SCALING THE MOBILE INTERNET

SAMP: Scalable Application-Layer Mobility Protocol

Sangheon Pack, University of Waterloo
Kunwoo Park, Taekyoung Kwon, and Yanghee Choi, Seoul National University

ABSTRACT

As wireless/mobile technologies evolve, the market for mobile Internet is rapidly growing. To proliferate mobile Internet market, scalable mobility support is a key question. In this article, we propose a scalable application-layer mobility protocol (SAMP) that is based on peer-to-peer (P2P) overlay networking and session initiation protocol (SIP). SAMP keeps track of locations of mobile nodes leveraging SIP mobility functions. In SAMP, all SIP servers form a P2P overlay network, which enables scalable, load balanced, and robust mobile Internet. In addition, two optimization techniques, hierarchical registration and two-tier caching schemes, are employed to localize signaling traffic for mobility and to reduce the session setup latency, respectively. Simulation results demonstrate that SAMP is an attractive choice for scalable mobile Internet and its limitations such as long handoff and session setup latencies are substantially mitigated by the optimization techniques.

INTRODUCTION

With the advance of wireless/mobile communication technologies, the market for mobile Internet is drastically growing. For example, cellular networks have ever increased the link bandwidth with recent standardization efforts on the IP multimedia subsystem (IMS). Also, Wi-Fi hotspot services-based IEEE 802.11 wireless local area networks (WLANs) have been widely deployed in public spaces such as airports, convention centers, cafes, and so forth. Furthermore, new wireless technologies (e.g., IEEE 802.16/20) are emerging and will be available in the near future. All of these technologies will accelerate the growth of the mobile Internet market, and mobility management is one of the key issues to proliferate mobile Internet services.

We focus on the scalability aspect of mobility management in mobile Internet services. In typical mobile Internet architectures, a few types of mobility agents are employed for mobility management: a foreign agent (FA) in Mobile IPv4 (MIPv4) [1], a home agent (HA) in Mobile IPv4 and Mobile IPv6 (MIPv6) [2], and a mobility anchor point (MAP) in Hierarchical Mobile IPv6 (HMIPv6) [3]. These mobility agents play important roles in mobility management and packet routing. Hence, if a high burden of these tasks is concentrated on a single mobility agent, the mobility agent will suffer from the increased processing load and this results in a long response time and even a system failure. Moreover, the overload at the mobility agent may lead to service unavailability. Consequently, how to provide a scalable service by distributing the network traffic load (for mobility management and packet routing) among multiple mobility agents is an important design issue to be resolved in mobile Internet services.

To address the scalability problem, a number of schemes such as a dynamic HA assignment mechanism [2] have been proposed in the literature. However, these approaches result in additional signaling overhead to learn the current load condition and to synchronize among multiple mobility agents. Furthermore, in the current mobile Internet architecture, it is not easy to design a self-organized and load-balanced mechanism among multiple mobility agents scattered across different network domains.

In this article, we propose a scalable application-layer mobility protocol (SAMP), which is based on peer-to-peer (P2P) overlay networking and session initiation protocol (SIP) [4]. SIP is a simple application-layer protocol that is originally designed for session management, but it can also be utilized to provide terminal, service, and personal mobility. In SAMP, each mobile node (MN) performs location registration and its location is tracked via SIP messages. In addition, all SIP servers in SAMP form a P2P overlay network, which addresses scalability, load balancing, and robustness issues incurred in the existing mobile Internet architecture. Furthermore, two optimization techniques, hierarchical registration (HR) and two-tier caching (TTC) schemes, are introduced to reduce the handoff latency and the session setup latency.

The remainder of this article is organized as follows. First, the background of SIP and P2P networking is presented. After that, the architecture and location registration/session setup procedures in SAMP are described. Also, two optimization techniques and simulation results are presented. Previous mobility-support solutions based on P2P networking are summarized and compared. Finally, the concluding remarks are given.
BACKGROUND

Two key elements of SAMP are SIP and P2P overlay networking, which are described briefly in this section. Actually, SIP-based P2P overlay networking was presented in [5] for IP telephony. However, SAMP focuses on mobility support rather than a specific application.

SESSION INITIATION PROTOCOL

SIP is an Internet standard protocol for initiating, modifying, and terminating an interactive multimedia session. The multimedia session involves various applications such as video, voice, instant messaging, and online games. Moreover, SIP is accepted as a call-control protocol in IMS. SIP can also be employed to support mobility at the application layer [6]. Especially, SIP is an appropriate mobility solution to interactive multimedia applications that need an explicit signaling for session management. In addition, SIP allows users to maintain access to their services while moving (i.e., service mobility) and to maintain sessions while changing terminals (i.e., session mobility).

A typical SIP architecture consists of SIP servers and user agents. SIP servers are classified into proxy, redirect, and registrar servers, depending on their functions. A proxy server relays received SIP messages to another SIP server or user agents, whereas a redirect server performs redirection of received SIP messages. A registrar maintains location information to support mobility. On the other hand, user agents are classified into user agent client (UAC) and user agent server (UAS). Each user agent is identified by a SIP universal resource identifier (URI) that follows a form similar to an email address (e.g., sip:shpack@domain.com). The UAC initiates a SIP session by sending an INVITE message, while the UAS responds with SIP reply messages that contain suitable status codes. Basically, the user agent registers its location at the registrar before establishing a SIP session.

P2P OVERLAY NETWORK

A P2P overlay network is a distributed network that relies on the computing power and bandwidth of peer nodes in the network. Unlike the client-server model, each node participates in the P2P overlay network as a peer with equal responsibility. In P2P overlay networks, since there is no central entity to control overall tasks, system unavailability due to the failure of the central entity is diminished. Also, P2P overlay networks provide self-organization and load-balancing functions in a distributed manner. Another attractive feature is the techniques for locating and retrieving a desired item (e.g., a file in file-sharing applications). For more efficient locating/retrieving operations, a distributed hash table (DHT) has been introduced. The DHT is a decentralized and distributed system where all items and peer nodes are identified by unique keys. In the DHT, the ownership of keys is distributed among participating peer nodes and hence the peer nodes can efficiently route messages to the owner of any given key. Therefore, the DHT is scalable to a large number of nodes and can handle continual node arrivals and failures. These features enable the DHT to be widely accepted for large-scale P2P networking.

Stoica et al. have proposed a novel DHT-based P2P networking protocol called Chord [7]. Chord is completely decentralized and can find an item using at most $\log_2(N)$ messages, where $N$ is the number of nodes in the system. In Chord, peer nodes and keys are arranged in a ring-shaped $m$-bit identifier space. A key $k$ is assigned to the first node whose identifier is equal to or follows $k$ in the identifier space. Chord uses a finger table to accelerate lookup procedures. Each node maintains a finger table with size of $O(\log(N))$ and resolves a lookup procedure within $O(\log(N))$ steps. Chord is a simple, efficient, and scalable P2P lookup primitive and hence SAMP employs Chord as a reference P2P lookup protocol.

SAMP: SCALABLE APPLICATION LAYER MOBILITY PROTOCOL

ARCHITECTURE

In SAMP, each SIP server participates to form a P2P overlay network, where the SIP server acts as a peer node in a DHT. As mentioned above, the DHT is a decentralized system where multiple keys are distributed among peer nodes. Unlike resource lookup applications, SAMP employs the DHT for location management in mobile Internet. Therefore, a key in the DHT corresponds to the location information of an MN in SAMP and the keys are maintained at SIP servers. To join the DHT-based overlay network, the SIP servers should perform the functions of a registrar as well as proxy/redirect servers. Hence, we define two SIP server modes: registrar and proxy (RP) mode and registrar and redirect (RR) mode. As its name implies, a SIP server in the RP mode handles SIP messages as a proxy server and also maintains location information as a registrar. Similarly, a SIP server in the RR mode redirects the received SIP messages and keeps track of MN’s location information. In this article, we describe the SAMP operations by means of RP-mode SIP servers.

LOCATION REGISTRATION

Before location registration, an MN should first find a SIP server, which acts as an entry point to trigger a location registration procedure in the P2P overlay network. The entry SIP server is called an anchor SIP server. To find available anchor SIP servers, SIP multicast can be utilized. When an MN enters a foreign domain, it first configures an IP address, that is, care-of address (CoA), using the Dynamic Host Configuration Protocol (DHCP). Then, the MN sends a REGISTER message to the well-known SIP multicast address (i.e., sip.mcast.net or 224.0.1.75). In the REGISTER message, the MN’s home SIP URI is included in the From field for identification, while the MN’s CoA is specified in the Contact field to receive response messages. SIP servers receiving the REGISTER message will respond with response messages (e.g., 200 OK). Since multicast is used, multiple responses may be received. In this case, the most intuitive way is...
for the MN to select a SIP server with the shortest response time as its anchor SIP server. However, in this approach, the MN may choose an overloaded SIP server. To achieve load-balanced server selection, SAMP employs a random-selection scheme based on application requirements. In most real-time applications, a tolerable handoff latency can be specified, for example, typically 30–50 ms for IP telephony. Accordingly, the MN randomly chooses a SIP server among multiple SIP servers with a response time less than a predefined threshold $\delta$, which is determined by the application requirements. By doing so, both load balancing and reduced response time are achieved. After deciding the anchor SIP server, the MN confirms it by sending an ACK message to the selected SIP server.

Figure 1a illustrates how to find an anchor SIP server. As shown in Fig. 1a, each SIP server acts as a peer node in the P2P overlay network and has a unique identifier (ID) (e.g., 4, 7, 12, 16, and 20). Assume that SIP servers 12 and 16 are located in the vicinity of the MN. By SIP multicast, SIP servers 12 and 16 receive the REGISTER message. Suppose that the response time of SIP server 12 is less than $\delta$ while that of SIP server 16 is greater than $\delta$. Therefore, SIP server 12 is chosen as an anchor SIP server. If no SIP servers satisfy the response time constraint, the MN may randomly select a SIP server among SIP servers that have responded to the REGISTER message.

After finding the anchor SIP server, the MN should register the current location with its home SIP server. As shown in Fig. 1b, SIP server 12 is selected as the anchor SIP server. Unlike the existing mobility protocols such as MIPv4/6 and HMIPv6, the home SIP server of an MN is determined by a consistent hash function $H(\cdot)$, (e.g., the SIP server 4 is selected in Fig. 1b). To perform location registration to the home SIP server, the MN first generates an ID by $H$ (Home SIP URI). After that, the MN sends a REGISTER message to its home SIP server that handles the corresponding ID through its anchor server. In the REGISTER message, the address of the previously chosen anchor SIP server is specified in the Contact field, which enables hierarchical registration (elaborated upon below).

**SESSION ESTABLISHMENT**

Unlike Mobile IP, SIP-based mobility-support protocols require an explicit session-establishment procedure. This session-establishment procedure is appropriate for session-based multimedia applications. During the session-establishment procedure, the MN and correspondent node (CN) exchange and negotiate session-related information.

A session establishment in SAMP is accomplished as follows. First, a CN generates the ID of a callee MN using the MN’s home SIP URI. Then, the CN sends an INVITE message toward the P2P overlay network through its anchor SIP server. The INVITE message is routed to the MN’s home SIP server by the routing algorithm employed in the SAMP overlay network. When the INVITE message arrives at the MN’s home SIP server, the home SIP server finds the MN’s current location using its registrar function and forwards the INVITE message to the found SIP server, i.e., the MN’s anchor SIP server. Since
the anchor SIP server knows the MN’s CoA, it relays the received INVITE message to the MN. Finally, the MN responds with a suitable response message and a SIP session is established. After the session establishment, the CN and MN can communicate directly without involving the overlay network.

Figure 2 shows a session-setup procedure in SAMP. First, a CN sends an INVITE message to SIP server 20, which is the CN’s anchor SIP server. Then, the INVITE message is forwarded to SIP server 4, that is, the home SIP server of the callee MN. Next, SIP server 4 relays the INVITE message to the MN’s anchor SIP server 12 and the MN receives the INVITE message from its anchor SIP server. As depicted in Fig. 2, since SAMP is based on the P2P overlay network, a session setup procedure requires a number of lookups in peer SIP servers and this may lead to a long session-setup latency. To reduce the session-setup latency, SAMP employs a two-tier caching scheme.

**OPTIMIZATION TECHNIQUES**

**HIERARCHICAL REGISTRATION SCHEME**

In wireless/mobile networks, reducing the handoff latency is important to achieve smooth and seamless services. In pure P2P systems, a home peer node in charge of maintaining the MN’s location information may be far from the current location of the MN. In such an environment, registering a new location with the home peer node for every movement significantly increases the signaling traffic and the session disruption time. To address this problem, several techniques were proposed in [8]. In SAMP, we adopt a hierarchical registration (HR) scheme that was also introduced in [6].

As mentioned above, the MN first finds the anchor SIP server when it enters a new foreign domain. The anchor SIP server maintains the mapping information between the MN’s SIP URI and the current CoA. Also, the anchor SIP server is randomly selected from multiple candidate SIP servers with a shorter response time than δ. Therefore, the registration with the anchor SIP server rather than the home SIP server can provide the reduced handoff latency. Consequently, in SAMP, when the MN moves to a new subnet, the MN needs to register its new CoA only with the anchor SIP server not the home SIP server (Fig. 3a). This localized registration guarantees correct session establishments because the home SIP server maintains the binding information between the MN’s home SIP URI and the anchor SIP server’s address. In short, the anchor SIP server acts similar to the MAP in HMIPv6 networks [3]. On the other hand, for mid-session mobility, the MN sends an INVITE message to the CN to notify its current location [6].

Actually, the response time from the current anchor SIP server varies as the MN moves. If the response time of the current anchor SIP server is still less than the delay requirement, the current anchor SIP server keeps being used. Only when the response time exceeds the delay requirement, the MN chooses a new anchor SIP server and updates its new anchor SIP server with the home SIP server.

**TWO-TIER CACHING SCHEME**

Since there is no central-location server in P2P overlay networks, location tracking of an MN requires multiple lookups at peer nodes. Therefore, P2P-based mobility-support protocols result in long session-setup latency [5], even though they provide desirable properties such as load balancing and fault tolerance. Furthermore, since the session-setup latency affects the session-blocking performance, it should be reduced as much as possible.

To reduce the session-setup latency, SAMP employs a two-tier caching (TTC) scheme. Before sending an INVITE message for session establishment, a CN looks for the callee MN’s current location CoA in its local cache. If there is a valid entry (i.e., lifetime is not expired), the CN sends the INVITE message to the address without any lookup over P2P overlay networks (phase I, Fig. 3b). If the MN resides still in the location, the CN will receive a successful 200 OK response message. Otherwise, the INVITE message will expire. If there is no entry in the local cache or the INVITE message expires, the CN sends the INVITE message to its anchor SIP server.

At the CN’s anchor SIP server, the second-phase cache lookup is triggered. Namely, the anchor SIP server also maintains a cache for callee MNs. When the anchor SIP server receives an INVITE message from the CN, it checks whether the message routing can be accomplished by its cache information. If there is a valid location information for the MN, the anchor SIP server directly routes the received INVITE message to the found location (phase II, Fig. 3b). Otherwise, a typical P2P lookup procedure is triggered. To improve the TTC performance, efficient cache management is necessary and cache management schemes adaptive to mobility and cache size can be adopted, which will be our future work.
SIMULATION RESULTS

To highlight the effectiveness of SAMP, we have developed an event-driven simulator and carried out comprehensive simulations. The simulation network has a grid topology consisting of 150 × 150 access routers (ARs). One HA and 100 MAPs (or SIP servers) are deployed over the simulation topology. Each MAP (or SIP server) covers 15 × 15 ARs in its domain. The number of MNs is 10,000 and their initial locations are uniformly selected in the simulation topology.

To investigate load balancing between mobility agents, we define a hotspot parameter \( \alpha \), where \( 0 \leq \alpha \leq 1 \). The hotspot parameter is assigned to a specific AR, and an MN conducts 200 movements. At each movement, each MN moves towards the hotspot AR with the probability \( \alpha \) (i.e., towards a neighbor AR nearest to the hotspot AR), whereas it moves to a randomly chosen neighbor AR with the probability of \( 1 - \alpha \) by the random walk mobility model.

Figure 4 shows the standard deviation (\( \sigma \)) of the number of MNs serviced by the mobility agents. In MIPv6, since there exists only the HA, all MNs are served by the HA. Hence, \( \sigma \) is the highest and the value is constant regardless of \( \alpha \). For HMIPv6, multiple MAPs act as local HAs, so that the overall load is distributed to multiple MAPs. Therefore, HMIPv6 shows less \( \sigma \) than MIPv6. However, as \( \alpha \) increases, more MNs move to the hotspot AR and therefore \( \sigma \) significantly increases. On the other hand, SIP servers in SAMP organize a P2P overlay network and the MN's home SIP server is determined evenly by the P2P routing algorithm. Consequently, the effect of \( \alpha \) in SAMP is not significant compared to HMIPv6.

Performance improvements via optimization techniques are indicated in Fig. 5. As shown in Fig. 5a, HMIPv6 significantly reduces the handoff latency incurred in MIPv6. This is because HMIPv6 introduces the MAP to localize the location registration procedure. When the HR scheme is not applied, SAMP exhibits the highest handoff latency. In this case all location registration messages should be delivered to the home SIP server. Furthermore, the routing of registration messages is performed over the P2P overlay network. Hence, a long handoff latency is observed. However, the increased handoff latency can be significantly reduced by adopting the HR scheme.

Figure 5b shows the average hop distance traveled by the first data packet of a session delivered from a CN to an MN. In MIPv6 and HMIPv6, the packet's delivery paths are CN → HA → MN and CN → HA → MAP → MN, respectively. On the other hand, SAMP needs additional hops for session establishment. It can...
be seen that SAMP without TTC has the longest average hop distance for the packet delivery because an INVITE message for session establishment should be delivered by the P2P routing algorithm. However, when the TTC scheme is used and a cache hit occurs, the CN can send an INVITE message to the MN without the P2P overlay routing. Consequently, SAMP with TTC can reduce the session-setup latency, and therefore shows a comparable hop distance for packet delivery to HMIPv6.

**RELATED WORK**

Recently, P2P overlay networking has been adopted for mobility support in several projects. Zhuang *et al.* proposed a mobility-support framework called Robust Overlay Architecture for Mobility (ROAM) [9]. ROAM is built on top of the Internet Indirection Infrastructure (i3). With i3, instead of explicitly sending a packet to a destination, each packet is associated with an identifier of the destination, which defines an indirection point in i3 and is also used to forward the packet. Via P2P overlay networking, ROAM provides robust and load-balanced mobility services. However, since there is no distinction in handling mobility-related signaling packets and data packets, all data packets should be indirected by i3, and therefore it induces significant packet-delivery latency.

Guo *et al.* proposed an end-system-based mobility solution for IPv6 (EMIPv6) [10]. For connection maintenance, EMIPv6 introduces an end-to-end approach to directly exchange mobility messages between two nodes. If the mobility messages cannot be directly delivered due to firewall or simultaneous movement, a distributed subscription/notification service built on top of a P2P overlay is utilized. In addition, for location management, an extended P2P overlay is used to locate an MN efficiently. However, EMIPv6 requires some modifications to the existing IPv6 stack (i.e., additional extension headers for mobility management are needed). Furthermore, since EMIPv6 does not support micromobility, a long handoff latency for frequent handoffs is expected.

Zhao *et al.* introduced another overlay-based mobility solution called Wrap [11], in which the problem of mobile crowds that generate storms of location update traffic was investigated. To address this problem, a novel aggregation technique that allows mobile crowds to roam as a single mobile entity using a structured P2P routing mesh architecture was developed. However, Wrap focuses on the efficient handling of handoff storms without addressing scalability in mobile Internet.

Mao *et al.* proposed a new approach called Distributed Home Agent for Robust Mobile Access (DHARMA) for intermittent connectivity and routing for mobile systems [12]. Specifically, a unified session-based mobility architecture for both end-to-end and proxy modes is designed. Also, a dynamic HA selection mechanism based on P2P overlay networking is employed to improve the routing performance. However, DHARMA has several limitations as a general mobility-support proto-

![Figure 5. Performance enhancements via optimization techniques: a) average hop distance for location registration; b) average hop distance for packet delivery.](image-url)
Lastly, since ROAM does not separate signaling and data deliveries, it may result in inefficient packet delivery, especially for long-lived sessions.

**CONCLUSION**

In this article we have proposed a novel mobility-support protocol called SAMP, which is based on P2P overlay networking. Due to the inherent advantages of P2P overlay networking, SAMP is highly scalable compared to MIPv6 and HMIPv6. Moreover, since SAMP is based on an Internet standard protocol (SIP), SAMP can be easily implemented and extended from the existing infrastructures. Two optimization techniques, hierarchical registration and two-tier caching schemes, efficiently mitigate the performance problems such as long location registration and session-setup latencies. Consequently, SAMP is expected to play a key role in proliferating the mobile Internet as a scalable mobility solution.

**ACKNOWLEDGMENT**

This work was supported in part by the Brain Korea 21 project of the Ministry of Education and in part by the Korean Research Foundation under grant no. M10-000-10073-0, 2005, Korea.

**REFERENCES**


**BIographies**

SANGHEON PACK (M) (spack@bbcr.uwaterloo.ca) received B.S. and Ph.D. degrees in 2000 and 2005 from Seoul National University, both in computer engineering, respectively. He is a post-doctoral fellow in the Centre for Wireless Communications, Department of Electrical and Computer Engineering, University of Waterloo. During 2002–2005 he was a recipient of the Korea Foundation for Advanced Studies (KFAS) Computer Science and Information Technology Scholarship. He has also been a member of Samsung Frontier Membership (SFM) from 1999. He received a student travel grant award for the IFIP Personal Wireless Conference 2003. He was a visiting researcher at Fraunhofer FOKUS, Germany, in 2003. His research interests include mobility management and multimedia transmission in next-generation wireless/mobile networks.

KUNWOO PARK (kwpark@mmlab.snu.ac.kr) received his B.S. degree in computer engineering from Korea Advanced Institute of Science (KAIST) in 2004. Currently he is working toward a Ph.D. degree at the School of Computer Science and Engineering, Seoul National University. His research interests include P2P networking, IPv6, and mobility management.

TAEKYOUNG KWON (tkkwon@snu.ac.kr) is an assistant professor in the Multimedia and Mobile Communications Laboratory, School of Computer Science and Engineering, Seoul National University. He received his B.S., M.S., and Ph.D. degrees in computer engineering from Seoul National University in 1993, 1995, and 2000, respectively. He was a visiting student at IBM T. J. Watson Research Center in 1998 and a visiting scholar at the University of North Texas in 1999. His recent research areas include radio resource management, wireless technology convergence, mobility management, and sensor networks.

YANGHEE CHOI [SM] (yhchoi@snu.ac.kr) received his B.S. degree in electronics engineering from Seoul National University, his M.S. degree in electrical engineering from KAIST, and his D.Eng. degree in computer science from Ecole Nationale Superieure des Telecommunications, Paris, France, in 1975, 1977, and 1984, respectively. Before joining the School of Computer Engineering, Seoul National University in 1991, he was with the Electronics and Telecommunications Research Institute during 1977–1991, where he served as director of the Data Communication Section and the Protocol Engineering Center. He was president of the Open Systems and Internet Association of Korea. He is now leading the Multimedia and Mobile Communications Laboratory at Seoul National University. His research interests lie in the field of multimedia systems and high-speed networking.

---

<table>
<thead>
<tr>
<th>Design consideration</th>
<th>SAMP</th>
<th>ROAM</th>
<th>EMIPv6</th>
<th>Wrap</th>
<th>DHARMA</th>
<th>DNSP2P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalability</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Localized mobility management</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Signaling/data separation</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Table 1. Comparison of P2P-based mobility solutions: yes (Y) or no (N).**

Due to the inherent advantages of P2P overlay networking, SAMP is highly scalable compared to MIPv6 and HMIPv6. Moreover, since SAMP is based on an Internet standard protocol (SIP), SAMP can be easily implemented and extended from the existing infrastructures.