Mobile Networks and Applications 10, 453–463, 2005

Performance Analysis of Location-Aware Mobile Service Proxies for Reducing Network Cost in Personal Communication Systems

BAOSHAN GU and ING-RAY CHEN
Department of Computer Science, Virginia Tech, Northern Virginia Center, 7054 Haycock Road, Falls Church, VA 22043

Abstract. We propose and analyze mobile service management schemes based on location-aware proxies with the objective to reduce the network signaling and communication cost in future personal communication systems (PCS). Under these schemes, a mobile user uses personal proxies as intelligent client-side agents to communication with services engaged by the mobile user. A personal proxy cooperates with the underlying location management system so that it is location-aware and can optimally decide when and how often it should move with the roaming user. We show that, when given a set of model parameters characterizing the network and workload conditions, there exists an optimal proxy service area size for service handoffs such that the overall network signaling and communication cost for servicing location and service operations is minimized. We demonstrate via Petri net models that our proposed proxy-based mobile service management schemes outperform non-proxy-based schemes over a wide range of identified conditions. Further, when the mobile user is concurrently engaged in multiple services, the per-service proxy scheme that uses a separate proxy for each service outperforms the aggregate proxy scheme that uses a single proxy to interface with multiple services taking their aggregate service characteristics into consideration.

Keywords: mobile service management, location management, service handoff, service proxy, cost optimization, performance analysis

1. Introduction

Future personal communication systems (PCS) will provide a wide range of personalized mobile services, such as personal banking, personalized stock services, location-aware travel advisory, etc. [7] with most of these mobile services based on the client-server computing paradigm. Future PCS networks will also likely to be IP-based such that the service (or server) will need to know the mobile user's location in order to deliver packets to the mobile user whose location changes from time to time. In general, there are two ways to obtain user location information. The first approach is to query the PCS network location databases. Over the years various location database management schemes [6,8,9] have been proposed to track the location of mobile users in PCS networks. The problem with this approach is the high overhead associated with maintaining the location database when the user moves and querying the location database when the service wants to deliver packets to the user. Another approach is to adopt external means, e.g., using a Global Positioning System (GPS) receiver, to track the location of the mobile user. However, frequently the location information detected is only the location from which the user's request was sent, so extra mechanisms are still required for the server to deliver the response to the mobile user's current location. Moreover, the GPS is not in a massive production scale to support this approach.

The use of a personal mobile service proxy associated with each mobile user has been suggested to track the location of the mobile user in previous studies [1,11,14,15]. The personal proxy will perform tasks such as location tracking, accepting data requests to access mobile services on behalf of the user, converting the request into various application formats, and forwarding the result data packets to the mobile user. The use of a personal mobile service proxy also achieves location privacy. However, since all communications to the mobile user must go through the personal proxy, a "static" proxy may utilize inefficient routes between the service and the mobile user once the user moves, thus incurring a high network signaling and communication cost to the PCS system.

This paper investigates mobile service management schemes based on location-aware "mobile" proxies with the objective to reduce the network cost for client-server personalized services in future PCS networks. Our approach is also based on the notion of personal proxy, that is, the proxy is created on a per-user basis; however, our personal proxy performs location tracking by cooperating with the underlying location management system with the objective to service both "location" and "service" management operations to the mobile user efficiently. To remedy the problem of inefficient routes, we consider the design of moving the personal proxy with the mobile user during location handoffs to minimize the network signaling cost while maintaining the mobile user's required Quality of Service (QoS). How often we move the personal proxy, that is, how often we perform *service handoffs*, depends on the user profile. We investigate the notion of "personal proxy service area." A fast-moving mobile user with low packet rate may require a large proxy area, while a slow–moving user with high packet rate may require a small proxy area.

We also differentiate an *aggregate* proxy from a *per-service* proxy. The former is a per-user proxy that interfaces with all mobile services that the mobile user concurrently engages, while the latter only interfaces with a specific mobile service, that is, a proxy is created for each service accessed by a mobile user. For the case in which the user only engages with one mobile service, the aggregate proxy degenerates into the

per-service proxy. Our per-service proxy performs *service handoff* optimally based on the specific characteristics of the mobile service involved in order to minimize the network signaling cost.

The *service handoff* in the paper refers to the process of moving the personal proxy (aggregate or per-service) as a user moves from one service area into another service area so as to move the proxy closer to the mobile user to reduce the network communication cost. The cost associated with a service handoff includes a reconnection cost to setup a new connection from the new proxy to the server and a context transfer cost to transfer service context from the old proxy to the new proxy. We aim to design and validate location-aware mobile service management schemes based on intelligent proxies that can optimally determine if a service handoff should occur during a location handoff as the user moves such that the overall signaling and communication cost incurred to the network is minimized.

With respect to previous work in proxy-based mobile service management, Pahlavan et al. [14] compared gateway/proxy-based handoff schemes with other schemes in hybrid networks. Endler et al. proposed a service delivery protocol [5] to provide reliable delivery of messages to mobile users. However, these studies considered that a proxy moves whenever the mobile user moves across a cell boundary regardless of specific user and service characteristics. Such indiscriminating service handoffs may incur a large network cost to the PCS system. Dunham and Kumar [3] investigated the impact of mobility on mobile transaction management. They considered various service handoff schemes and examined their costs. Their mobile service management schemes were not integrated with location management. Bellavista et al. [1] introduced a domain-based proxy scheme. The service area of a proxy, however, is a statically pre-defined domain. Joshi and Brewer et al. [2,10] discussed server-side proxies concerned with content transformation and adaptation, each acting like a gateway to interact with a number of clients, while the proxies discussed in our paper are client-side personal proxies concerned with mobility and service characteristics of individual users to reduce the network cost. Jain and Krishnakumar [7] were the first to suggest that location and service handoffs be integrated to reduce the overall cost but no analysis was given. Our work considers integrating service management with location management such that our personal proxies are location-aware and can decide optimal service areas for service handoffs based on mobile user and service characteristics to minimize the overall network signaling and communication cost due to location and service management operations.

The rest of the paper is organized as follows. Section 2 gives a description of the system model and assumptions used. Section 3 describes in detail our location-aware proxy-based schemes for mobile service management. Section 4 analyzes the network signaling cost incurred under our proposed schemes by means of Petri net performance models. Section 5 compares the performance of our proxy-based mobile service management schemes with non-proxy-based ones and gives numerical data; it also reveals conditions (in terms of model parameters such as the proxy area size) under which the overall communication cost incurred is minimized with physical interpretations given. Finally, Section 6 summarizes the paper and outlines some future research areas.

2. System model

In this section, we describe the system model. The PCS system is modeled by a homogeneous hexagonal network coverage model as shown in figure 1 where the PCS service areas are divided into cells. An N-ring area contains $3N^2 - 3N + 1$ cells. For example, when $N = 2$, an area will contain 7 cells and when $N = 3$ it will contain 19 cells. In our proposed scheme, we consider cells being grouped into a proxy N-ring service area with the optimal N to be determined based on the user mobility and service characteristics so as to minimize the network signaling cost. A user proxy for performance reasons may store a cache copy of the user profile.

We assume that a mobile user will stay in a cell before moving to another. For simplicity, the residence time is assumed to be exponentially distributed with an average rate of σ. Such a parameter can be estimated using the approach described in [12] on a per-user basis. A wireless data service is characterized by the service request rate between the client and server (i.e., the incoming packet rate for data) and a service context. The service context records the state of services, e.g., status of a mobile transaction service, which must be moved with the personal proxy during a service handoff to continue services. The inter-arrival time between two consecutive service requests from a mobile user is assumed to be exponentially distributed with an average rate of λ. Note that these exponential distribution assumptions can be relaxed, if desired, in the SPN model developed in this paper by using SPN evaluation tools that allow general time distributions to be specified, such as SPNP version 6 [16] and TimeNET version 3 [17].

In our proxy-based scheme, a mobile user will create a client-side mobile proxy migrating with the mobile user but sitting at the fixed side of the wireless network to keep track of the mobile user's current location for data delivery. That is, a data packet from the server will be delivered to the proxy

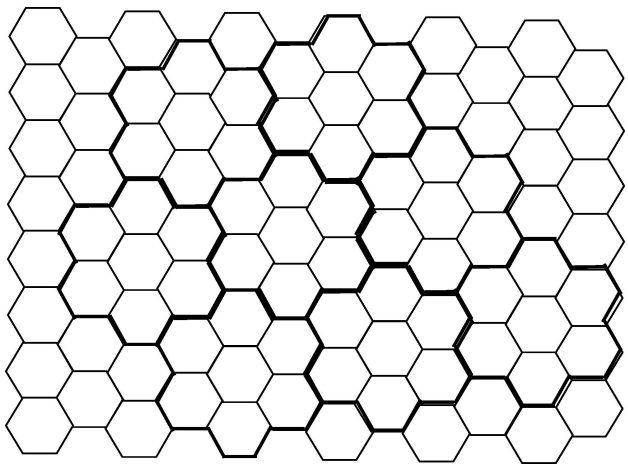


Figure 1. A hexagonal architecture for PCS cellular networks.

who in turn will forward the packet to the mobile user in the current cell. The personal proxy at all times knows the location of the mobile user by cooperating with the underlying location management system.

We model the proxy service area by a N-ring area such that a service handoff will occur when the N-ring service area is crossed. Follow our model for the PCS network, an *N-ring* service area covers many cells, depending on the value of N as illustrated in figure 1. When a mobile user crosses a cell boundary, i.e., when a location handoff occurs, the underlying location management system will inform the mobile user's personal proxy of the new address. If the service area is also crossed as a result of this location handoff operation, the proxy will also perform a service handoff operation to move into the new service area, in which case a reconnection operation is performed to inform the server of the new address of the proxy. If the service area is not crossed, on the other hand, the proxy will stay put, in which case the proxy records the current location of the mobile user in its internal database. The server at all times only communicates with the mobile user's proxy for data packets to be delivered to the mobile user. Similarly, the mobile user communicates with the server through its proxy at all times. Since the proxy knows the current location of the mobile user, it can provide location-aware and personalized service to the mobile user.

We follow the assumption in [7] for the overhead involved in performing a service handoff, namely, a reconnection cost and a service context transfer cost. The physical reconnection cost refers to the communication cost for the proxy to inform the server of the new network address (and session reestablishment for connection-oriented services such as those based on TCP), while the service context transfer cost refers to the communication cost to move the service context with the moving proxy. The amount of service context information is application-dependent and may include both static context information (such as user profile and authentication information) and dynamic context information (such as files opened, objects updated, locks and time-stamps, and status of execution). Our scheme aims to find the optimal proxy service area, when given a set of parameter values characterizing the network and workload conditions, such that the overall communication cost due to location and service management operations (including location management cost for servicing service handoffs and location handoffs, and service management cost for servicing data delivery) is minimized.

System parameters that characterize the network and user workload condition of a PCS system are summarized in Table 1 for easy reference. Here we note that three set of parameters are considered, namely, user parameters (e.g. σ), application-specific parameters (e.g. λ, α, β) and network parameters (e.g. T, τ).

Table 1
Parameters.

λ	The aggregate service request rate, i.e. the data packet rate for all services currently accessed by a mobile user.
σ	The average rate at which the mobile user moves across cell boundaries.
SMR	Service rate to mobility rate ratio, e.g., λ/σ.
T	The average communication cost between a proxy and a server per packet.
τ	The average communication cost between two neighboring cells per packet.
α	The reconnect parameter, i.e., the communication cost parameter to setup a new connection with the server when the personal proxy moves. For example, for an application on TCP, α is the number of messages to tear down the old TCP connection, setup a new TCP connection and any application specific messages in a service handoff.
β	The context transfer parameter, i.e., the communication cost parameter to transfer service context when the personal proxy moves.
C_{pt}	The proxy-move cost, including the connection setup cost and context transfer cost, i.e. $\alpha T + \beta N\tau$.
λ_i	The service request rate for a particular service, i.e., the data packet delivery rate for service i.
α_i	The service-specific communication cost parameter related with the physical connection with service i when its service proxy moves.
β_i	The service-specific communication cost parameter related with context transfer with service i when its service proxy moves.

3. Personal proxy schemes for reducing network cost of personalized services

In this section, we first describe a mobile service management scheme based on the notion of aggregate personal proxy, that is, a single personal proxy is used for all mobile services engaged by the mobile user. Then we extend the discussion to the case of per-service personal proxy for performance optimization. Later in Section 4, we will develop analytical models based on Petri nets to analyze their performance characteristics.

3.1. Aggregate personal proxy scheme

Under the aggregate personal proxy scheme, each mobile user on power up creates a client-side personal proxy that acts on behalf of the mobile user. Initially the aggregate proxy will reside in the base station of the cell in which the mobile user resides. All messages exchanged between the client and any service will go through the personal proxy. The personal proxy performs tasks such as location tracking, accepting user requests to access services, converting communication data in different application formats, and forwarding data packets to the mobile user. There is only a single proxy regardless of the number of services engaged by the user. All servers at all times only know the personal proxy. The personal proxy may move when the user moves across a cell boundary if justified, in which case the proxy will move from the base station it had resided to the base station which the mobile user just entered. When a proxy moves, a cost incurs for reestablishing the connection and transferring service context. In return, the proxy is moved nearer to the mobile user so the communication cost from the proxy to the mobile user is reduced. Thus, there exists a tradeoff between

the cost incurred due to moving the proxy vs. the cost saved due to close proximity between the proxy and the mobile user.

The aggregate personal proxy scheme has its root derived from the notion of "local anchor" (LA) proposed by Ho and Akyildiz [6] in the context of location management. The basic idea is that within a personal proxy service area, we use the user's personal proxy to keep track of the location of the mobile user within the area. The underlying location management system informs the proxy whenever the mobile user crosses a cell boundary, so the proxy at all times knows the current cell of the user. As a personal proxy area normally covers a large geographic region spanning several cells, so when a mobile user crosses a cell boundary it may be still within the same service area. In this case, the personal proxy stays in the same location without moving with the mobile user. On the other hand, if the mobile user moves out of the current service area into another service area upon a cell boundary crossing, then the proxy will move with the mobile user into the new area. In this latter case, in addition to the location management cost incurred for the system to inform the proxy of the location change of the mobile user, there is also a cost to inform the server of the network address of the new proxy and to transfer the service context to the new proxy.

A mobile user's personal proxy, in addition to keeping service context information for each service accessed by the mobile user, also keeps the mobile user's statistics information, such as the mobility rate, the packet rate for each service, and characteristics of services currently accessed by the mobile user to determine the optimal service area. Upon being informed of the new location of the mobile user when the mobile user moves into a new cell, the proxy will check if the service area is crossed. If yes, after the proxy moves into the new service area, a new optimal personal proxy service area size will be determined by executing a computational procedure developed in the paper based on the up-to-date statistics information maintained.

It should be noted that in the aggregate personal proxy scheme, there is only a single user proxy that acts on behalf of the user in the fixed network serving as the client-side agent for all services engaged by the mobile user. As a result, the optimal personal proxy service area determined by the proxy to minimize the network signaling and communication cost will be based on the aggregate service characteristics exhibited by all services, e.g., an aggregate packet rate, as the service area determined by the proxy will apply to all services engaged by the mobile user.

Figure 2 illustrates a scenario in which a mobile user moves under the aggregate personal proxy scheme. Initially the optimal proxy area size is determined to be $N_{\text{opt}} = 3$. A mobile user resides in cell A together with the proxy who resides at the center cell of the service area. When the mobile user makes a move from cell A to B, the proxy in cell A is notified. A similar location management operation is performed when the mobile user subsequently moves from cell B to C. In the meantime, all packets from S1 and S2 to the mobile user will be delivered to the proxy in A first, and then forwarded by the proxy to the mobile user. When the mobile user moves to D, which is outside of the proxy's service area, the proxy, along with the services context, is moved to D, triggering a service handoff to inform all services (S1 and S2) of the new network address of the proxy (now in cell D) and to transfer context information from cell A to cell D. Depending on the current state information, the new proxy service area may or may not be the same as before. It will be determined dynamically by the proxy after a service handoff based on a computational procedure which we will discuss later. Figure 2 shows that the new proxy service area size is now $N_{\text{opt}} = 2$ after the proxy moves to cell D.

3.2. Per-service personal proxy scheme

Unlike the aggregate proxy scheme where the optimal proxy service area is determined based on aggregate characteristics of services being accessed by the mobile user, the per-service personal proxy scheme creates a separate proxy for each client-server application engaged by the mobile user. Each proxy created is application-specific and, as it knows specific service characteristics of the application, can optimally determine the best service area for the application. For example, a high-speed data service with a small handoff cost may dictate a different optimal proxy area from the one having a low speed data service with a high handoff cost. The disadvantage of the scheme is a small processing overhead added to the mobile user since each mobile user needs to keep a list of proxies for multiple services that it is currently accessing. The advantage in return is that each service can have its own service-tailored optimal proxy service area, thus collectively reducing the overall network signaling and communication cost compared with the aggregate personal proxy scheme.

Each personal proxy behaves the same as the one in the aggregate scheme except that it only maintains its own service-specific context and statistics information. Note that it is possible that different proxies may have different optimal proxy

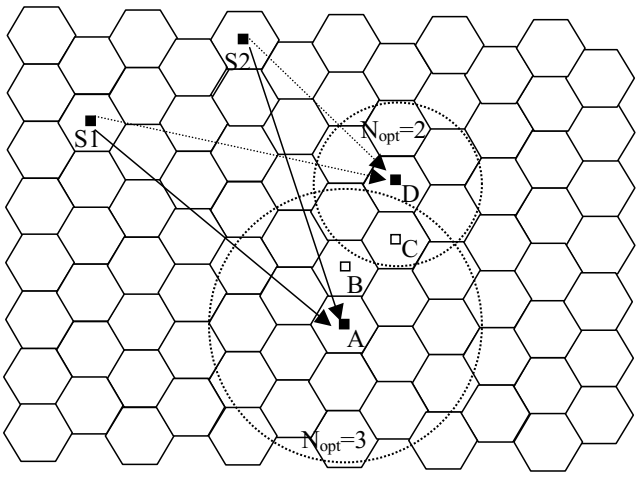


Figure 2. Aggregate personal proxy scheme.

service areas since in general different services exhibit different service characteristics.

We illustrate a user movement scenario under the per-service personal proxy scheme in Figure 3. Initally the optimal proxy service area sizes for S1 and S2 are $N_{\text{opt}} = 3$ and 2, respectively. The service area for S1 is marked dashed, while that for S2 is marked solid. Initially assume that the mobile user resides in cell A together with the two proxies of S1 and S2. When the mobile user makes a move from A to B, both proxies are notified of the movement. When the mobile user subsequently moves to C, it is still inside S1's proxy service area but outside of S2's proxy area. Thus, a service handoff for S2 is triggered, after which the proxy of S2 moves to cell C and a new proxy area size is determined (still 2 in the diagram). When the mobile user moves to D, a service handoff for S1 is triggered, after which the proxy of S1 moves to D and a new optimal proxy service area size is calculated (2 in the diagram). The proxy for S2 remains in C since the current location of the mobile user, namely, D, is still within S2's service area. All packets from S1 and S2 to the mobile user are delivered to their respective per-service proxies who in turn forward them to the mobile user.

4. Performance model

In this section, we develop analytical models for evaluating the aggregate and per-service personal proxy schemes introduced in Section 3. We first define the performance metric used as the basis for evaluation. Then, we show how the performance metric can be evaluated through our analytical model.

4.1. Performance metric

Our performance metric used for evaluating location-aware proxy-based mobile service management schemes is based on the *total communication cost per time unit* for the network to service location and service management operations. Specifically, our performance metric considers two cost parameters:

- Location management related cost C_{move} – this includes the cost for tracking the location of the mobile user and the cost for moving the proxy to stay closer to the mobile user if necessary as a mobile user moves across a cell boundary.
- Service management related cost C_{service} – this is the cost for the proxy to deliver data packets to the mobile user.

Note that the above two cost parameters refer to the average cost. Let C_{total} be the average cost of the PCS network in servicing the above two types of operations per time unit. Then, our performance metric C_{total}, defined as the *total cost* incurred to the PCS network *per time unit* for servicing location and service management operations, is given by:

$$C_{\text{total}} = C_{\text{move}} \times \sigma + C_{\text{service}} \times \lambda \tag{1}$$

Here σ and λ are mobile user's cell boundary crossing rate and service request rate, respectively, as described in Table 1. Note that the paging cost for locating the location of the mobile user within the current cell is not considered in the cost model because the paging cost is the same in all schemes.

In this paper, we consider that the communication cost between a mobile user and its proxy is proportional to the separating distance. The proxy is always located at the center base station of the N-ring structure as shown in Figure 1, so the distance between a mobile user in ring i and the proxy (located in ring 0) is exactly i cells apart. Note that an N-ring structure contains N rings, with ring id from 0 to $N - 1$.

4.2. Model for aggregate personal proxy scheme

For the aggregate personal proxy scheme, a Petri net model as shown in Figure 4 has been developed to analyze its behavior. Table 2 gives the meanings of places and transitions defined in the Petri net model. Here `mark(p)` returns the number of tokens held in place `p`.

The Petri net model describes the behavior of the mobile user in a PCS system operating under the aggregate personal proxy scheme for which the personal proxy area is an N-ring structure in the PCS network. It is constructed as follows:

- When a mobile user moves across a cell boundary, a token is placed in place `m`.

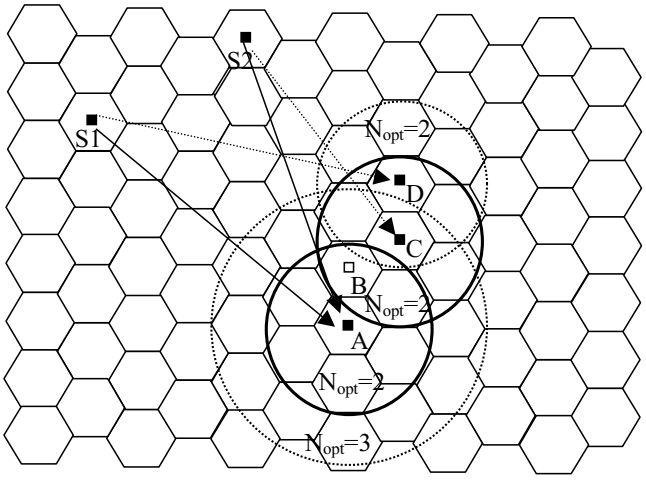


Figure 3. Per-service personal proxy scheme.

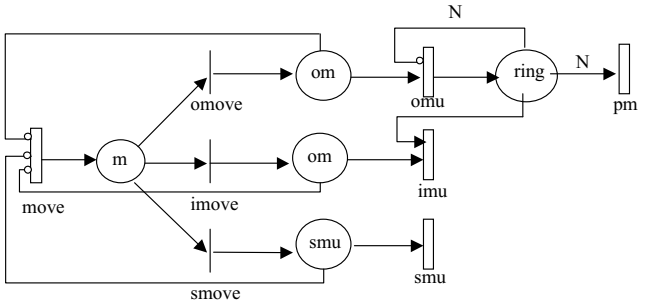


Figure 4. Petri bet model for the aggregate personal proxy scheme.

Table 2
Meaning of places and transitions in the petri net model.

`m`	`mark(m)` indicates that the mobile user has just moved across a cell boundary.
`om`	`mark(om)` indicates that the mobile user has just moved outwards, i.e., from ring i to ring $i+1$.
`im`	`mark(om)` indicates that the mobile user has just moved inwards, i.e., from ring i to ring $i-1$.
`sm`	`mark(om)` indicates that the mobile user has just moved, but still remains in the same ring.
`ring`	`mark(ring)` indicates the ring number at which the mobile user currently resides; it also is the distance between the mobile user and its proxy located at ring 0.
`move`	A timed transition representing a cell boundary crossing with a rate of σ.
`omove`	An immediate transition following a move event with P_{omove} representing the probability that the movement is an outward movement.
`imove`	An immediate transition following a move event with P_{imove} representing the probability that a movement is an inward movement.
`smove`	An immediate transition following a move event with P_{smove} representing the probability that a movement is an inside-the-same-ring movement.
`omu`	A timed transition to service an outward movement.
`imu`	A timed transition to service an inward movement.
`smu`	A timed transition to service an inside-the-same-ring movement.
`pm`	A timed transition to service a proxy transfer.

- If the movement is an "outward movement" (i.e. the user moves from ring i to ring $i+1$) with probability P_{omove} (to be parameterized), then transition `omove` will consume the token immediately, after which a token will be placed in `om` which subsequently disables transition `m` and enables transition `omu`, meaning that a local proxy update operation is being performed by the underlying location management system to inform the proxy of the current location of the mobile user. After that, a token is placed in place `ring`, meaning that the mobile user has moved from ring i to ring $i+1$. (Note: the number of tokens in place `ring` represents the id of the ring the mobile user now resides.)
- If the token number in place `ring` is equal to N then it means that the mobile user has just moved out of the personal proxy service area, in which case a service handoff occurs and the personal proxy, along with the service context, will move to a new N-ring proxy area centered at the cell which the mobile user just entered into.
- If the movement is an "inward movement" (i.e. the user moves from ring i to ring $i-1$) with probability P_{imove}(to be parameterized), then transition `imove` will consume the token immediately, after which a token will be placed in `im` which subsequently disables transition `m` and enables transition `omu`, thus triggering a local proxy update operation to be performed. After that, one token in place `ring` will be consumed, meaning that the ring number has been reduced by 1 as a result of the inward movement. Note that there must exist at least a token in place `ring` when an inward movement occurs.
- If the movement is an "inside-the-same-ring movement" (i.e. the user moves from one cell to another cell in the same ring i) with probability P_{smove} (to be parameterized), then transition `smove` will consume the token immediately, after which a token will be placed in `sm` which subsequently disables transition `m` and enables transition `smu`, representing that the proxy has been informed of the location change without a service handoff. After that, the token in place `sm` is consumed while the number of tokens in place `ring` remains the same, meaning that the ring number is not changed (since the mobile user stays at the same ring) as a result of this movement.

Note that there is no service request being modeled in the Petri net. The reason is that place `ring` keeps track of the current status, i.e. the current ring that a mobile user currently resides, and the service cost only depends on this status. Thus we are able to calculate the service request cost without having to model the service request behavior explicitly.

Suppose the personal proxy area size is N (from ring 0 to ring $N-1$). Let P_i be the steady state probability that the system is found to contain i tokens in place `ring`. Let π be the steady state average number of tokens found in place `ring`. Then the service management cost per user request, C_{service}, can be calculated by:

$$\begin{aligned} C_{\text{service}} &= \sum_{i=0}^{N-1} P_i \times C_{i,\text{service}} \\ &= \sum_{i=0}^{N-1} P_i \times (T + \tau \times i) \\ &= \sum_{i=0}^{N-1} P_i \times T + \sum_{i=0}^{N-1} P_i \times (\tau \times i) = T + \tau \times \pi \end{aligned} \quad (2)$$

where $C_{i,\text{service}}$ is the service management cost per service request when the mobile user is in ring i and is equal to the communication cost between the proxy and server (T) plus the communication cost ($\tau \times i$) between the proxy and the mobile user which are i cells apart in distance. Similarly, let C_{move} be the location management cost per move, including location update and possible context transfer costs. We have:

$$\begin{aligned} C_{\text{move}} &= \sum_{i=0}^{N} P_i \times C_{i,\text{move}} \\ &= P_0 \times \tau + \sum_{i=1}^{N-2} P_i \times (\tau \times (i+1) \\ &\quad \times P_{\text{omove}} + \tau \times i \times P_{\text{smove}} + \tau \times (i-1) \times P_{\text{imove}}) \\ &\quad + P_{N-1} \times (C_{pt} \times P_{\text{omove}} + \tau \times i \times P_{\text{smove}} + \tau \\ &\quad \times (i-1) \times P_{\text{imove}}) + P_N \times C_{pt} \end{aligned} \quad (3)$$

where C_{pt} is the proxy-move cost, including the context transfer cost and connection re-establishment cost with remote servers. We assume it has the form $\alpha T + \beta N \tau$ as described in Table 1, with α, β being proxy-move parameters dependent on

services characteristics. The *total cost per time unit* incurred to PCS network under personal proxy scheme, C_{total}, then can be calculated by:

$$C_{\text{total}} = C_{\text{service}} \times \lambda + C_{\text{move}} \times \sigma \tag{4}$$

4.3. Model for the per-service personal proxy scheme

In the per-service personal proxy scheme, a proxy is used for each service accessed by a mobile user. We can use the same performance model in Figure 4 to analyze the scheme with some adjustment made on service parameters values to account for the fact that each mobile service has its own set of service parameters. These parameters include service-specific request rate λ_i and proxy-move parameters α_i and β_i (see Table 1). Specifically, for each service accessed by the mobile user we will analyze its behavior separately utilizing the performance model shown in Figure 4. Thus, the performance model will be utilized as many times as the number of services concurrently accessed by the mobile user to obtain the overall cost incurred due to location and service management activities.

Recall that the advantage of the per-service personal proxy scheme over the aggregate personal proxy scheme is that optimizing conditions in terms of the best personal proxy areas to reduce the per-service communication cost can be separately determined for different services concurrently accessed by the mobile user. Thus in calculating the per-service cost, the parameters in equations (2), (3) and (4) will be service-specific, e.g., replacing λ, α and β by λ_i, α_i and β_i respectively. Suppose we have M services concurrently being accessed by the mobile user. Then the overall cost incurred will be:

$$C_{\text{total}} = \sum_{i=1}^{M} C_{\text{total}}(\text{service}_i)$$

where C_{total} (service$_i$) is calculated from equation (4) above for each separate service with λ, α and β being replaced by λ_i, α_i and β_i respectively.

5. Analysis

In this section, we show how to parameterize (i.e., give values of model parameters of) the SPN models developed in Section 4 by means of a hexagonal network coverage model for describing the PCS network under consideration, and devise a computational procedure for computing the total cost by the proxy during run time. Our objective is to reveal design conditions under which the network signaling and communication cost due to location and service management operations can be minimized for the PCS system operating under the aggregate and per-service personal proxy schemes and compare their performance characteristics with those by non-proxy schemes. We use SPNP [16] as a tool to define and evaluate the SPN models developed to yield numerical results with physical interpretations given.

5.1. Parameterization

Consider a hexagonal network coverage model for modeling a PCS network in which an N-layer proxy area covers $3N^2 - 3N + 1$ cells as illustrated in 1 with the center cell in ring 0 and the outmost cells in ring $N-1$. Assuming n is the current ring number at which the mobile user resides in a particular time as given by `mark(ring)` in the Petri net model, it can be shown that [13] P_{omove}, P_{imove} and P_{smove} are calculated as:

$$P_{\text{omove}} = \begin{cases} 1.0 & \text{if } n = 0 \\ \dfrac{2n+1}{6n} & \text{otherwise} \end{cases}$$

$$P_{imove} = \begin{cases} 0 & \text{if } n = 0 \\ \dfrac{2n-1}{6n} & \text{otherwise} \end{cases}$$

$$P_{smove} = \begin{cases} 0 & \text{if } n = 0 \\ \dfrac{2n}{6n} = \dfrac{1}{3} & \text{otherwise} \end{cases}$$

We use σ to represent the user mobility rate, thus the rate of the transition `move`, R_{move}, is equal to σ. Assuming the mobile user locates at ring n, the communication cost involved in an outward movement from ring n to ring $n+1$ is $(n+1)\tau$ since the mobile user is $n+1$ cells in distance from the proxy. Thus the transition rate for transition `omu` can be paramerized as $R_{\text{omu}} = 1/((n+1)\tau)$. Similarly, the transition rate for transition `imu` from ring n to ring $n-1$ is $R_{\text{imu}} = 1/((n-1)\tau)$ and the transition rate for transition `smu` for an inside-the-same-ring movement is $1/(n\tau)$.

When the mobile user moves across an N-ring proxy area, thus triggering a service handoff, the communication cost involves a context transfer operation from one N-ring proxy to another with the cost of $\beta \times N \times \tau$, and a connection transfer operation from the proxy to the service with the cost of $\alpha \times T$. Thus we can parameterize the transition rate for transition `pm` as $R_{pm} = 1/(\alpha \times T + \beta \times N \times \tau)$.

5.2. Computational procedure for calculating C_{total}

To calculate the total communication cost C_{total} based on Equations (2), (3) and (4), we need to obtain the steady state probability that i tokens are found in place ring, P_i, and the steady-state average number of tokens in place ring, π. SPNP was used to help obtaining these when given a set of parameter values characterizing the network and workload conditions. Specifically, we used the following reward assignment to calculate P_i:

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

In effect, this will calculate the *average reward* weighted by the state probabilities, which in this case, is exactly the probability that i tokens are found in place `ring`. To calculate π, we used the following reward assignment: $r =$ `mark(ring)`.

5.3. Numerical data

We report numerical data to (a) show that there exists an optimal personal proxy area in our proposed service management schemes based on location-aware personal proxies to minimize the overall network signaling and communication cost, (b) compare our proposed schemes with non-proxy-based schemes in the PCS system and (c) study the effects of certain model parameters, including the SMR and context transfer parameters, on the optimal personal proxy area size. The numerical data are obtained by using SPNP as a tool to define and evaluate the SPN models developed following the parameterization process explained in Section 5.1 and the computational procedure in Section 5.2.

We first compare location-aware personal proxy schemes with non-proxy service management schemes, as a function of model parameters to analyze conditions under which, if any, non-proxy can perform better than proxy-based schemes. Figure 5 compares the cost of non-proxy scheme and personal proxy-based schemes under varying SMR (i.e., λ/σ) ratios with mobility rate σ set at 0.1 and proxy-move parameters chosen at $\alpha = 4$ and $\beta = 2$. The effect of proxy-move parameters α and β on the system performance will be analyzed later. The top curve shows the total cost obtained under the non-proxy scheme. The bottom curve shows the total cost obtained under the location-aware proxy scheme when operating at optimizing proxy service areas (that is, at N_{opt}) as identified from our model. There are several middle curves in between these two curves showing the total cost obtained at various proxy service areas. Of particular interest is the case when N=1 for which the proxy always moves with the user whenever the user moves across a cell boundary.

We observe that the non-proxy scheme possibly could perform better than the proxy-based scheme under low SMR ratios and large proxy areas (that is, large N). However, if the proxy service area is optimally selected at N_{opt}, the proxy-based scheme always performs better than the non-proxy scheme. Further, the advantage of the proxy-based scheme becomes more and more pronounced with the increase of SMR. The reason is that when SMR is low, the packet arrival rate is low compared with the user mobility rate; thus, the service management cost incurred in the non-proxy scheme due to triangular routing in servicing packets is minimal. This factor, when coupled with a large service area which incurs a large service handoff cost in the proxy-based scheme, can make the non-proxy scheme perform better than the proxy-based scheme in terms of the overall service and location management cost incurred to the network. On the other hand, as SMR increases the high service management cost for packet delivery due to the triangular routing in the non-proxy scheme dominates the location management cost, making non-proxy schemes perform worse than proxy-based schemes, regardless of the service area in the proxy-based scheme in this case.

Next we study the effect of proxy-move parameters α and β. Figure 6 plots the total cost incurred per unit time as a function of proxy-move parameters α and β with SMR set at 10, for three different schemes, namely, non-proxy (top-plane), proxy-based at N=1 (middle plane) and proxy-based at N_{opt} (bottom plane). We see that the personal proxy scheme at N_{opt} incurs the least cost; the non-proxy scheme incurs the most cost among the three schemes; and the personal proxy scheme with N=1 falls within between. Recall that when a proxy moves in the personal proxy scheme, the service handoff cost is modeled by $C_{pt} = \alpha \times T + \beta \times N \times \tau$. We see that the total cost is more sensitive to α as it is to β. The total cost changes from 1.1 to 1.361 when α increases from 1 to 9 under fixed $\beta = 1$, while it only changes from 1.1 to 1.207 when β increases from 1 to 9 under fixed $\alpha = 1$. The reason is that τ (the communication cost between two cells) is relatively small compared to T (the communication cost between the server

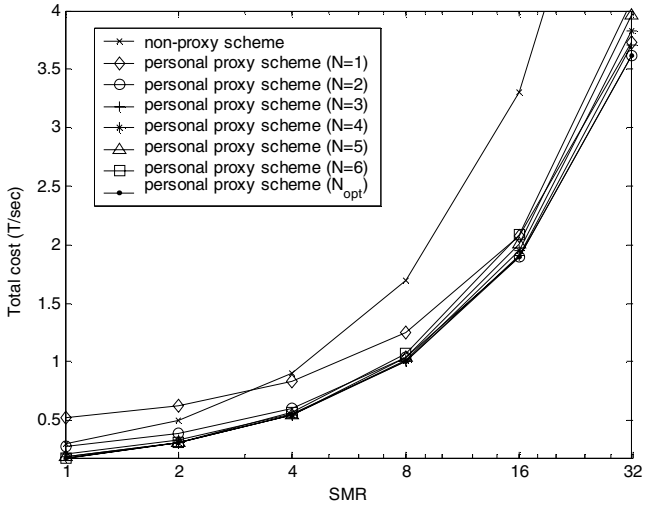


Figure 5. Comparison of proxy-based cs. Non-proxy service management schemes.

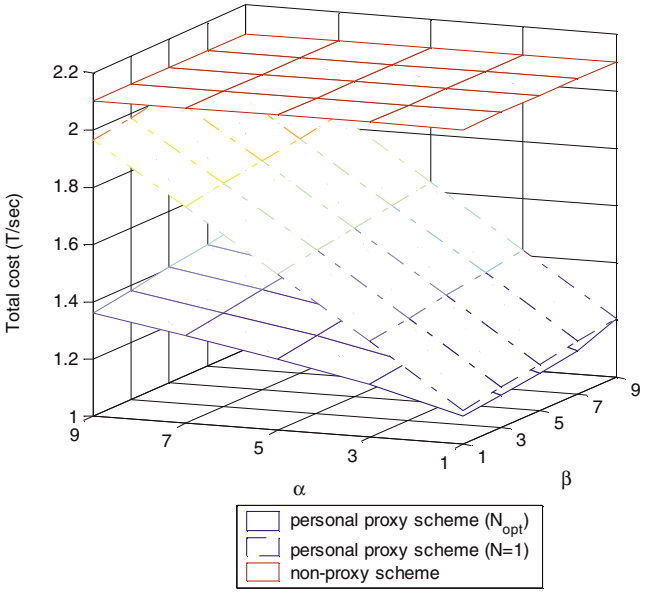


Figure 6. Total cost as a function of proxy-move parameters α and β.

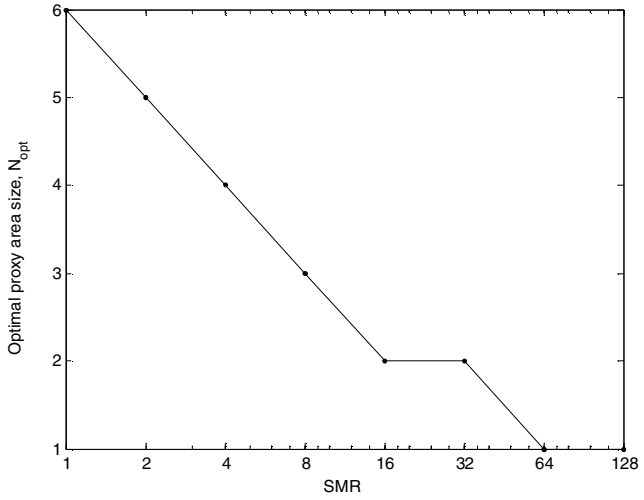


Figure 7. Optimal proxy area size under different SMR values.

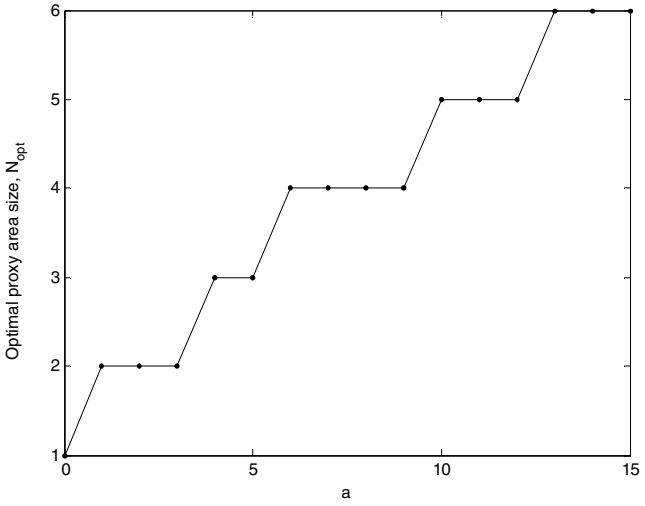


Figure 8. Optimal proxy area size under different α values.

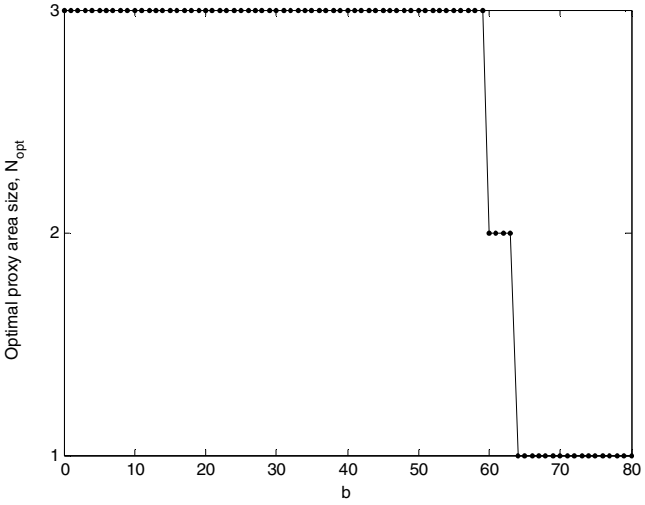


Figure 9. Optimal proxy area size under different β values.

and proxy) with $\tau = 0.1T$ in the case study. Thus the effect of β on the proxy-move cost through the cost contributing term $\beta \times N \times \tau$ is small compared with the effect of α through the other cost contributing term $\alpha \times T$. Another observation is that though the total cost increases with the increase of α and β, the proxy scheme performs much better than the non-proxy scheme over a wide range of α, β considered. Finally, it is noteworthy that moving the proxy with the mobile user whenever the mobile user moves across a cell boundary (i.e., at $N = 1$) is not necessarily the best proxy-based scheme.

Next we analyze the effect of model parameters on the optimal service area size N_{opt}. Figure 7 plots the optimal proxy area size N_{opt} under different SMR (i.e., λ/σ) ratios. To isolate the effect of SMR, we again set the context transfer parameters $\alpha = 4$ and $\beta = 2$. For the aggregate personal proxy schemes investigated in the paper, the optimal proxy area size N_{opt} is a design parameter determined by the characteristics of the operating and workload conditions of the PCN system. As shown in the figure, the N_{opt} value decreases as the SMR ratio increases. When the SMR is low, the mobility rate is high compared to the packet arrival rate, thus the location management cost dominates the service management cost. A larger proxy area reduces the number of costly proxy-move operations for service context transfer and reconnection at the expense of an increased packet/call delivery cost due to a larger distance separating the proxy to the mobile user's current location. Since the packet rate is low compared to the mobility rate when the SMR ratio is low, the total system cost is reduced with a larger proxy area. Conversely, when the SMR ratio is high, i.e., the packet rate is much higher than the mobility rate, a smaller proxy area will reduce the packet delivery cost, making the service management cost dominate the location management cost, and, as a result, would reduce the total network cost.

Figures 8 and 9 show the effect of parameters α and β on the optimal proxy area size N_{opt} with $SMR = 10$. Figure 8 shows that the optimal proxy area size increases as α increases. The reason is that with a large α value, the proxy-move cost between the proxy and server is high. Thus, the proxy will tend to stay in the same service area to avoid costly service handoff operations. This effect favors a large proxy area. Conversely, Figure 9 shows that the optimal proxy area size decreases as β increases. The effect of β on the optimal proxy area is counterintuitive. A large β value represents a high context transfer cost and, thus, a large proxy area seems preferable. However the result shows that optimal proxy area is quite insensitive to β and actually decreases when β increases to a high enough value. The reason is the context transfer cost, i.e., $\beta \times N \times \tau$, not only depends on β but also depends on $N$and τ since the network communication cost is proportional to the distance separating the proxy and the mobile user. As a result, although a large β value indirectly prefers a large proxy area, i.e. a large N, the context transfer cost directly favors a small proxy area.

All the above analysis is based on the aggregate proxy-based service management schemes where there is a single

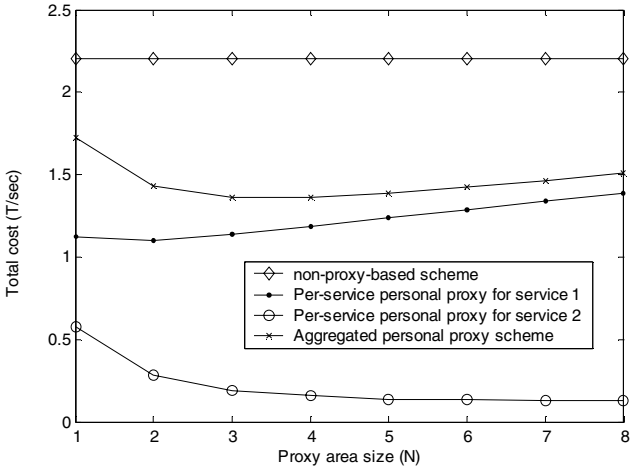


Figure 10. C_{total} under different proxy area sizes.

service, or there are multiple services but only one proxy is used to interface with all services taking the aggregate characteristics into consideration. Below we construct a case study to compare aggregate vs. per-service proxy-based mobile service management schemes.

Figure 10 shows the effect of the size of the proxy service area (an N-ring area) on the overall cost incurred to the system due to location and service management operations under the proxy schemes, with $\tau = 0.1T$ and $\sigma = 0.1$. We consider the case in which there are two services being accessed by a mobile user concurrently. One service is a UDP-like multimedia service with packet delivery rate $\lambda_1 = 1$ packet/second, and proxy-move cost parameters $\alpha_1 = 1$, $\beta_1 = 1$. Another is a TCP-based service (e.g. telnet) with packet delivery rate $\lambda_2 = 0.05$ packet/second, and proxy-move cost parameters $\alpha_2 = 5$, $\beta_2 = 2$. Thus the aggregate packet delivery rate is $\lambda = \lambda_1 + \lambda_2 = 1.05$, $\alpha = \alpha_1 + \alpha_2 = 6$ and $\beta = \beta_1 + \beta_2 = 3$.

We also consider the cost of a non-proxy-based scheme for which the HLR is being informed of the new network address of the mobile user whenever the mobile user moves across a cell boundary (with a cost of T per move), and a triangular routing is incurred for packet delivery as in Cellular Digital Packet Data (CDPD) systems, that is, each packet to the mobile host will travel from the server to the HLR (with cost T) and then from the HLR to the current address (with another cost T). Consequently for non-proxy-based scheme, the total cost is $T \times \sigma + (T + T) \times \lambda = 2.2T$ and remains a constant with the change of N.

From Figure 10 we first see that there exists an optimal proxy service area size that highlights the tradeoff between location management cost and service management cost. On the one hand, with the increase of N, and thus a larger service area, the service management cost (e.g. packets delivery) is higher due to the higher communication cost from the proxy to the current location of mobile user. On the other hand, with a larger service area, the location management cost is reduced more because a user movement crossing a cell boundary is more likely to be within the same service area and thus the proxy needs not to be moved, thus resulting in a lower location management cost. The optimal value is reached when N=4 for the aggregate personal proxy scheme, at which the network signaling and communication cost incurred to the PCS network is minimized while maintaining the required service and location functionality. For the per-service personal proxy scheme, service1 and service2 have the optimal N values at 2 and 7, respectively.

Figure 10 also demonstrates the superiority of the per-service personal proxy scheme over the aggregate personal proxy scheme. The total costs incurred for service1, service2 and the aggregate service under optimal proxy area sizes are 1.0998, 0.1294 and 1.3624, respectively. Thus, the cost is reduced by (1.3624-1.0998-0.1294)/1.3624 = 9.8% in the per-service personal proxy scheme compared to the aggregate personal proxy scheme. Note that here the cost metric is amount of cost incurred to the system per time unit (second), so even a 10% difference is considered significant.

Lastly, we observed that the improvement of the per-service personal proxy scheme over the aggregate personal proxy scheme is more pronounced when the services characteristics (e.g. packet rate, context transfer cost, etc.) of multiple services accessed by the mobile user are dramatically distinct, because otherwise the aggregate service characteristics would be close to those of individual ones and the optimal service area found by the aggregate scheme would be close to those individually found by separate services, making the performance behavior virtually the same between these two schemes.

6. Conclusion

In this paper, we investigated the concept of location-aware mobile service management schemes based on personal proxies with the objective to reduce the overall communication cost for servicing location and service management operations in the PCS network environment. Two location-aware mobile service management schemes were introduced: the aggregate and per-service personal proxy schemes. We developed SPN performance models to help identify the optimal proxy service areas for the proposed proxy-based schemes and devised a computational procedure to be utilized by a mobile host to dynamically determine the best proxy service area per service in order to minimize the network cost based on runtime estimates of model parameter values characterizing a client-server application in the wireless PCS system. We showed that (a) there exists an optimal proxy service area under which the network signaling and communication cost for location and service management operations can be minimized; (b) these two schemes operating at optimizing service areas outperformed non-proxy schemes over a wide range of parameter values for which the conditions under which proxy-based schemes perform better than non-proxy ones are characterized and identified; (c) the per-service proxy scheme performed better the aggregate proxy scheme since separate services can operate at their respective optimal proxy service

areas for service handoffs, resulting in the collective overall cost incurred to the network smaller than that based on the aggregate proxy scheme which considers only aggregate service characteristics.

Future research areas extending from this work include (a) applying the location-aware mobile personal proxy concept for handling location and service handoffs in Mobile IP and/or SIP environments; (b) investigating the possibility of designing a more tightly integrated location and service management scheme such that per-user per-service proxies for service management are collocated with per-user per-service location databases for location management to further reduce the overall network cost for location and service management.

References

[1] P. Bellavista, A. Corradi and C. Stefanelli, The ubiquitous provisioning of Internet services to portable devices, IEEE Pervasive Computing 1(3) (2002) 81–87.

[2] E. Brewer et al., A network architecture for heterogeneous mobile computing," IEEE Personal Communications 5(5) (1998) 8–24.

[3] M.H. Dunham and V. Kumar, Impact of mobility on transaction management, in: *International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '99, Seattle, WA, USA (1999) pp. 14–21.

[4] *EIA/TIA*, Cellular Radio Telecommunication Inter system Operations, Technical Report IS-41 (Revision B), EIA/TIA, (July 1991).

[5] M. Endler, D.M. Silva and K. Okuda, RDP: A result delivery protocol for mobile computing, in: *Int. Workshop on Wireless Networks and Mobile Computing (WNMC) at the 20th Int. Conference on Distributed Computing Systems (ICDCS)*, Taiwan (2000).

[6] J.S.M. Ho and I.F. Akyilidiz, Local anchor scheme for reducing signaling costs in personal communications networks, IEEE/ACM Transactions on Networking 4(5) (1996) 709–725.

[7] R. Jain and N. Krishnakumar, Network support for personal information services to PCS users, in: *IEEE Conference on Networks for Personal Communications* (1994) pp. 1–7.

[8] R. Jain, Y.B. Lin, C. Lo and S. Mohan, A caching strategy to reduce network impacts of PCS, IEEE Journal on Selected Areas in Communications 12(8) (1994) 1434–1444.

[9] R. Jain, Y.B. Lin, C. Lo and S. Mohan, A forwarding strategy to reduce network impacts of PCS, in: *14th Ann. Joint Conf. of the IEEE Computer and Communications Societies (IEEE INFOCOM '95)*, Boston, MA (1995) pp. 481–489.

[10] A. Joshi, On proxy agents, mobility, and web access, ACM Journal on Mobile Networks and Application 5(4) (2000) 233–241.

[11] J. Kammann and T. Blachnitzky, Split-proxy concept for application layer handover in mobile communication systems, in: *4th IEEE Conference on Mobile and Wireless Communications Networks*, Stockholm, Sweden (2002).

[12] W.R. Lai and Y.B. Lin, Mobility database planning for PCS, in: *1996 Workshop on Distributed System Technologies and Applications*, Tainan, Taiwan (1996) pp. 263–269.

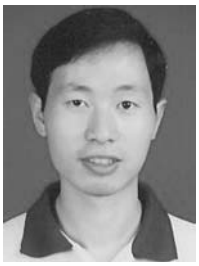
[13] Y.B. Lin, L.F. Chang and A. Noerpel, Modeling hierarchical microcell and macrocell PCS architecture, in: *1995 IEEE International Conference on Communications (ICC)*, Seattle (1995) pp. 18–22.

[14] K. Pahlavan et al. Handoff in hybrid mobile data networks, IEEE Personal Communications 7(2) (2000) 34–47.

[15] M. Roussopoulos et al. Personal-level routing in the mobile people architecture, in: *USENIX Symposium on Internet Technologies and Systems*, Boulder, CO(1999).

[16] K.S. Trivedi, G. Ciardo and J. Muppala, *SPNP Version 6 User Manual* (Dept. of Electrical Engineering Duke University, Durham, NC, 1999).

[17] A. Zimmermann, User Manual 3.0, *TimeNET: A Software Tool for the Performability Evaluation with Stochastic Petri Nets* (TU Berlin, 2001).



Baoshan Gu received the BS degree from University of Science and Technology of China, Hefei, China, in 1992 and the MS degree in computer science from Institute of Computing Technology, Chinese Academia of Science, Beijing, China, in 1995. From 1995 to 2000, he was a research and development engineer in Institute of Computing Technology, Chinese Academia of Science. He is currently pursuing his PhD degree in the Department of Computer Science, Virginia Tech, where he is a research assistant in the Systems and Software Engineering Laboratory. His research interests include next-generation wireless system architectures, design and evaluation of location and service management schemes in mobile computing environments, and mobile database systems.
E-mail: bgu@vt.edu

Ing-Ray Chen received the BS degree from the National Taiwan University, Taipei, Taiwan, and the MS and PhD degrees in computer science from the University of Houston, Texas. He is currently an associate professor in the Department of Computer Science at Virginia Tech. His research interests include mobile computing, pervasive computing, multimedia, distributed systems, real-time intelligent systems, and reliability and performance analysis. Dr. Chen has served on the program committee of numerous conferences, including being as program chair of 14th IEEE International Conference on Tools with Artificial Intelligence in 2002, and 3rd IEEE Symposium on Application-Specific Systems and Software Engineering Technology in 2000. Dr. Chen currently serves as an Associate Editor for *IEEE Transactions on Knowledge and Data Engineering*, *The Computer Journal*, and *International Journal on Artificial Intelligence Tools*. He is a member of the IEEE/CS and ACM.
E-mail: irchen@vt.edu