

Proxy-based Regional Registration for Integrated Mobility and Service Management in Mobile IP Systems

ING-RAY CHEN*, WEIPING HE AND BAOSHAN GU

Department of Computer Science, Virginia Tech, VA, USA

**Corresponding author: irchen@vt.edu*

We propose and analyse a proxy-based regional registration scheme for integrated mobility and service management with the goal to minimize the network signaling and packet delivery cost in Mobile Internet Protocol (MIP) systems. Under the proposed proxy-based regional registration scheme, a client-side proxy is created on a per-user basis to serve as a gateway between a mobile node (MN) and all services engaged by the MN. From the perspective of these services, the proxy behaves as if it were the MN. From the perspective of the MN, the proxy behaves as if it were the services. Leveraging MIP with route optimization, the proxy runs on a foreign agent node and cooperates with the home agent and foreign agents of the MN in the MIP network to maintain the location information of the MN in order to facilitate data delivery by services engaged by the MN. Further the proxy can optimally determine when to move with the MN so as to minimize the network cost associated with the user's mobility and service management. We investigate the notion of 'service areas' for the proxy to perform 'service handoffs' in our scheme. We show that, when given a set of parameters characterizing the operational and workload conditions of an MN, there exists an optimal service area size for the MN such that the network communication cost is minimized for serving mobility and service management operations of the MN. We demonstrate via Petri net modeling and analysis that our proposed scheme outperforms both basic MIP and MIP regional registration that do not consider integrated mobility and service management.

Keywords: Mobile IP, regional registration, integrated mobility and service management, service handoff, location handoff, proxy, performance analysis, Petri net

Received 9 March 2006; revised 5 December 2006

INTRODUCTION

The next generation wireless network will provide not only voice but also data services. With the success of the Internet, it is widely believed that internet protocol (IP) will become the foundation of next generation wireless networks. With the help of Internet Engineering Task Force standardization, IP-based wireless networks can benefit from the existing and emerging IP-related technologies and services. One key issue is how to provide uninterrupted, reliable and efficient data services to a mobile node (MN) in wireless networks.

The Mobile Internet Protocol (MIP) protocol [1, 2] has been designed to partially address this issue. An MN is identified by its permanent IP home address. If the MN is not in its home area, it has another address named care of address (CoA) associated with its current foreign location. The home agent (HA) maintains a dynamic mapping between the home

address and the CoA of the current foreign agent (FA). When a corresponding node (CN) sends packets to the MN by its home address, the HA intercepts them and tunnels them to the current FA which forwards to the MN. The advantage of MIP is transparency, that is, mobile applications can reach the MN by the same home IP address without having to track the MN's location. However, there is a significant overhead in signaling the HA of CoA changes due to mobility handoffs and in packet delivery due to triangular routing (CN-HA-FA).

The issue of triangular routing has been partially addressed in Mobile IPv4 (MIPv4) with route optimization [3, 4] and has been effectively solved in Mobile IPv6 (MIPv6) [5], by allowing the MN to inform the CN of CoA change, so packets can be delivered directly from the CN to the MN.

The issue of reducing the handoff signaling overhead in MIP systems also has been well researched [6]. A number of

micro–macro (intra–inter-domain) mobility protocols have been proposed in recent years, including Cellular IP [7], Hawaii [8], IDMP [9], Micro Mobile MPLS [10], MIP-RR [11], WIGS [12] and HMIPv6 [13]. While these micro–macro mobility protocols vary in the way routing is performed, the basic idea is to setup regional registration areas or domains (hierarchically in HMIPv6) with a local regional register in each area. When an MN moves across a subnet within a regional registration area, it only informs the local regional register of the location change. The HA and CNs only know the address of the MN’s regional register, which is responsible for maintaining a pointer (or a path as in Hawaii or Cellular IP) to the MN’s current location and forwarding packets to the MN. As an example, Figure 1 illustrates how MIP regional registration (MIP-RR) [11] with route optimization operates. When an MN enters a new regional registration area, it informs the HA and CNs of the address of a gateway foreign agent (GFA) as the Regional CoA (RCoA). When an MN moves within the regional registration area and changes its FA, it only informs its current CoA to the GFA. Data packets sent from a CN to the MN will be intercepted by the GFA and tunneled to the current FA under which the MN resides. The MIP-RR protocol effectively reduces the overhead of location handoffs. However, if a GFA covers too many FAs, then it tends to be away from the current FA so it introduces the cost of triangular routing for data delivery, viz., data packets will take a CN–GFA–FA route instead of a CN–FA route as in MIP with route optimization. These micro–macro mobility management protocols consider the design of routers and protocols to setup regional mobility management, but do not consider the determination of optimal regional area size that would balance and minimize signaling costs due to mobility handoffs vs. packet delivery costs due to triangular routing (CN–regional register–FA).

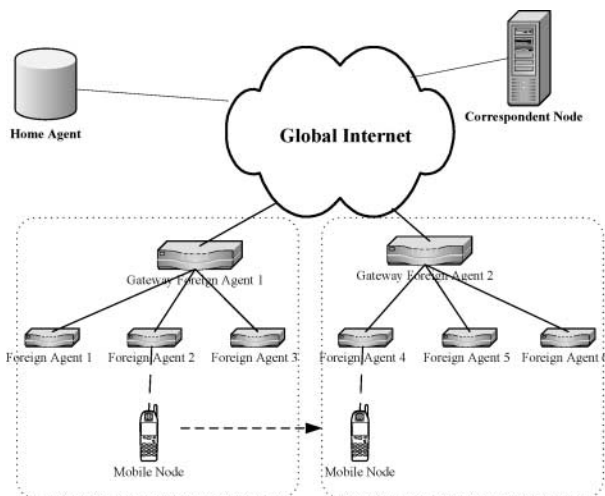


FIGURE 1. Basic infrastructure of MIP regional registration systems.

This article concerns performance optimization of regional registration in MIP systems. Our work is distinct from previous works in three aspects. First, our regional registration scheme deals with both mobility and service management to minimize the overall network cost due to mobility signaling and packet delivery. Second, our regional registration scheme is per-user based. Specifically, regional registers are not shared by all MNs at fixed locations. Rather, regional registers in our scheme are per-user based, taking into consideration of individual MN’s mobility and service characteristics. Lastly, we determine the optimal regional area size that will minimize the overall cost due to mobility handoffs and packet delivery for individual MNs, and to apply the optimal regional area size by individual MNs at runtime. Under our proposed regional registration scheme, a client-side proxy is created on a per-user basis to serve as the regional register of the MN to keep track of the location of the MN within the region, while at the same time acts as the service proxy for application services engaged by the MN. The proxy runs on an FA and will move with the MN when the MN crosses a ‘service area’. The benefits of our proxy-based registration scheme are: (i) the optimal service area size can be dynamically determined to minimize the overall mobility and service management cost based on the MN’s mobility and service characteristics at runtime; (ii) the proxy optionally can carry service context information such as cached data items [14] and Web processing objects [15], and perform context-aware functions such as content adaptation [16] for services engaged by the MN to help application executions; (iii) since there are no ‘designated’ regional registers at fixed locations used for mapping RCoA to CoA, FAs (on which regional registers are run) are less likely to be overloaded even in cases the density of MNs is high in an area.

A major contribution of our work is that we determine the optimal number of FAs in a regional registration area on a per MN basis based on an MN’s mobility and service characteristics to minimize the overall network communication cost, including the signaling cost due to mobility management and the data packet transmission cost due to service management. Lately, Mtika and Takawira [17] have analysed the optimal number of access routers (ARs) in a region in MIPv6 environments. However, they considered the use of regional registers for all MNs and adopted a uniform flow model assuming that all MNs have an average mobility and data packet rates, based on which the optimal number of ARs in a regional area is derived to be applicable to all MNs in the system. Our approach is per-user based, taking each MN’s dynamic mobility and service characteristics into consideration to derive the optimal number of FAs in a regional registration area specifically applicable to individual MNs. Also unlike the uniform flow model used in [17] which simply takes the hop-distance between the MN and the GFA as a constant, we develop a state-based Petri net model to keep track of the actual hop-distance between the MN and the GFA based on the state

the MN is in. This allows the local registration cost to be accurately calculated and thus the optimal number of FAs in a region to be more accurately determined.

The use of proxies for service management has been researched before, but the idea of using proxies for integrated mobility and service management has been considered only lately in the context of wireless cellular networks [18]. Marshall *et al.* [19] proposed Alpine to examine dynamic proxy configuration and placement for managing services. MARCH [16] leverages service proxies for content adaptation in client-server environments for service management so that content transmission matches network capabilities of the mobile device and the user preferences. Service proxies as such potentially can be applied to our scheme which leverages proxies for both mobility and service management.

In this article, we also investigate the notion of ‘service areas’ for the proxy to perform ‘service handoffs’ in our scheme. The size of a service area is defined by the number of FA areas and is determined on a per-user basis depending on the MN’s mobility and service characteristics. A service handoff is the process by which the proxy migrates from one service area to another to stay close to the MN, carrying the service context information (if any) with it during the migration. Since our proxy is responsible for both mobility and service management, a service handoff incurs the cost of informing the HA and CNs of the address change of the proxy, and the cost of service context transfer if the proxy carries service context information. The concept of service handoffs has been brought up by researchers [20–22]. The mobile personal architecture [23] proposes the use of personal proxies to allow people to receive services regardless of the networks, devices or applications they use, while maintaining their privacy. However, it does not deal with mobility management to maintain the connectivity of ongoing applications. Chen *et al.* [18, 24] explored the benefit of proxies for service handoffs in the context of cellular wireless networks. No prior research has been done to integrate mobility management with service management for MIP networks. Endler *et al.* [20] proposed a service delivery protocol using a service proxy to provide reliable message delivery to MNs. However, the proxy moves whenever the MN moves across a location boundary, so it may incur a high communication cost. Joshi [22] adopts server side proxies with content adaptation to interact with a number of clients. The proxy proposed in this article is a per-user, client-side proxy for both mobility and service management for cost minimization in MIP networks.

Our proposed scheme will not constrain the deployment of MIP as emerging programmable IP routers such as those for active networks [25–27] are powerful and flexible to run proxy code. Our proposed proxy will run on network routers (acting as FAs) in the network infrastructure. Some examples for testbed implementations are described in [25–27] in which IP routers are capable of not only passively forwarding packets but also actively executing mobile code and hosting proxies.

Our scheme would facilitate MIP deployment, as it would greatly reduce the overall network cost associated with mobility handoffs and packet delivery for mobile applications in MIP environments.

The rest of the article is organized as follows. Section 2 describes the system model and assumptions. Section 3 develops a performance model based on Petri nets to analyse the network communication cost associated with mobility and service management under the proposed proxy-based regional registration scheme. The objective is to identify the optimal service area to minimize the network communication cost when the mobility and service characteristics of an MN are given. In Section 4, we compare the proxy-based scheme with basic MIP and MIP-RR and provide physical interpretation of the results obtained as well as design suggestions. Finally, Section 5 summarizes the article, suggests a way to use the analysis results obtained at runtime and outlines some future research areas.

SYSTEM MODEL

We assume that when an MN starts in an MIP environment, a *client-side* proxy is created to serve as a GFA as in the MIP-RR protocol [11] to maintain the location information of the MN. Moreover, when the MN invokes a server application (through communicating with a CN), the client-side proxy will communicate with the CN on behalf of the MN as if it were the MN. When the MN crosses an FA boundary thus incurring a location handoff, the proxy acting as a GFA will be informed of the new FA address but may not move with the MN. The proxy will move only when the MN crosses a service area, thus incurring a service handoff. The size of the service area in terms of the number of FA areas covered depends on the mobility and service characteristics of the MN such that the network cost associated with mobility and service handoffs will be minimized. Once a proxy moves into a new service area, it again acts as a GFA for mobility management and a service proxy for all server applications that the MN currently engages.

Figure 2 illustrates the system model where an FA area corresponds to an IP subnet area. The client-side proxy initially runs on the FA node of subnet A. When the MN moves within service area 1 from subnet A to subnet C through B (with two location handoffs), the proxy, acting as a GFA for mobility management, remains at the same location and is informed of the address change by the FA’s. The CN and the HA are not informed of these address changes due to location handoffs in this case. When the MN crosses a service area boundary into service area 2, a service handoff occurs by which the proxy moves into subnet D and runs on the FA node of subnet D. The proxy after the move will stay closer to the MN, so the communication cost for data delivery along the path of CN–proxy–FA–MN is lower. Note that the

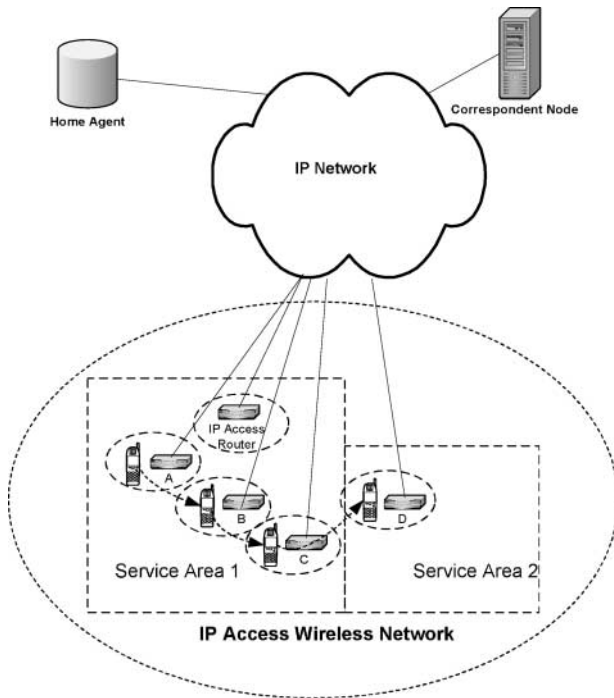


FIGURE 2. Proxy-based regional registration for integrated location and service management.

first FA (i.e. node D) that the proxy moves into is actually at the center of the new service area since starting from that FA, the proxy will move again only when the MN moves across a number of subnets starting from the first FA in the new service area. A proxy move involves the cost of informing the HA and all the CNs of the proxy address change, as well as the context transfer cost. Therefore, there is a tradeoff between these two cost factors. The optimal service area size is dictated by the MN’s mobility and service characteristics. One should note that the service area size of the proxy is not necessarily uniform. Figure 2 shows that service area 1 is larger in size than service area 2 since the mobility and service

characteristics of the MN may be drastically different in different service locations. Figure 2 also shows that a service handoff coincides with a location handoff as the MN moves from subnet C into subnet D.

The proxy is per-user based. Data packets from a server application are sent to the proxy to forward to the MN. From the perspective of a CN, the proxy represents the MN. Conversely, from the perspective of the MN, the proxy represents the server applications (or the CNs) engaged by the MN and all communications between the MN and CNs will go through the proxy. From the perspective of service management, we like to keep the proxy close to the MN since this will reduce the communication overhead for data delivery. This factor favors a small service area. From the perspective of mobility management, we like to keep a large service area to reduce the cost of mobility and service handoffs.

The proxy acting on behalf of the MN to communicate with application servers keeps service context information for services engaged by the MN. Since the proxy is created by the MN, the MN could supply the proxy its mobility and service characteristics the moment it crosses a service area. When the MN moves across an FA boundary, the proxy will check if the service area is crossed. If yes, after the proxy moves into the new service area, a new optimal service area size can be determined by executing a computational procedure developed in this article based on the MN’s mobility and service characteristics in the new service area.

Figure 3 shows the process by which an MN submits a service request. The MN will submit the service request to the proxy through its current FA. The proxy forwards the service request to the CN on behalf of the MN. The responses from the CN are returned to the proxy first and then forwarded to the MN through the current FA. The proxy acting as a GFA knows the location of the current FA and the MN all the time, so data delivery is very efficient in our scheme without incurring the overhead of triangular routing through the HA.

Figure 4 shows the process by which an MN performs a location handoff within a service area during a service

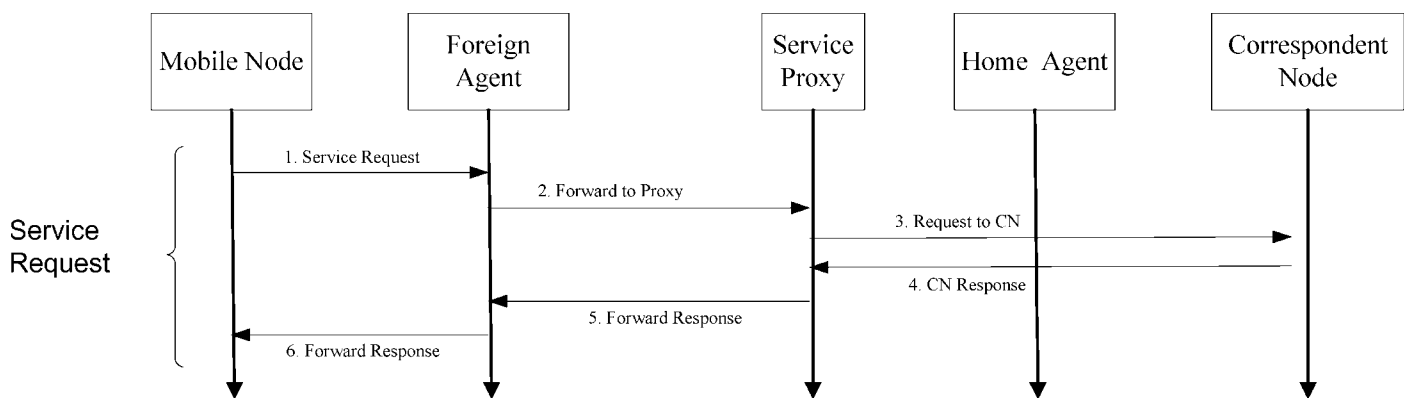


FIGURE 3. Service request process.

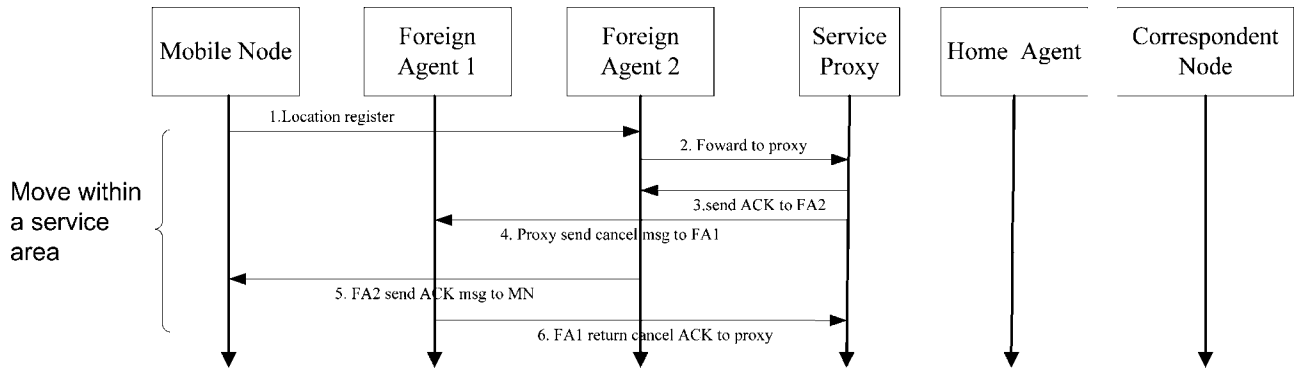


FIGURE 4. Location handoff process within a service area by an MN.

session with a CN. In this case, the proxy is informed of the address change of the FA which the MN moves into but does not migrate itself. The HA and the CN are not involved in the location handoff; they still know the MN by the proxy’s current address which is not changed.

Figure 5 shows the process by which an MN moves across a service area thus incurring a service handoff during an ongoing service session with a CN. The MN first updates its location information with the proxy through the FA it moves into (FA2 shown in the figure). When the proxy realizes that a service area has been crossed, it initiates a migration process to move to FA2. Once the proxy moves into the new service area and runs on FA2 with a new IP address, it informs the HA and the CN of its new address to complete the service handoff.

The implementation of the client-side proxy for an MN can be realized by a middleware running on the MN which initially creates a proxy acting as a GFA to interact with MIP when the

MN starts up. The client-side proxy has two components. One component is at network layer dealing with mobility management in MIP systems by interacting with MIP software running on the network router. The other component is at the application layer dealing with service management with respect to services currently engaged by the MN. These two components cooperate with each other as necessary for a cross-layer implementation of the proxy.

The service and mobility characteristics of an MN are summarized by two parameters. The first parameter is the residence time that the MN stays in a subnet. This parameter can be collected by each MN based on statistical analysis [28]. We expect that future MNs are reasonably powerful for collecting data and doing statistical analysis. The residence time would be characterized by a general distribution in general. We use the MN’s mobility rate (σ) to represent this parameter. The second parameter is the service traffic between the MN and server applications. The MN can also

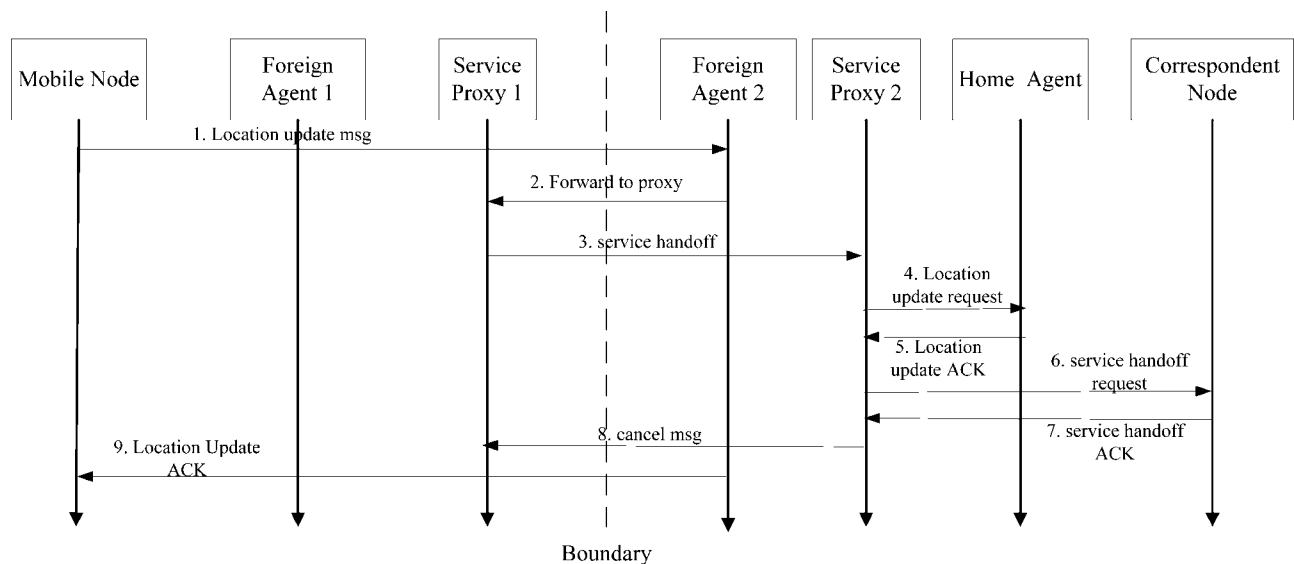


FIGURE 5. Service handoff process when crossing a service area by an MN.

collect data statistically to parameterize this. We use the data packet rate (λ) between the MN and CNs to represent this parameter. Both of these parameters will be periodically determined by the MN. For efficiency, the MN could also build a table to lookup its mobility and service rates as a function of its location, time of the day and day of the week, based on statistical analysis.

When an MN moves across a service boundary, the service proxy moves, thus incurring a service handoff. This overhead involved includes the reconnection cost and the service context transfer cost. The reconnection cost refers to the communication cost for the proxy to inform the HA and the server applications of the new network address. The service context transfer cost is the communication cost to move the service context to the new proxy. We denote the number of packets carrying the context information for context transfer by n_{CT} , the magnitude of which is application-dependent.

In IP-based systems, the communication overhead between two communicating processes is measured by the number of hops, since the number of subnets separating the two processes does not properly measure the distance. In our analysis, we follow the fluid flow model [29] assuming that the average number of hops between two communicating processes separated by k subnets is equal to \sqrt{k} . This assumption is later relaxed by examining a more general function $F(k)$ that returns the number of hops as a function of the number of subnets k .

The system parameters that characterize the mobility and service characteristics of an MN in an MIP system are summarized in Table 1 for easy reference.

PERFORMANCE MODEL

A stochastic Petri net (SPN) model as showed in Figure 6 is developed to analyse the behavior of an MN in an MIP system under our proxy-based regional registration scheme. For ease of referencing, Table 2 gives the meaning of places

and transitions defined in the SPN model. In particular, K represents the number of FAs under a regional registration area for which we want to obtain its optimal value K_{opt} through SPN modeling. The justification of having a different K_{opt} value for each MN in regional registration is that each MN has its specific mobility and service characteristics, thus requiring the use of a different K_{opt} value (through a proxy) to define its optimal regional registration area to minimize the overall cost due to mobility signaling and packet delivery.

We choose SPN as the modeling technique because of its concise representation of the underlying state machine to deal with a large number of states and its ability to reason about an MN's behavior, as it migrates from one state to another in response to events occurring in the system during the MN's lifetime. Moreover, SPN models allow the residence time of an MN in a subnet or the packet inter-arrival time between an MN and its CNs to be generally distributed.

A token in the SPN model represents an FA area crossing (or a location handoff) event by the MN. The function $\text{Mark}(P)$ returns the number of tokens in place P . The number of tokens accumulated in place X_S , i.e. $\text{Mark}(X_S)$, represents the number of FA's crossed (or the number of location handoffs) in a service area. SPN modeling is state-based, with a state being defined by the token distribution into places in the net. Thus, in a particular state we know exactly how many FAs an MN has already crossed by looking at the number of tokens in place X_S . Below, we explain how the SPN model is constructed.

- When an MN moves across an FA area, thus incurring a location handoff, a token is put in place Move . The rate at which location handoffs occur is σ which is the transition rate assigned to Move .
- The MN will register with the new FA. This is modeled by enabling and firing transition MN2FA while disabling transition Move , after which a token flows from place Move to place MFAS , meaning that the MN has just registered with the new FA.

Table 1. Parameters for proxy-based integrated location and service management

Symbol	Meaning
λ	data packet rate, i.e. the data packet rate for all services currently engaged by the MN
σ	mobility rate at which the MN moves across FA boundaries
SMR	service rate to mobility rate ratio, i.e. λ/σ
n_{CT}	number of packets required for content transfer
N	number of server applications currently engaged by the MN
$F(k)$	a general function relating the number of subnets k to the number of hops
K	number of subnets (or FAs) in one service area
τ	1-hop communication delay per packet in wired networks
α	average distance (in hops) between the HA and the proxy
β	average distance (in hops) between a CN and the proxy
γ	ratio between the communication cost in a wireless network to the communication cost in a wired network

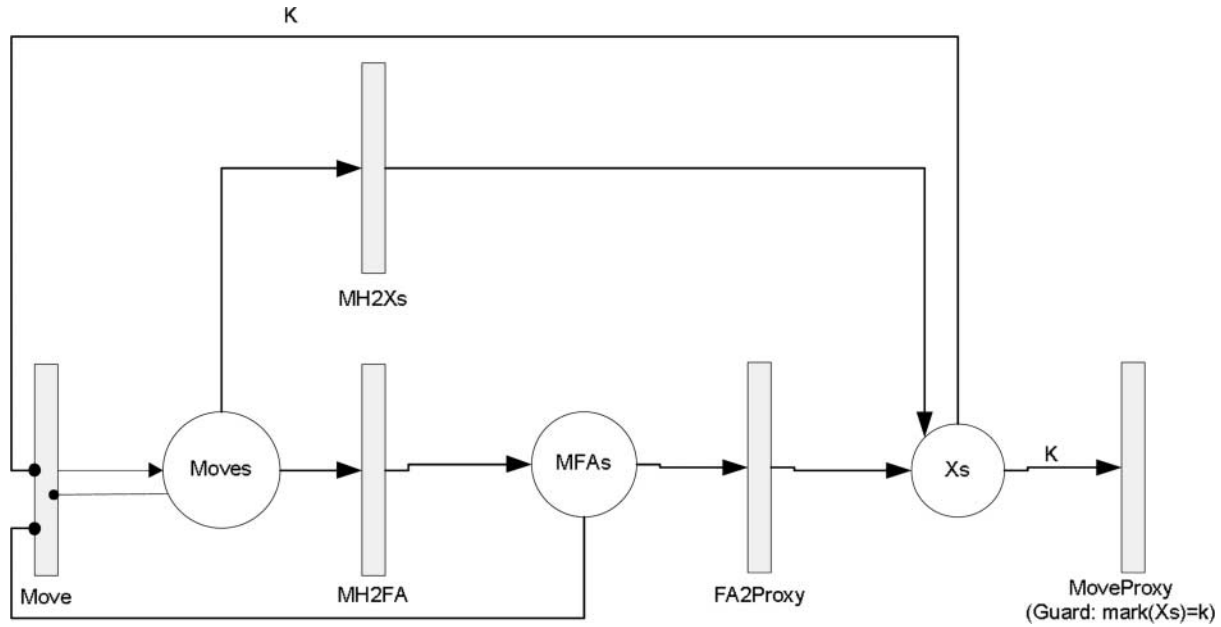


FIGURE 6. SPN model for proxy-based integrated location and service management.

- The MN’s new FA will communicate with the proxy which acts as a GFA. This is modeled by enabling and firing transition FA2Proxy while disabling transition Move. After FA2Proxy is fired, a token in place MFAs flows to place Xs, representing that a location handoff has been completed and the proxy has been informed of the FA address change.
- If the number of tokens in place Xs has accumulated to K, a threshold set by the system to represent the size of a service area, then it means that the MN has just moved into a new service area and a service handoff ensues. This is modeled by assigning an enabling function that will enable transition MoveProxy when there are K tokens in place Xs. After transition MoveProxy is fired, all K tokens are consumed and place Xs contains no token, representing that the proxy has just moved into a new service area.

Parameterization

Here we parameterize (give values to) transition rates of transitions MN2FA, FA2Proxy and MoveProxy based on the set of base parameters defined in Table 1:

- The firing time of transition MN2FA stands for the communication time of the MN registering with the current FA through the wireless network. Thus, the transition rate of transition MN2FA is calculated as:

$$\frac{1}{\gamma\tau}$$

where τ stands for the one-hop communication delay per packet in the wired network and γ is a proportionality constant representing the ratio of the communication

Table 2. Meanings of places and transitions in the SPN model.

Symbol	Meaning
Moves	Mark (Moves) = 1 means that the MN has just moved across an FA area
Move	a timed transition for the MN to move across FA areas
MFAs	Mark (MFAs) = 1 means that the MN has just changed its FA
MN2FA	a timed transition for the MN to register with a new FA
FA2Proxy	a timed transition for the FA to communicate with the proxy
Xs	Mark (Xs) indicates the number of FA’s crossed in a service area
MoveProxy	a timed transition for the proxy to move into a new server area
K	number of FA’s crossed after which a service handoff will occur
Guard:Mark (Xs) = K	enabled if the number of tokens in place Xs is K

delay in the wireless network to the communication delay in the wired network.

- When transition `FA2Proxy` fires, the FA under which the MN resides will inform the proxy of the FA address change. The transition rate of transition `FA2Proxy` depends on how far the MN's current FA is away from the proxy in terms of the number of hops. The number of hops the current FA is away from the proxy depends on the number of FA's crossed by the MN since the last time the proxy moved into the service area. Therefore we calculate the transition rate of transition `FA2Proxy` as:

$$\frac{1}{F(\text{Mark}(\mathbb{X}_S) + 1) \times \tau}$$

where $F(\text{Mark}(\mathbb{X}_S) + 1)$ returns the number of hops between the current FA and the proxy separated by $\text{Mark}(\mathbb{X}_S) + 1$ FAs (or subnets). The argument of the $\text{Mark}()$ function is added by 1 to satisfy the initial condition when $\text{Mark}(\mathbb{X}_S) = 0$ in which the proxy has just moved into a new service area, so at the first FA crossing event, the distance between the proxy and the FA is one FA apart. Note that this transition rate is state-dependent because the number of tokens in place \mathbb{X}_S changes dynamically over time.

- When transition `MoveProxy` fires, the proxy will move into a service area involving a context transfer cost. The proxy also needs to inform the HA and CNs of the address change. The transition rate of transition `MoveProxy` thus can be calculated as:

$$\frac{1}{n_{\text{CT}}F(K)\tau + (\alpha + N\beta)\tau}$$

where $F(K)$ returns the number of hops for two FAs separated by K subnets, n_{CT} is the number of packets required to carry the service context information during a proxy transfer, α the average distance between the proxy and the HA, N the number of server applications (or CNs) which the MN engages concurrently and β is the average distance between the proxy and a CN. The proxy could determine the values of α and β based on statistical data collected on the fly to provide the best estimates of these two parameters.

Cost Model and Measurement

The cost metric considered in this article is the communication cost due to mobility management (i.e. the use of the proxy as a GFA in MIP) and service management (the use of the same proxy for data communication activities with the application servers).

The stochastic model underlying the SPN model is a continuous-time semi-Markov chain¹ with the state representation of (a, b, c) , where a is the number of tokens in place `Moves`, b the number of tokens in place `MFA`s and c the number of tokens in place `Xs`. The stochastic process will reach equilibrium eventually such that there is a non-zero probability that the system will be found in one of the states in a finite set. Let P_i be the steady-state probability that the system is found to contain i tokens in place `Xs` such that $\text{Mark}(\mathbb{X}_S) = i$.

Let $C_{i,\text{service}}$ be the communication cost for the network to service a data packet when the MN is in the i th subnet in the service area. Let C_{service} be the average communication cost to service a data packet weighted by the respective P_i probabilities. This service management cost includes the communication cost between the CN and proxy, the cost between the proxy and the current FA (which depends on the number of hops separating the proxy and the current FA) and the cost for wireless communication between the current FA and the MN. Thus C_{service} is calculated as follows:

$$\begin{aligned} C_{\text{service}} &= \sum_{i=0}^K (P_i \times C_{i,\text{service}}) = \sum_{i=0}^K P_i (\beta\tau + F(i)\tau + \gamma\tau) \\ &= \beta\tau + \gamma\tau + \sum_{i=0}^K (P_i \times F(i)\tau) \end{aligned} \quad (1)$$

Let $C_{i,\text{location}}$ be the communication cost to service a location handoff operation given that the MN is in the i th subnet in the service area. If $i < K$, then the location handoff would only involve the communication cost for the MN to register with the new FA and for the FA to inform the proxy of the address change. On the other hand, if $i = K$, then the location handoff also triggers a service handoff, which in addition to the FA registration cost mentioned above, will also incur a context transfer cost to move the proxy to the new service area and the communication cost for the proxy to inform the HA and the CNs (or application servers) of the address change of the proxy. Let C_{location} be the average communication cost to service a move operation by the MN weighted by the respective P_i probabilities. Then, C_{location} is calculated as follows:

$$\begin{aligned} C_{\text{location}} &= \sum_{i=0}^K (P_i \times C_{i,\text{location}}) = \sum_{i=0}^{K-1} \{P_i \times (\gamma\tau + F(i)\tau)\} \\ &\quad + P_K \times (\gamma\tau + \alpha\tau + N\beta\tau + F(K)\tau n_{\text{CT}}) \end{aligned} \quad (2)$$

¹If all times are exponentially distributed, then the underlying model is a Markov chain. However, since our SPN model allows times to be generally distributed, the underlying model is semi-Markov.

The total cost *per time unit* for the MIP network to service operations associated mobility and service management of the MN is calculated as

$$C_{\text{total}} = C_{\text{service}} \times \lambda + C_{\text{location}} \times \sigma \quad (3)$$

where λ is the data packet rate and σ the mobility rate.

To calculate the total communication cost C_{total} based on Equations (1–3), we need to obtain the steady-state probability that i tokens are found in the place X_s . We utilize Stochastic Petri Net Package [30] to define and evaluate the SPN, and to obtain P_i 's, given a set of parameters characterizing the MN's mobility and service conditions. Specifically, we use the following reward assignment to calculate P_i :

$$r_i = \begin{cases} 1 & \text{if Mark}(X_s) = i \\ 0 & \text{otherwise} \end{cases}$$

RESULTS AND ANALYSIS

Here we present numerical data obtained based on the SPN model developed and Equations (1–3) with physical interpretations given. The numerical analysis is used to (i) show that there exists an optimal service area for MIP systems operating under the proposed proxy-based regional registration scheme for network cost minimization; (ii) compare our proxy-based scheme with basic MIP and MIP-RR and (iii) study the effect of model parameters, including the SMR and n_{CT} , on the optimal service area size.

First, we observe from numerical data that there exists an optimal proxy service area size K_{opt} to minimize the overall communication cost when given a set of parameter values characterizing the mobility and service behaviors of the MN and the network conditions of the MIP network. Figure 7 shows that there exists an optimal size $K_{\text{opt}} = 17$ at $\text{SMR} = 2$, $N = 1$, $\alpha = \beta = 30$, $\gamma = 10$ and $n_{\text{CT}} = 4$, and an optimal size $K_{\text{opt}} = 15$ at $\text{SMR} = 2$, $N = 1$, $\alpha = \beta = 30$, $\gamma = 10$ and $n_{\text{CT}} = 1$ under which the overall cost for the network to service the associated mobility and service management operations with the MN is minimized.

To provide a better sense of the performance improvement of our proposed proxy-based scheme for integrated mobility and service management, we compare our proxy-based scheme with basic MIP without route optimization and MIP-RR with route optimization. Below we first compare our scheme with basic MIP. For the basic MIP scheme, there is no proxy. Thus, for a client-server application, the CN (server application) itself keeps the service context information without the overhead of context transfer. The communication cost $C_{\text{service}}^{\text{MIP}}$ for servicing a packet delivery in basic MIP includes a communication delay from the CN to the HA, a delay from the HA to the current FA and a delay

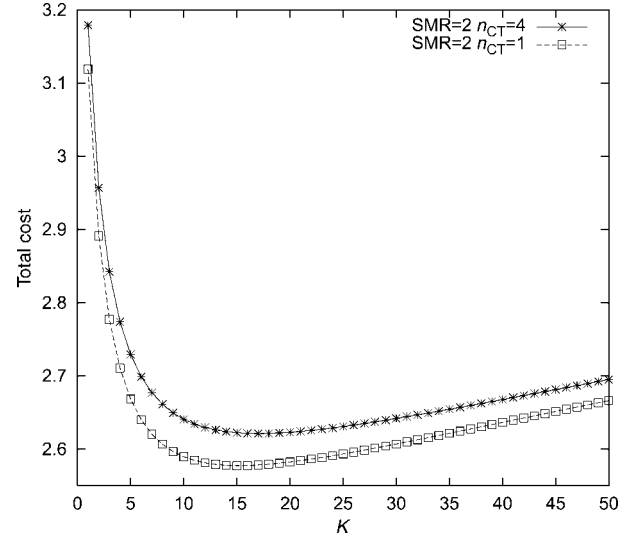


FIGURE 7. Optimal service area size K_{opt} with varying SMR and n_{CT} .

in the wireless link from the FA to the MN as follows:

$$C_{\text{service}}^{\text{MIP}} = \beta\tau + \alpha\tau + \gamma\tau \quad (4)$$

where the average distance between the CN and the HA is assumed to be about the same as that between the CN and the proxy in our proxy-based regional registration scheme.

The cost $C_{\text{location}}^{\text{MIP}}$ for servicing a location handoff under basic MIP consists of a delay in the wireless link from the MN to the FA that it enters into and a delay from the current FA to the HA as follows:

$$C_{\text{location}}^{\text{MIP}} = \gamma\tau + \alpha\tau \quad (5)$$

The total cost assuming $N = 1$ is

$$C_{\text{total}}^{\text{MIP}} = C_{\text{service}}^{\text{MIP}} \times \lambda + C_{\text{location}}^{\text{MIP}} \times \sigma \quad (6)$$

Figure 8 compares the cost of our proxy-based scheme with basic MIP under various SMR. All other parameters are fixed ($N = 1$, $\alpha = \beta = 30$, $\gamma = 10$ and $n_{\text{CT}} = 2$) to isolate the effect of SMR. Both the residence time and packet interarrival time are assumed to be exponentially distributed with varying rates of σ and λ respectively. We observe that our proxy-based regional registration scheme incurs much less communication overhead, the effect of which is especially pronounced when SMR is high. Another observation is that the total cost increases with the increase of SMR for both schemes. We also observe that basic MIP performs comparably well under very low SMR values. The reason is that when SMR is low, the data packet rate is low compared with the user mobility rate. Thus, the overhead of basic MIP due to triangular

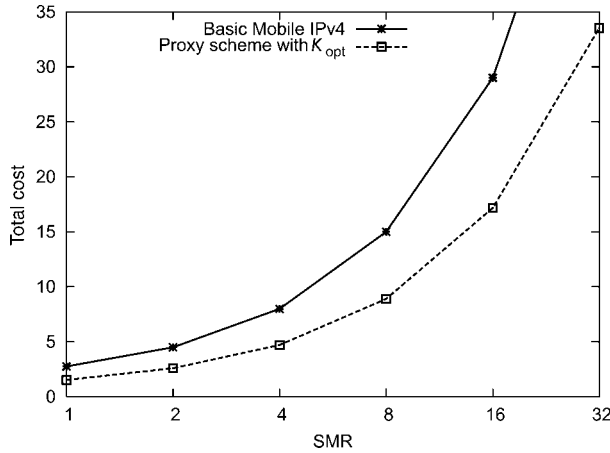


FIGURE 8. Comparison of the proxy-based scheme with basic MIP.

routing (CN-HA-FA) for data packet delivery is negligible. Under a low SMR, our proxy-based scheme would favor a large service area (to be shown later) under which a service handoff incurs a high cost. Thus, these two factors make our proxy-based regional registration schemes' performance comparable to basic MIP under low SMR values. When SMR increases, the cost due to service management also increases. In basic MIP, the service management cost quickly dominates the mobility management cost because of the high overhead associated with the triangular routing for data packet delivery. In our proxy-based regional registration scheme, the balance between mobility and service management is obtained by moving the proxy close to the MN more frequently (thus favoring a smaller service area) to avoid the costly triangular routing (which in our case is CN-proxy-FA) for data packet delivery. Consequently, when SMR is reasonably high, our scheme grossly outperforms basic MIP.

Our proxy-based regional registration scheme can be viewed as an extension of MIP-RR with route optimization, except that for each individual MN, we determine the optimal placement of its *personal* GFA through proxy migration to minimize the network cost. For MIP-RR, the placement of GFAs is pre-determined at fixed locations being applied to all MNs, and each GFA covers a fixed number of subnets, say, K_H . Thus, when an MN crosses a subnet within a GFA area of K_H subnets, the MN only informs its CoA change to the GFA. When the MN crosses a GFA domain of K_H subnets, the address binding of the new GFA's CoA is also sent to the HA and CNs through route optimization in MIP-RR to minimize the signaling cost for data delivery.

Figure 9 compares the total network signaling cost of our proxy-based scheme with MIP-RR with $K_H = 4$ under different combinations of λ and σ in the same setting. The cost is normalized with respect to the per-hop communication cost, $\tau = 1$. We first observe that the total network signaling cost increases with the increase of either the mobility rate or the data packet arrival rate. We next observe that our proxy

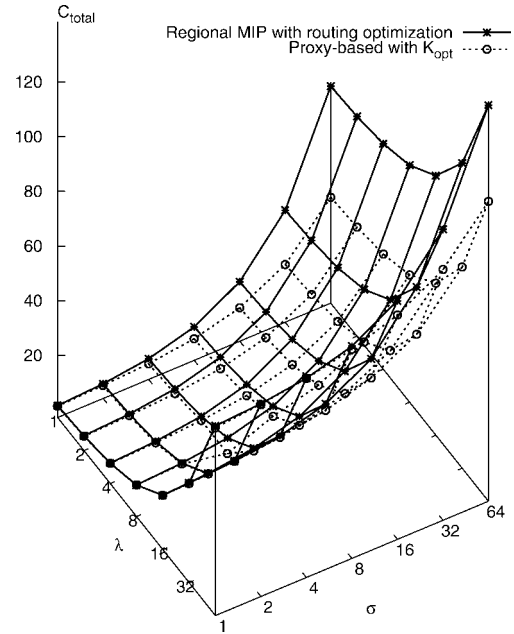


FIGURE 9. Comparison of the proxy-based scheme with MIP-RR with route optimization.

scheme always incurs less communication overhead than MIP-RR, the effect of which is especially pronounced when the MN's mobility rate is high and/or when the data packet arrival rate is high. It is noteworthy that a small cost difference is considered significant because the metric is expressed in terms of the network signaling cost per time unit (normalized to the per-hop cost $\tau = 1$), so that the cumulative effect would be significant over the lifetime of a single MN. It is also noteworthy that the cumulative effect of cost saving for a large number of MNs would be even more significant.

Below we analyse the effect of several model parameters on the optimal service area size, K_{opt} . Figure 10 shows the effect of n_{CT} on the optimal proxy area size K_{opt} with SMR fixed at 2 for the case $N = 1$. Other cases exhibiting similar trends are not shown here. Figure 10 shows that the optimal size K_{opt} initially increases as n_{CT} increases. The reason is that as n_{CT} increases, the context transfer cost becomes high upon a service handoff. Thus, the proxy likes to stay in a large service area to avoid the high cost associated with a service handoff. However, as n_{CT} continues to increase past a threshold value, K_{opt} decreases and eventually $K_{opt} = 1$. This is because the context transfer cost is proportional to $n_{CT} F(K)$, so a large n_{CT} makes the context transfer cost very large by a factor of $F(K)$. Thus, if we allow $K > 1$ during a proxy move operation, the cost of context transfer would dominate the cost of informing the CN and the HA of the address change and the effect of cost saving for informing the CN and the HA of the address change only after a few FA crossing events have occurred would not be significant. Consequently, allowing $K > 1$ does not gain much cost saving as far as the cost of proxy move is

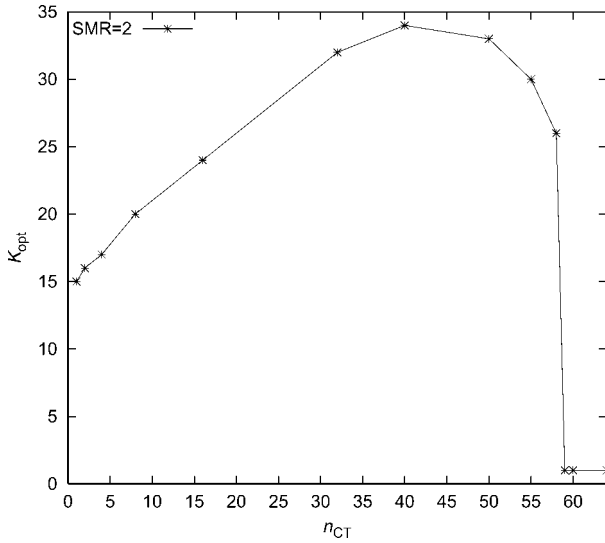


FIGURE 10. Optimal service area size K_{opt} as a function of n_{CT} .

concerned when n_{CT} is very large. This factor coupled with the factor that when $K > 1$, the cost of informing the proxy of the current FA and the cost of delivering packets both would increase by a factor of $F(K)$, favors a small K_{opt} . As shown in Figure 10 when n_{CT} is sufficiently large the system is better-off with $K_{opt} = 1$.

Figure 11 shows the effect of SMR on the optimal service area size K_{opt} with n_{CT} fixed at 2 and $N = 1$. Figure 11 shows that the optimal service area size K_{opt} decreases as SMR increases. The reason is that when SMR is small, the mobility rate is high compared to the data packet rate; thus, the mobility management cost is much larger than the service management cost. The proxy likes to stay at a large service area to reduce the location handoff cost such that a location handoff will

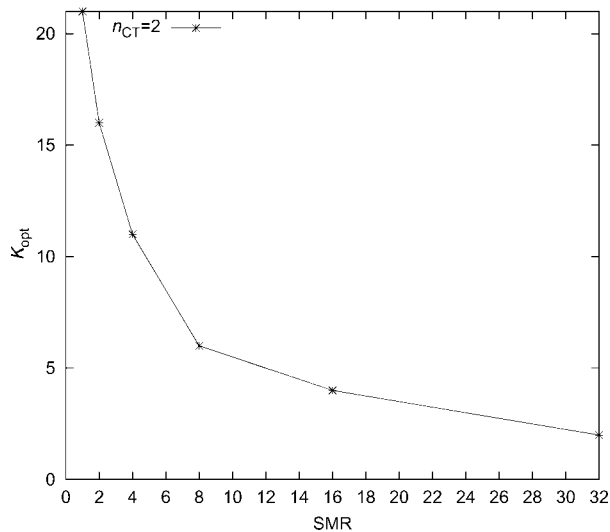


FIGURE 11. Optimal service area size K_{opt} as a function of SMR.

most likely only involve informing the proxy of the location change without incurring a service handoff to migrate the proxy. However, as the data packet rate increases compared with the mobility rate (i.e. as SMR increases), a small service area will reduce the packet delivery cost because the proxy will tend to stay closer to the MN, thus reducing the overhead of triangular routing (CN-proxy-FA) for data packet delivery. As a result, when SMR increases, K_{opt} decreases.

All the above results are obtained based on the assumption that the average number of hops between two FA's separated by k subnets is given by $F(k) = \sqrt{k}$ adopted from the fluid flow model [29]. Below, we test the sensitivity of the results with respect to the form of $F(k)$. Figure 12 shows the total cost obtained under the optimal service area size K_{opt} for $F(k) = \sqrt{k}$, $F(k) = k$ and $F(k) = k^2$. The trends exhibited by these three forms are very similar and are not sensitive to the form of $F(k)$. Thus, all the conclusions drawn earlier from the case $F(k) = \sqrt{k}$ are valid. Here we observe that, however, as the number of hops increases from \sqrt{k} , k to k^2 for two FA's separated by k subnets, the total cost also slightly increases. Thus, this function $F(k)$ does affect the performance of our proposed proxy-based scheme for integrated mobility and service management. Since the performance metric is a rate parameter (amount of cost incurred per time unit), a small difference is also not negligible.

Figure 13 shows the sensitivity of $F(k)$ on the optimal service area size, K_{opt} . The data show that as the number of hops separating two FA's involved in a service handoff increases, e.g. $F(k) = k^2$, the system would favor a smaller service area because a smaller service area tends to mitigate the negative effect of k^2 so as to reduce the cost of context transfer in a service handoff.

APPLICABILITY AND CONCLUSION

In this article, we have investigated the concept of integrating mobility management with service management in MIP

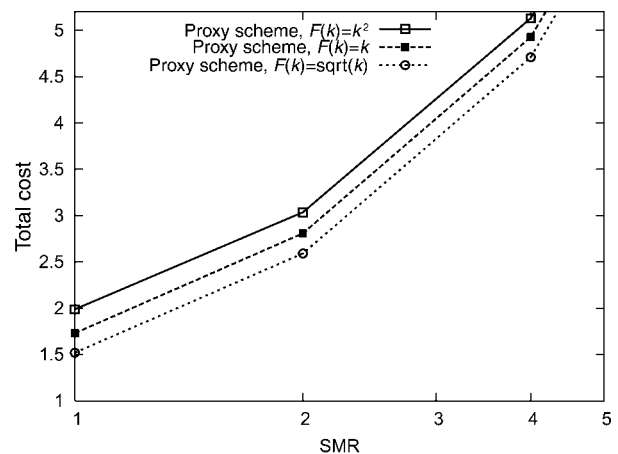


FIGURE 12. Effect of $F(K)$ on communication cost: a comparison.

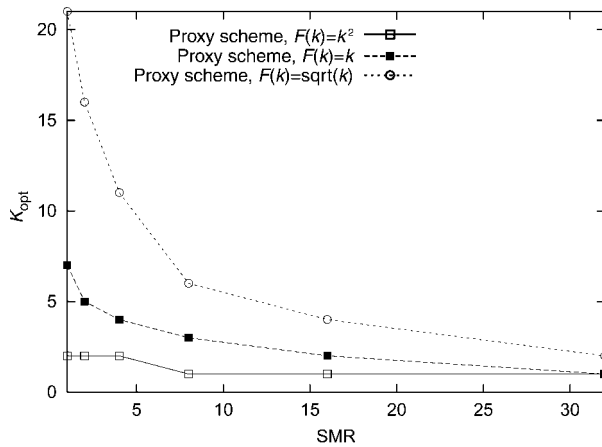


FIGURE 13. Effect of $F(k)$ to K_{opt} .

systems. We proposed using a client-side proxy that serves as a GFA for mobility management and as a service proxy for service management by interacting with all server applications engaged by the MN. We discussed mobility and service characteristics of a MN under the proposed proxy-based regional registration scheme and developed a performance model to identify the optimal service area for performing service handoffs such that the overall communication cost incurred due to mobility and service management operations by the MN is minimized. We showed that when given a set of parameter values characterizing the MN's mobility and service characteristics, as well as the network conditions, there exists an optimal service proxy area under which the network communication cost is minimized. This optimizing proxy area size is determined on a per-user basis. We compared our proxy-based regional registration scheme with basic MIP and MIP-RR and concluded that our scheme consistently outperforms both basic MIP and MIP-RR. The effect is especially pronounced when SMR is high for basic MIP and when either the packet arrival rate or the mobility rate is high for MIP-RR. Our contribution is in the design and evaluation of the proxy-based regional registration scheme which intelligently determines the optimal, dynamic regional registration area for each MN based on the MN's mobility and service characteristics and which enables a proxy to be delegated as the regional register of an MN dynamically to integrate service management with mobility management when needed, and, to be de-allocated when not needed. Our design effectively minimizes the overall signaling and packet delivery cost by all of the MNs in MIP systems by minimizing individual MN's mobility signaling and packet delivery costs.

One way to apply the analysis result reported in the article is to build a table at static time taking a possible range of parameter values. Then at runtime the proxy can perform a simple look-up operation to determine the optimal service area size based on the MN's mobility and service

characteristics supplied to it by the MN in the service area. The presence of multiple MNs can be reflected by adjusting the value of γ to account for the contention of the wireless channel by multiple MNs in the wireless network. This will allow the system to dynamically determine the best service area size even the operating condition (including mobility and service characteristics) of a mobile user in different locations may be drastically different. The performance gain due to the employment of the analysis result is in the amount of communication cost saved per time unit per user so the saving due to a proper selection of the best service area dynamically may have significant impacts since the cumulative effect over all mobile users over time would be significant.

Route optimization is required to apply our proxy scheme. Thus our proposed proxy scheme is beneficial when route optimization in MIP systems becomes widespread such as in MIPv6 systems. In the future, we plan to consider the implementation issue by building a testbed system based on MIP systems supporting route optimization to validate the analytical results obtained in this article. We also plan to consider fault-tolerance and security issues in the design so as to provide reliable, secure and efficient mobility and service management for all future IP-based systems.

REFERENCES

- [1] IETF RFC 3344 (2002) *IP Mobility Support for IPv4*. <http://www.ietf.org/rfc/rfc3344.txt>.
- [2] Wisely, D., Eardley, P. and Burness, L. (2002) *IP for 3G: Networking Technologies for Mobile Communications*. John Wiley & Sons Ltd. (ISBN: 0471486973), Chichester, West Sussex, England.
- [3] HP Release Notes A.02.01 (2003) *HP-UX Mobile IPv4*. Hewlett-Packard, Palo Alto, CA.
- [4] IETF Internet Draft (2001) *Route Optimization in Mobile IP*. <http://www.watersprings.org/pub/id/draft-ietfmobileip-optim-11.txt>.
- [5] IETF RFC 3775 (2004) *Mobility Support in IPv6*. <http://www.ietf.org/rfc/rfc3775.txt>.
- [6] Campbell, A., Gomez, J., Kim, S., Turanyi, Z., Wan, C. and Valko, A. (2002) Comparison of IP micromobility protocols. *IEEE Wirel. Commun.*, **9**, 72–82.
- [7] Campbell, A., Gomez, J., Kim, S., Valko, A., Wan, C. and Turanyi, Z. (2000) Design, implementation, and evaluation of cellular IP. *IEEE Personal Commun. Mag.*, **7**, 709–725.
- [8] Ramjee, R., Varadhan, K., Salgarelli, L., Thuel, S., Wang, S. and Porta, T. L. (2002) HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks. *IEEE/ACM Trans. Netw.*, **10**, 396–410.
- [9] Misra, A., Das, S., McAuley, A., Dutta, A. and Das, S. K. (2002) IDMP-based fast handoffs and paging in IP-based 4g mobile networks. *IEEE Commun.*, **40**, 138–145.

- [10] Langar, R., Tohme, S. and Bouabdallah, N. (2006) Mobility management support and performance analysis for wireless MPLS networks. *Int. J. Netw. Manage.*, **16**, 279–294.
- [11] IETF MIP4 Working Group Internet Draft (2006) *Mobile IPv4 Regional Registration*. <http://www.ietf.org/internet-drafts/draft-ietf-mip4-reg-tunnel-04.txt>, IETF.
- [12] Chan, J.B., Landfeldt, R. L. and Seneviratne, A. (2001) A home-proxy based wireless internet framework in supporting mobility and roaming of real-time services. *IEICE Trans. Commun.*, **E84-B**, 873–884.
- [13] IETF RFC 4140 (2005) *Hierarchical Mobile IPv6 Mobility Management*. <http://www.ietf.org/rfc/rfc4140.txt>.
- [14] Kahol, A., Khurana, S., Gupta, S. and Srimani, P. (2001) A strategy to manage cache consistency in a disconnected distributed environment. *IEEE Trans. Parallel Distrib. Syst.*, **12**, 686–700.
- [15] Hao, W., Fu, J., He, J., Yen, I., Bastani, F. and Chen, I. R. (2006) Extending proxy caching capability: issues and performance. *World Wide Web J.*, **9**, 253–275.
- [16] Ardon, S., Gunningberg, P., Landfeldt, B., Ismailov, Y., Portmann, M. and Seneviratne, A. (2003) MARCH: a distributed content adaptation architecture. *Int. J. Commun. Syst.*, **16**, 97–115.
- [17] Mtika, S. and Takawira, F. (2005) Mobile IPv6 regional mobility management. *ACM 4th Int. Symp. Information and Communication Technologies*, Cape Town, South Africa, January, pp. 93–98. Trinity College Dublin.
- [18] Chen, I.R., Gu, B. and Cheng, S. (2006) On integrated location and service handoff schemes for reducing network cost in personal communication systems. *IEEE Trans. Mobile Comput.*, **5**, 179–192.
- [19] Marshall, I.W., Crowcroft, J., Fry, M., Ghosh, A., Hutchison, D., Parish, D.J., Phillips, I.W., Pryce, N.G., Sloman, M.S. and Waddington, D. (1999) Application-level programmable internetwork environment. *BT Technol. J.*, **17**, 82–94.
- [20] Endler, M., Silva, D. and Okuda, K. (2000) RDP: A result delivery protocol for mobile computing. *ICDCS Workshop on Wireless Networks and Mobile Computing*, Taipei, Taiwan, April, pp. D36–D43. IEEE.
- [21] Jain, R. and Krishnakumar, N. (1994) Network support for personal information services to PCS users. *IEEE Conf. on Networks for Personal Communications*, Long Branch, NJ, USA, March, pp. 1–7. IEEE.
- [22] Joshi, A. (2000) On proxy agents, mobility, and web access. *Mobile Net. and Appl.*, **5**, 233–241.
- [23] Maniatis, P., Roussopoulos, M., Swierk, E., Lai, K., Appenzeller, G., Zhao, X. and Baker, M. (1999) The mobile people architecture. *ACM Mobile Comput. Commun. Rev.*, **3**, 36–42.
- [24] Gu, B. and Chen, I. R. (2005) Performance analysis of location-aware mobile service proxies for reducing network cost in personal communication systems. *ACM Mobile Netw. Appl.*, **10**, 453–463.
- [25] Chan, A., He, D., Chuang, S. and Cao, J. (2001) Towards a programmable mobile IP. *ACM 2nd Int. Conf. on Mobile Data Management*, Hong Kong, January, pp. 210–221. ACM.
- [26] Hussain, S. (2004) An active scheduling paradigm for open adaptive network environments. *Int. J. Commun. Syst.*, **17**, 491–506.
- [27] Raz, D. and Shavitt, Y. (2000) Active networks for efficient distributed network management. *IEEE Commun. Mag.*, **38**, 138–143.
- [28] Chen, I. R. and Verma, N. (2003) Simulation study of a class of autonomous host-centric mobility prediction algorithms for cellular and ad hoc networks. *36th Annual Simulation Symposium*, Orlando, USA, April, pp. 65–72. IEEE.
- [29] Zhang, X., Castellanos, J. and Campbell, A. (2002) P-MIP: paging extensions for Mobile IP. *Mobile Netw. Appl.*, **7**, 127–141.
- [30] SPNP 6.0 (1999) *SPNP Version 6 User Manual*. Department of Electrical Engineering, Duke University, Durham, NC, USA.