f)$ under various scenarios. Note that $k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f)$ is pre-computed before communications. Once it is set, it will not change. But $k_{opt\_dd}(\lambda_a, T_f)$ is dynamically adapted to the user parameters during the communications. Since the cost functions of the two schemes are the same, the advantages of the dynamic scheme over the fixed scheme are reflected when the user parameters are different and changing from time to time. Therefore, we investigate the impacts of user-variant and time-variant user parameters.

1) The Impact of User-Variant Residence Time: We first investigate the impact of user-variant mobility. Let packet arrival rate λ_a be a fixed number, i.e., $\lambda_a = \bar{\lambda}_a =$ constant. Assume there are two groups of MNs [7]. One represents "*active*" users with average residence time in each subnet $\bar{T}_{f_1} = 1.0$. The other group is for "*passive*" users with average residence time in each subnet $\bar{T}_{f_2} = 100$. The residence time of group 1 users follows an exponential distribution, i.e.,

$$f_1(T_f) = \frac{1}{\bar{T}_{f_1}} e^{-T_f/\bar{T}_{f_1}}, \quad T_f \geq 0 \tag{27}$$

and the residence time of group 2 users follows a Gaussian distribution:

$$f_2(T_f) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(T_f - \bar{T}_{f_2}\right)^2/2\sigma^2}, \quad T_f \geq 0 \tag{28}$$

where $\sigma = 10$. Assume that each group has 50% of total users. The residence time T_f of a randomly selected user has pdf as:

$$f(T_f) = 0.5\left(f_1(T_f) + f_2(T_f)\right) \tag{29}$$

and the overall average residence time is:

$$\bar{T}_f = 0.5\bar{T}_{f_1} + 0.5\bar{T}_{f_2} \tag{30}$$

Therefore, the total signaling cost of the distributed fixed scheme is:

$$\begin{aligned} &C_{df} \\ &= 0.5\int_0^\infty f_1(T_f) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_{f_1}), \lambda_a, T_f) dT_f \\ &+ 0.5\int_0^\infty f_2(T_f) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_{f_2}), \lambda_a, T_f) dT_f \end{aligned} \tag{31}$$

where k_{opt} of group 1 users is computed based on their average residence time $\bar{T}_{f_1}$ and k_{opt} of group 2 users is computed based on $\bar{T}_{f_2}$. Note that for distributed fixed scheme, the optimal regional network size may be user-variant or the same for all the users. Fig. 4 gives an example of user-variant k_{opt} and (31) indicates that group 1 and group 2 users adopt different fixed optimal regional network size. The total signaling cost of the distributed fixed scheme using fixed k_{opt} for all the users is:

$$\widetilde{C}_{df} = \int_0^\infty f(T_f) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f), \lambda_a, T_f) dT_f \tag{32}$$

and the total signaling cost of our distributed dynamic scheme is:

$$C_{dd} = \int_0^\infty f(T_f) C_{TOT\_dd}(k_{opt\_dd}(\lambda_a, T_f), \lambda_a, T_f) dT_f \tag{33}$$

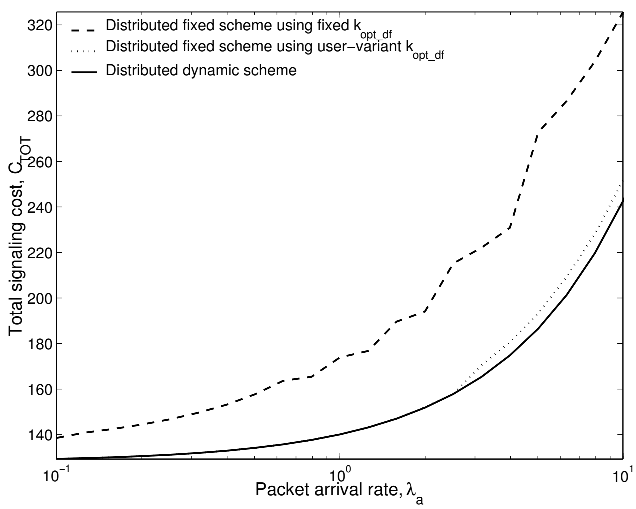

Fig. 8. Comparison of total signaling cost under user-variant residence time.

Fig. 8 shows the total signaling cost of distributed dynamic scheme and distributed fixed scheme under user-variant residence time T_f. The dashed line in the figure is the signaling cost of the distributed fixed scheme using fixed k_{opt_df}, which is actually the case shown in Fig. 7 with solid line. It is observed in Fig. 8 that the signaling cost of distributed dynamic scheme is less than that of both the distributed fixed scheme using fixed optimal regional network size and using user-variant optimal size. Our results demonstrate that C_{TOT} is reduced by up to 33% using dynamic scheme instead of fixed scheme with fixed k_{opt}. Although the performance improvement of the distributed dynamic scheme is not large compared to the distributed fixed scheme under user-variant k_{opt}, in the following time-variant residence time situation, the dynamic scheme will demonstrate its advantage.

2) The Impact of Time-Variant Residence time: Packet arrival rate λ_a is still a constant. The residence time of all MNs, T_f, is of exponential distribution:

$$f(T_f) = \frac{1}{\bar{T}_f} e^{-T_f/\bar{T}_f} \tag{34}$$

where $\bar{T}_f$ is the mean residence time and $\bar{T}_f$ is time-variant. The overall signaling cost of distributed fixed scheme is:

$$C_{df}(\bar{T}_f) = \int_0^\infty f(T_f) C_{TOT\_df}(k_{opt\_df}, \lambda_a, T_f) dT_f \tag{35}$$

Note that although $\bar{T}_f$ is varying during the communications, the optimal value for the fixed scheme k_{opt_df} is pre-computed as a designed value and is a fixed all the time during the communications. The signaling cost of the distributed dynamic scheme is given by (33) using the new pdf function $f(T_f)$ in (34).

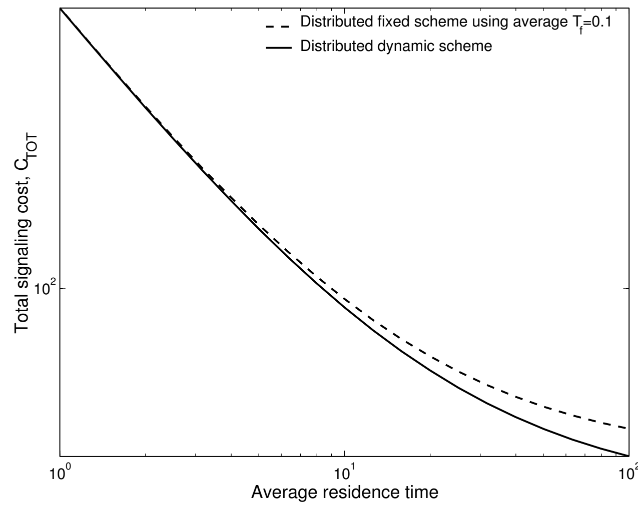

Fig. 9. Comparison of total signaling cost under time-variant residence time.

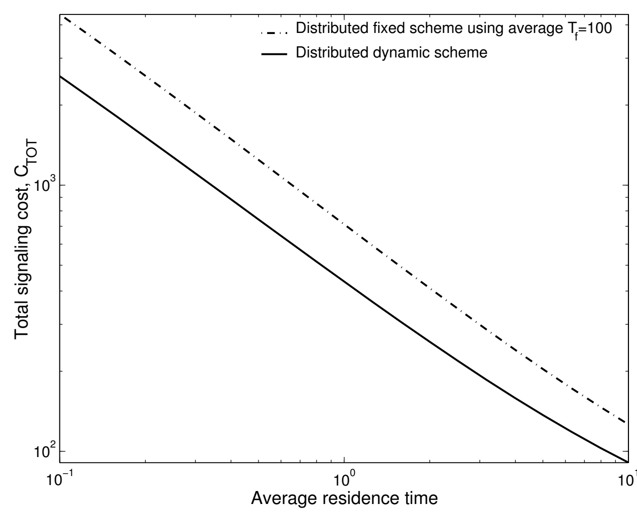

Fig. 10. Comparison of total signaling cost under time-variant residence time.

Fig. 9 and Fig. 10 show the total signaling cost as a function of average residence time $\bar{T}_f$, when $\bar{\lambda}_a = 3.0$. Two cases of the distributed fixed scheme are shown. One is with the optimal regional network size k_{opt_df} pre-computed using $\bar{T}_f = 0.1$ as the average residence time over all users. The other is with the optimal size k_{opt_df} pre-computed using $\bar{T}_f = 100$. Note that the distributed fixed system always pays higher cost than the distributed dynamic system. Our results show that up to 15% cost can be saved by our distributed dynamic scheme compared to the distributed fixed scheme using $\bar{T}_f = 0.1$ for the optimal regional network size computation, and up to 44% cost can be saved compared to the distributed fixed scheme using $\bar{T}_f = 100$ for the computation. We can see from the figures that the distributed fixed system using $\bar{T}_f = 0.1$ for optimal size computation may perform well when the user residence time is small, but when the residence time is large, the fixed scheme consumes more network resource. Similarly, the cost gap of dynamic system and fixed system using $\bar{T}_f = 100$ for computation is smaller when $\bar{T}_f$ is large, but the fixed system pays much more extra bandwidth when $\bar{T}_f$ is small. Therefore, it is a difficult task to design an optimal regional network size beforehand for the distributed fixed scheme. If the user mobility has some unusual big changes to its normal average value, the system with a pre-designed fixed regional network size will consume much more bandwidth and the network may be congested.

3) The Impact of User-Variant Packet Arrival Rate: Now we investigate the impact of user-variant packet arrival rate. Let user residence time T_f be a constant, i.e., $T_f = \bar{T}_f =$ constant. Similar to the discussion in Section V-B.1, we assume there are two groups of MNs. One represents normal users with average packet arrival rate $\bar{\lambda}_{a_1} = 0.1$. The other group is for special users with average packet arrival rate $\bar{\lambda}_{a_2} = 10.0$. The packet arrival rates of group 1 normal users follow an exponential distribution, i.e.,

$$f_1(\lambda_a) = \frac{1}{\bar{\lambda}_{a_1}} e^{-\lambda_a/\bar{\lambda}_{a_1}}, \quad \lambda_a \geq 0 \tag{36}$$

and the packet arrival rates of group 2 special users follow a Gaussian distribution:

$$f_2(\lambda_a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\lambda_a - \bar{\lambda}_{a_2}\right)^2/2\sigma^2}, \quad \lambda_a \geq 0 \tag{37}$$

where $\sigma = 4.0$. Assume that each group contributes 50% of total users. For an arbitrary MN, the packet arrival rate has pdf as:

$$f(\lambda_a) = 0.5\left(f_1(\lambda_a) + f_2(\lambda_a)\right) \tag{38}$$

and the overall average packet arrival rate is:

$$\bar{\lambda}_a = 0.5\bar{\lambda}_{a_1} + 0.5\bar{\lambda}_{a_2} \tag{39}$$

Therefore, the total signaling costs of the distributed fixed scheme using fixed k_{opt} for all the MNs and using different k_{opt} for group 1 and group 2 users are:

$$\widetilde{C}_{df} = \int_0^\infty f(\lambda_a) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f), \lambda_a, T_f) d\lambda_a \tag{40}$$

$$\begin{aligned} &C_{df} \\ &= 0.5\int_0^\infty f_1(\lambda_a) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_{a_1}, \bar{T}_f), \lambda_a, T_f) d\lambda_a \\ &+ 0.5\int_0^\infty f_2(\lambda_a) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_{a_2}, \bar{T}_f), \lambda_a, T_f) d\lambda_a \end{aligned} \tag{41}$$

and the total signaling cost of our distributed dynamic scheme is:

$$C_{dd} = \int_0^\infty f(\lambda_a) C_{TOT\_dd}(k_{opt\_dd}(\lambda_a, T_f), \lambda_a, T_f) d\lambda_a \tag{42}$$

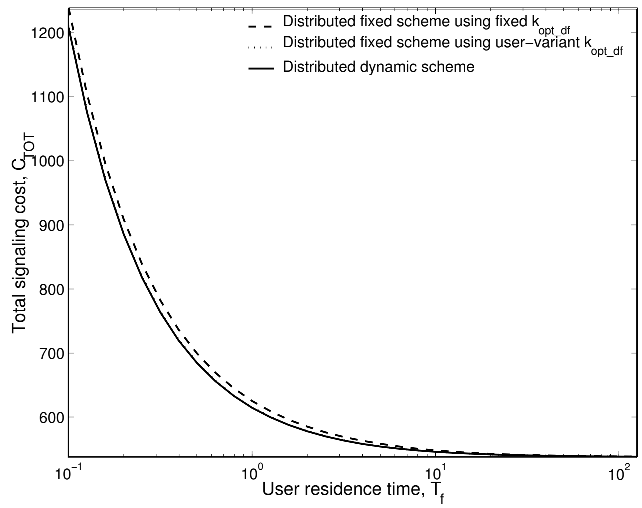

Fig. 11. Comparison of total signaling cost under user-variant packet arrival rate.

Fig. 11 shows the total signaling cost of distributed dynamic scheme and distributed fixed scheme under user-variant packet arrival rate λ_a. The signaling cost of the distributed dynamic scheme is almost the same as that of both the distributed fixed scheme using fixed optimal regional network size and using user-variant optimal size. Only 3% cost can be reduced using the distributed dynamic scheme. It indicates that the optimal regional network size is relatively insensitive to the packet arrival rate. Although different users have widely ranged traffic load, their optimized regional network sizes do not vary much.

4) The Impact of Time-Variant Packet Arrival Rate: Finally, we study the impact of time-variant packet arrival rate. The user residence time T_f is still fixed. The packet arrival rates of all MNs are exponentially distributed:

$$f(\lambda_a) = \frac{1}{\bar{\lambda}_a} e^{-\lambda_a/\bar{\lambda}_a} \tag{43}$$

where $\bar{\lambda}_a$ is the mean arrival rate and $\bar{\lambda}_a$ is time-variant. The overall signaling cost of distributed fixed scheme is given by:

$$C_{df}(\bar{\lambda}_a) = \int_0^\infty f(\lambda_a) C_{TOT\_df}(k_{opt\_df}, \lambda_a, T_f) d\lambda_a \tag{44}$$

where k_{opt_df} is pre-computed and is fixed all the time. The signaling cost of the distributed dynamic scheme is given by (42) using $f(\lambda_a)$ in (43).

Fig. 12 and Fig. 13 plot the total signaling cost as a function of time-variant average packet arrival rate $\bar{\lambda}_a$, when $\bar{T}_f = 10$. The dashed line in Fig. 12 is based on k_{opt} calculated using $\bar{\lambda}_a = 0.01$. The dash-dot line in Fig. 13 is based on k_{opt} calculated using $\bar{\lambda}_a = 100$. The solid line in both figures is for our proposed distributed dynamic scheme where k_{opt} varies according to the up-to-date parameters. The figures show that the fixed system always pays higher cost than the dynamic system. The cost gap is larger when $\bar{\lambda}_a < 0.1$ in Fig. 13 and when $\bar{\lambda}_a > 10$ in Fig. 12. The dynamic system saves up to 19% and 36% cost compared to the fixed system using $\bar{\lambda}_a = 0.01$ and $\bar{\lambda}_a = 100$ for optimal value computation, respectively. This result is similar to that in Section V-B.2. It indicates that our dis-

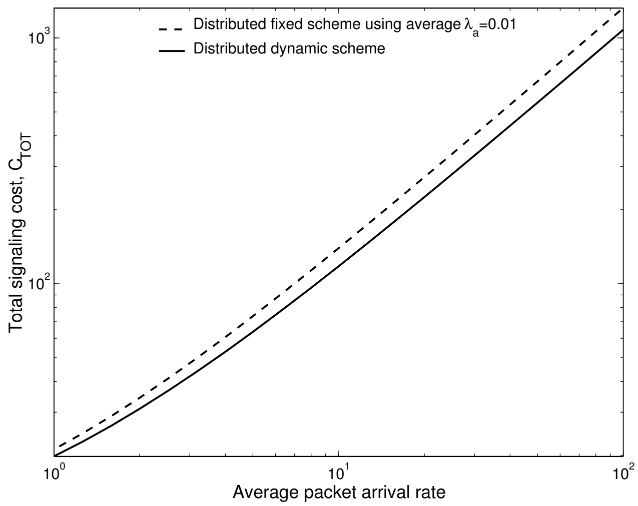

Fig. 12. Comparison of total signaling cost under time-variant packet arrival rate.

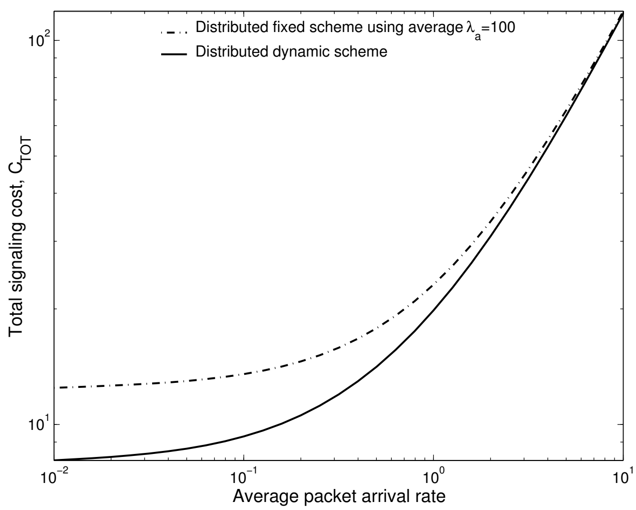

Fig. 13. Comparison of total signaling cost under time-variant packet arrival rate.

tributed dynamic scheme is more cost-efficient when the user parameters are time-variant.

VI. CONCLUSION

In this paper, we introduced a distributed and dynamic regional location mechanism for Mobile IP. We proposed a distributed GFA system architecture where each FA can function either as an FA or a GFA. This distributed system may allocate signaling burden more evenly. A dynamic scheme is adopted by the distributed system to dynamically optimize the regional network size of each MN according to its current traffic load and mobility. We also presented the operation protocols of our distributed dynamic scheme for MNs. The proposed distributed and dynamic scheme is able to perform optimally for all users from time to time and the system robustness is enhanced. Since the movement of MNs does not follow a Markov process, we introduced a novel discrete analytical model for cost analysis and an iterative algorithm to find out the optimal number of FAs in a regional network which consumes the minimal network resource. Our model does not have constraints on the shape and the geographic location of Internet subnets. Analytical results demonstrated that the signaling bandwidth is significantly reduced through our proposed distributed system architecture compared with the IETF Mobile IP regional registration scheme. It is also demonstrated that our dynamic scheme has great advantages under time-variant user parameters when it is not obvious to pre-determine the optimal regional network size.

REFERENCES

[1] C. E. Perkins, “IP mobility support,” Request for Comments (RFC) 2002-2006.

[2] C. E. Perkins, “Mobile IP”, *IEEE Communication Magazine*, pp. 84-99, May 1997.

[3] R. Caceres and V. N. Padmanabhan, “Fast and scalable handoffs for wireless internetworks,” in *Proc. ACM Mobicom’96*, pp. 56-66, 1996.

[4] C. Castelluccia, “Extending Mobile IP with adaptive individual paging: a performance analysis”, in *Proc. IEEE Symposium on Computer and Communications*, pp. 113-118, 2000.

[5] E. Gustafsson, A. Jonsson, and C. E. Perkins, “Mobile IP regional registration (work in progress),” Internet Draft, Internet Engineering Task Force, draft-ietf-mobileip-reg-tunnel-04.txt, March 2001.

[6] H. Omar, T. Saadawi, and M. Lee, “Supporting reduced location management overhead and fault tolerance in Mobile-IP systems,” in *Proc. IEEE Symposium on Computer and Communications*, pp.347-353, 1999.

[7] H. Xie, S. Tabbane, and D. J. Goodman, “Dynamic location area management and performance analysis,” in *Proc. 43rd IEEE Vehicular Technology Conference*, pp. 536-539, 1993.

[8] W. S. Wong and C. M. Leung, “An adaptive distance-based location update algorithm for next-generation PCS networks,” *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 19, no. 10, pp. 1942-1952, October 2001.

[9] W. Wang and I. F. Akyildiz, “Intersystem location update and paging schemes for multitier wireless networks,” in *Proc. ACM MobiCom’2000*, pp.99-109, 2000.

[10] Y-B Lin, W-R Lai, and R-J Chen, “Performance analysis for dual band PCS networks,” *IEEE Transactions on Computers*, vol. 49, no. 2, pp. 148-159, February 2000.

[11] I. F. Akyildiz, Y-B Lin, W-R Lai, and R-J Chen, “A new random walk model for PCS networks,” *IEEE Journal on Selected Areas in Communications (JSAC): Wireless Series*, vol. 18, no. 7, pp. 1254-1261, July 2000.

[12] J. S. M. Ho and I. F. Akyildiz, “Mobile user location update and paging under delay constraints,” *ACM-Baltzer Journal of Wireless Networks (WINET)*, vol. 1, no. 4, pp.413-425, December 1995.

[13] J. Xie and I. F. Akyildiz, “An optimal location management scheme for minimizing signaling cost in Mobile IP,” in *Proc. IEEE International Conference on Communications (ICC 2002)*, April 2002.

[14] Y. Wang, W. Chen, and J. S. Ho, “Performance analysis of Mobile IP extended with routing agents,” *Technical Report 97-CSE-13*, Southern Methodist University, 1997.

[15] A. Bar-Noy, I. Kessler, and M. Sidi, “Mobile users: to update or not to update?,” *ACM-Baltzer Journal of Wireless Networks (WINET)*, vol. 1, no. 2, pp.175-185, July 1995.

[16] I. F. Akyildiz, J. S. M. Ho, Y-B Lin, “Movement-based location update and selective paging for PCS networks,” *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, pp.629-638, August 1996.

[17] B. Lampson, V. Srinivasan, and G. Varghese, “IP lookups using multiway and multicolumn search,” *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp.324-334, June 1999.

[18] H-Y Tzeng, and T. Przygienda, “On fast address-lookup algorithms,” *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 17, no. 6, pp.1067-1082, June 1999.

[19] W. R. Stevens, *TCP/IP Illustrated, Volume 1: The Protocols*, Addison Wesley Longman, Inc., 1994.