Distributed Mobility Management Based on Flat Network Architecture

Liu Yu, Zhao Zhijun, Lin Tao, Tang Hui
High performance network laboratory, Institute of Acoustics, Chinese Academy of Sciences,
NO.21, Bei-Si-huan-Xi Road, Beijing, China, 100190,
Email: liuy@hpnl.ac.cn

***Abstract*—Mobility management is crucial for mobile networks to support certain quality of service (QoS) requirements. In this paper, we developed a flat network architecture rather than traditional two-layer mobility management architecture to reduce network elements and lighten networking load. Based on the proposed flat network structure the concept of separation of location and identity (Loc/ID), which comes from "clean-slate" design idea, was used in the mobility management scheme. To maximally reduce the handoff latency, bicasting mechanism was adopted and the consumption of bicasting was limited locally rather than globally. As analyzed in this paper, the bicasting scheme forms a lower bound for handoff latency in single network interface configuration case. To evaluate the performance of the proposed mobility management system analytical model was established by which the signaling cost, handoff delay and packet loss were evaluated. Compared with other fast handoff approach, we showed that our handoff approach is better in terms of signaling cost, handoff latency and bicasting cost.**

I. Introduction

As mobile users and mobility applications increasing, mobility management efficiency and performances become more and more important due to increasing mobility traffic and mobile users. Mobility management includes two aspects: location management and handoff management [1]. Location management allows network to locate a mobile terminal when there is a call to the mobile terminal while handoff management tries to keep a user's connection when it moves from an attachment point to another. The object of mobility management is to design network architecture and mobility management scheme to implement it at a lower cost and high efficiency.

There were various mobility management schemes in previous researches. Mobile IP series tried to solve this problem in network layer including Mobile IPv6 (MIPv6) [2], Fast Handovers for MIPv6 (FMIPv6) [3], Hierarchical MIPv6 (HMIPv6) [4]. These schemes had the two-layer structure, home agent (HA) and foreign agent (FA). In FMIPv6 the link layer information was used to generate trigger to layer 3 to accelerate mobility detection. Paper [5] analyzed and compared these schemes. [6] gave a distributed dynamic regional location management scheme in which the gateway foreign agent was dynamically selected to maintain signaling burden balance among FAs and most of the signaling traffic was restricted locally. A dynamic hierarchical mobility management scheme was developed in [7] in which the signaling burden was evenly distributed in the network and the authors gave an algorithm to determine optimal FA chain length. Mobility management solutions on higher layers include TCP-Migrate [8], MSOCKS [9], Session Initiation Protocol (SIP) [10]. These solutions were all end to end mode since mobility management protocols being implemented on end terminals the basic idea of which can be found in [11]. Others, such as IDMP [12], Cellular IP [13], and HAWAII [14], were all micro-mobility solutions to reduce global location update signaling cost. In [15] the authors proposed a scheme of distributed HA to avoid inefficient routing.

The solutions above were all based on existing network structures and existing Internet name space — IP space. Current mobility management network architecture is mainly HA-FA two-level hierarchy. However, Future Internet Network Design (FIND) and Global Environment for Network Innovations (GENI) have initiated new projects to change the current Internet network architecture to adapt future needs. Flat network architecture is one of the trends and distributed management is one of the future management schemes. In this paper we propose a new flat network architecture as an substitution of the present two-level structure in which a new network body — access gateway (AGW) is introduced to implement mobility management functionalities. Inspired by the clean-slate idea in [16], host ID is developed and Loc/ID separation is used in our mobility management design. The advantages of these changes include: a) flat network architecture reduces the number of network elements and the processing cost compared to hierarchical architecture which results in overall efficient management. b) Loc/ID separation relieve the overloading IP name space which make mobility management more flexible and efficient. In the meantime, the bicasting scheme is employed to further improve the handoff performance. In [17], bicasting has been used in SIP based scheme, but the bicast cost is high for its global manner. The bicasting in this paper is local manner which limits the bicasting locally to minimize cost. The handoff latency here forms a lower bound. The proposed mobility management scheme is called distributed mobility management (DMM) for the information of Loc/ID is stored and looked up in a distributed manner.

The rest of this paper is organized as follows. The new flat network architecture is presented in section II. Based on the flat architecture and combined with Loc/ID separation the distributed mobility management scheme is developed in

Digital Object Identifier: 10.4108/ICST.WICON2010.8728
http://dx.doi.org/10.4108/ICST.WICON2010.8728

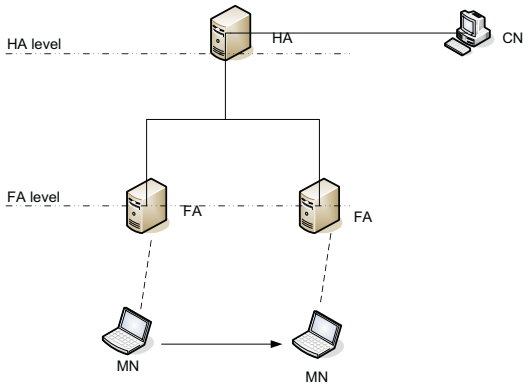

Fig. 1. Two-level network architecture.

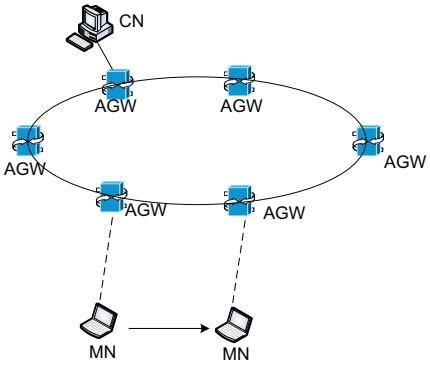

Fig. 2. Flat network architecture.

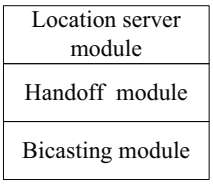

Fig. 3. AGW components.

section III. Analytical model is established in section IV for performance evaluation. In section V the performance is evaluated and analyzed. The last section is the conclusion part.

II. Flat Network Architecture

To present the flat network architecture, first the two-level architecture is illustrated as in Fig. 1 which is also used in MIP series. We briefly state the mobility management procedure in this architecture here:

- Mobility agents (i.e., FA and HA) advertise their presence by sending Agent Advertisement messages.
- A mobile node (MN) determines whether it is on its home network or a foreign network via these Agent Advertisement messages.
- When MN finds that it is in a foreign network, it obtains a care-of address on the foreign network and initiate a register procedure with HA. MN sends register message to HA which is relayed by FA.
- HA receives this register message and sets a binding item, i.e. an home address and care-of address pair, for this MN. HA sends register reply message to MN.
- When HA receives data packet destined to this MN, it intercepts the data packet and tunnels the data packet to the corresponding care-of address. Then the MN can get the data packet for it.

The whole procedure needs HA and FA two levels participating which increases signaling cost, processing cost and latency.

Flat network architecture is proposed to reduce the network architecture levels from two to one to reduce the above mentioned disadvantages. Instead of HA-FA two-level architecture, we use a single AGW level architecture as shown in Fig. 2. The whole network consists of two kinds of elements, AGW on the network side and terminals including corresponding node (CN) and MN. Loc/ID separation is adopted in this architecture which can be explained as: every MN has a globally unique ID keeping unchanged and when MN changes its location, i.e. MN moves to a new network, it merely changes its IP which represents the current location of a MN. Each AGW manages a AGW domain which can be understood as forming a network. MN moving between AGWs crosses different networks. Every AGW has a location server module inside it which provides location functionality and the location servers on different AGW are organized as distributed hash table (DHT) manner. Each location server stores (ID,IP) pairs where ID is the unique ID of a MN and IP is the current IP that MN uses. AGW supports bicasting function which will be explained in detail in next section.

Every CN and MN has a handoff module to tackle handoff task. The mobility management related function modules in AGW are shown in Fig. 3. The function of AGW contains:

- Provide location service, i.e. find the current IP of a given ID,
- Provide handoff support for mobile terminals,
- Provide ID bicasting, i.e. bicast a data packet for an ID as needed.

III. Distributed Mobility Management Scheme

In this section, we provide the details of distributed mobility management scheme.

As described above, each MN has a unique ID keeping unchanged and a current using IP. The (ID,IP) pair of each MN is stored in the location servers inside AGW which are organized as DHT manner. When MN moves to a new network it changes its IP according to the new network and updates the (ID,IP) pair in the DHT to make sure that the MN owning the ID can be achieved by looking up the IP in the (ID,IP) pair. When an CN wants to connect a MN, the calling procedure is:

Digital Object Identifier: 10.4108/ICST.WICON2010.8728
http://dx.doi.org/10.4108/ICST.WICON2010.8728

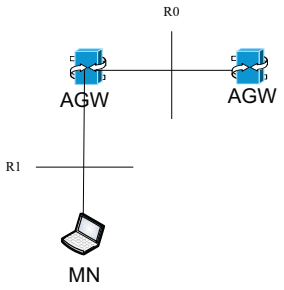

Fig. 4. Interfaces.

1) CN looks up current IP of the called MN using MN's ID as index in location server DHT and gets the corresponding (ID,IP) pair
2) CN connects the MN using the IP address gotten from step 1
3) MN responses the connecting request as it receives the connecting request which complete the calling procedure.

Interfaces between AGWs and between AGW and MN are shown in Fig. 4. R0 is the interface via which information is exchanged between AGWs while R1 is the interface that is used to exchange information between AGW and MN. The whole handoff procedure can be divided into two phases, handoff preparation phase and handoff execution phase. We make use of link layer information to aid handoff procedure. When link layer detects that it has come to handoff region it sends trigger message to upper layer to initiate handoff preparation phase. The handoff module on MN sends handoff preparation message to its serving AGW as it receives the handoff trigger from link layer. Upon receiving handoff preparation message from MN the handoff module of AGW implements following things: 1) exchange messages with its neighboring AGWs to determine which is the target AGW to handoff to, 2) transfer the MN related context with target AGW, 3) acquire available IP from target AGW and relay it to MN, 4) start bicast service for MN. So far, MN has obtained available IP at targeting AGW and it can handoff to the new AGW now. After handoff to new AGW, MN needs to update its (ID, IP) pair stored in location server DHT. MN notifies CN of changing IP and CN sends data packet to the new IP but the same ID. Upon receiving data packet directly from CN, new AGW sends stop bicasting message to previous AGW to terminate bicasting and previous AGW reclaims all resources occupied by the MN. Bicasting means when the serving AGW receives data packet from CN it replicates the data packet and modifies its destination IP to the available IP allocated by target AGW for MN using, then sends the replicative data packet to target AGW. So there are two data packets simultaneously in space which have same data, same destination ID but different destination IP. The sequence of messages of handoff procedures is illustrated in Fig. 5. The performance of the proposed mobility management scheme

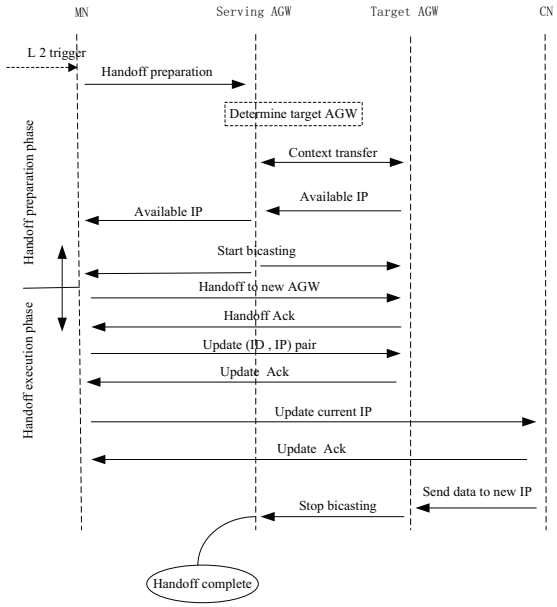

Fig. 5. Signaling message sequence.

will be evaluated in following sections.

IV. Analytical Models

Analytical model is a useful approach to evaluate system performance. For analyzing system performance mobility and traffic model are crucial factors. As in [18], MN mobility can be modeled via the cell residence time and we replace cell residence time by AGW residence time in this paper. Assume that arriving call to MN follows Poisson process. Although the inter-call time may not be exponentially distributed due to busy line effect, we use exponential distribution here for simplicity and it has good approximation. Let $\alpha(i)$ represent the probability that MN crosses i AGWs between two consecutive calls. The average number of AGW crossings of MN between two consecutive calls is,

$$E(N_m) = \sum_{i=0}^{\infty} i\alpha(i) \tag{1}$$

Where N_m represents the number of AGW crossing between two consecutive calls. To derive $\alpha(i)$, assume the coverage area of AGW is circle, as in [19], the AGW crossing rate of an MN is,

$$\mu_c = 2\frac{v}{\sqrt{\pi a_{AGW}}} \tag{2}$$

Where v is the average velocity of an MN, $a_{AGW} = \pi R^2$ and R is referred as the coverage radius of an AGW. Note that the coverage area of AGW is a logical concept which implies an AGW domain compared to cell coverage area in wireless cellular networks.

The notations we use in the followings are illustrated in Table. I. Fig. 6 shows the time diagram for an MN crossing an

Digital Object Identifier: 10.4108/ICST.WICON2010.8728
http://dx.doi.org/10.4108/ICST.WICON2010.8728

TABLE I
NOTATION.

t_c	inter-call time between two consecutive calls
$t_i(t_j)$	residence time in an AGW domain
t_r	residual residence time in an AGW domain
f_r	probability density function (pdf) of AGW residence time

AGW domain boundary and moves to another AGW domain during inter-call time. As in [5] [20],The probability of AGW domain crossing during inter-call time is,

$$P_c = Pr(t_c > t_i) = \int_0^\infty Pr(t_c > r) f_r(r) dr \quad (3)$$

The probability that MN crosses i AGWs during inter-call time is,

$$\alpha(i) = P_c^i (1 - P_c) \quad (4)$$

With the assumption that the AGW domain residence time is exponentially distributed with parameter μ_c and Poisson call arrival process with parameter λ_s, we can get,

$$P_c = \frac{\mu_c}{\mu_c + \lambda_s} \quad (5)$$

Combining (1), (4), (5), the average number of AGW crossing during inter-call time can be obtained under the exponential assumptions,

$$E(N_m) = \frac{\mu_c}{\lambda_s} \quad (6)$$

Call-to-mobility ratio (CMR) is a significant value in mobility management system performance evaluation and is defined as the relative ratio of call arrival rate to MN mobility rate. If data packets arrive at an MN at rate λ and the mean residence time of MN in an AGW domain is $1/\mu$, then CMR is given as,

$$\rho = \frac{\lambda}{\mu} \quad (7)$$

A. Signaling cost

Signaling cost is an important aspect of system performance. In the distributed mobility management system, when MN crosses an AGW domain and goes to another AGW domain the location server DHT and CN need to be updated. These updates are similar with the binding update in MIP schemes. For unifying expression reason we name location server and CN update in our scheme binding update too. We focus on the mobility incurred signaling performance, since the signaling costs due to binding time expiration are approximately the same for mobility management schemes.

The notations used in the expressions of signaling cost are listed in Table. II. For comparison we first give the expression of update signaling cost of FMIP.

$$\begin{aligned} C_{FMIP} &= E(N_m) \cdot [(6L_{wl} + 3L_w + 5P_{FA}) \\ &\quad + (2L_{wl} + 2d_{MN,HA} L_w + P_{HA}) \\ &\quad + (2L_{wl} + 2d_{MN,CN} L_w + P_{CN})] \end{aligned} \quad (8)$$

TABLE II
NOTATION.

L_{wl}	the link cost for relaying signaling message in a wireless link
L_w	the link cost for relaying signaling message in a wired link
P_{FA}	the processing cost of signaling message in FA
P_{HA}	the processing cost of signaling message in HA
P_{CN}	the processing cost of signaling message in CN
P_{AGW}	the processing cost of signaling message in AGW
$d_{MN,HA}$	the number of hops between MN and HA (exclude the last wireless hop to MN)
$d_{MN,CN}$	the number of hops between MN and CN (exclude the last wireless hop to MN)

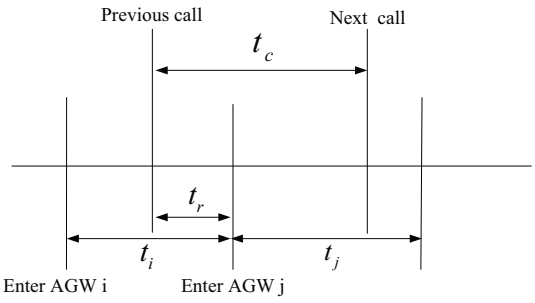

Fig. 6. The time diagram for AGW boundary crossing.

The update signaling cost of our distributed mobility management, referred to the sequence of message of our handoff procedure in Fig. 5, is,

$$\begin{aligned} C_{DMM} &= E(N_m) \cdot [(4L_{wl} + 3L_w + 4P_{AGW}) + \\ &\quad + (2L_{wl} + P_{AGW}) \\ &\quad + (2L_{wl} + 2d_{MN,CN} L_w + P_{CN})] \end{aligned} \quad (9)$$

When the distribution is assumed to be exponential, from (6)(7) we can rewrite above equations via CMR as,

$$\begin{aligned} C_{FMIP} &= \frac{1}{\rho} \cdot [(6L_{wl} + 3L_w + 5P_{FA}) \\ &\quad + (2L_{wl} + 2d_{MN,HA} L_w + P_{HA}) \\ &\quad + (2L_{wl} + 2d_{MN,CN} L_w + P_{CN})] \end{aligned} \quad (10)$$

$$\begin{aligned} C_{DMM} &= \frac{1}{\rho} \cdot [(4L_{wl} + 3L_w + 4P_{AGW}) + \\ &\quad + (2L_{wl} + P_{AGW}) \\ &\quad + (2L_{wl} + 2d_{MN,CN} L_w + P_{CN})] \end{aligned} \quad (11)$$

These are the analytical formulas of binding update signaling cost.

V. PERFORMANCE EVALUATION AND ANALYSIS

A. Update signaling cost

The topology used to compute numerical results is shown in Fig. 7. Set the analytical system parameters as: $L_{wl} = 8, L_w = 1$ and all the processing cost $P_X = 12$ for comparing reason.

Digital Object Identifier: 10.4108/ICST.WICON2010.8728
http://dx.doi.org/10.4108/ICST.WICON2010.8728

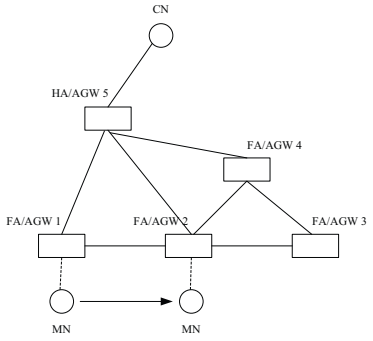

Fig. 7. Network topology.

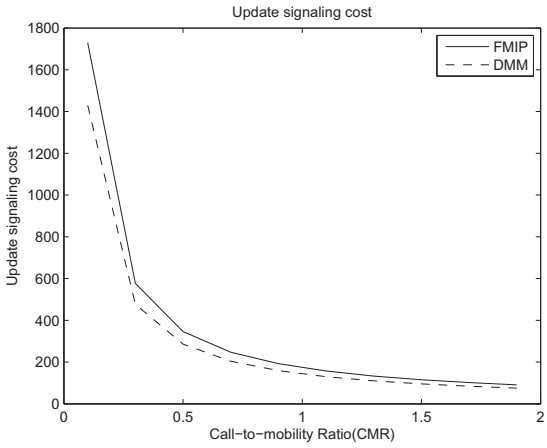

Fig. 8. Update signaling cost.

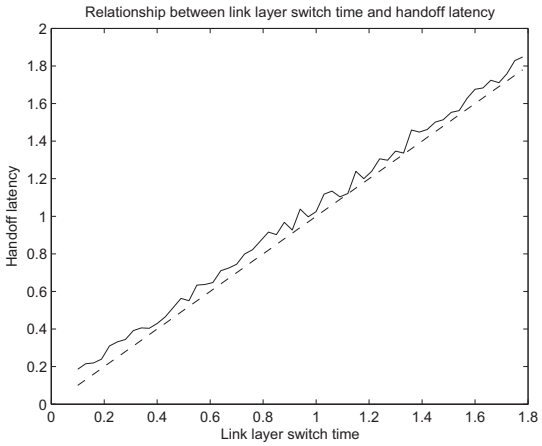

Fig. 9. Handoff latency and link layer switch.

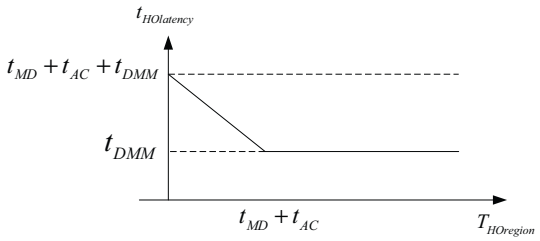

Fig. 10. Handoff latency.

The update signaling costs of FMIP and DMM are depicted in Fig. 8.

We indicate that the difference between FMIP and DMM actually reflects the difference between flat architecture and hierarchical architecture. The advantage of the flat network architecture can be seen.

B. Handoff latency

If the handoff preparation phase succeed, the handoff latency is just the linker layer switch time,

$$t_{HOlatency} = t_{l2switch} \tag{12}$$

This is defined as DMM handoff latency, $t_{DMM} = t_{l2switch}$. In the handoff preparation failure case the whole three handoff phases, movement detection, address configuration and handoff execution must all be implemented and the handoff latency is:

$$t_{HOlatency} = t_{MD} + t_{AC} + t_{DMM} \tag{13}$$

Where t_{MD} is movement detection time, t_{AC} is address configuration time and t_{DMM} is handoff latency of DMM in handoff preparation success case. In the handoff preparation phase failure case the handoff procedure is reduced to the procedure similar with MIP which experience the whole mobility detection, IP address configuration and binding update procedures resulting in large handoff latency. A primary reason that results in handoff preparation phase failure is that the residence time in handoff region is too short to complete the handoff preparation phase. The relation can be represented as,

$$T_{HOregion} < t_{MD} + t_{AC} \tag{14}$$

Where $T_{HOregion}$ is the handoff region residence time. In simulation we find that in the case of handoff preparation success the handoff latency is approximately the link layer switch time as shown in Fig. 9. The whole handoff latency curve is similar with the one in [17], shown in Fig. 10

If MN is equipped with one network interface, the handoff latency here forms the lower bound of handoff latency which merely depends on the link layer switch time. If MN is equipped with two network interfaces, MN can start up both of the two interfaces in handoff period which results in zero handoff latency.

C. Packet Loss

Based on the analysis results of handoff latency, packet loss is easy to analyze. We divide packet loss into two classes.

Digital Object Identifier: 10.4108/ICST.WICON2010.8728
http://dx.doi.org/10.4108/ICST.WICON2010.8728

1) When AGW is not configured with bicasting packet buffer, the packet loss number can be presented as,

$$N_{loss} = \lambda_p \cdot t_{DMM} \tag{15}$$

2) When AGW has a buffer with capacity C_b, the packet loss number is,

$$N_{loss} = \lambda_p \cdot t_{DMM} - C_b \tag{16}$$

Where λ_p is the packet arriving rate. The buffer size also can be configured according to the dynamic traffic arriving rate which is not discussed in this paper.

D. Bicasting packets delivery cost

The bicasting packet delivery cost in this paper is limited locally rather than globally in [17]. In [17] the bicasting packets are generated and delivered from CN to MN, however in our DMM scheme the bicasting packets are generated in the serving AGW and delivered from AGW to MN which limit the bicasting process locally and the bicasting packet delivery cost is reduced.

VI. Conclusion

In this paper a new distributed mobility management scheme has been proposed. The flat network architecture is developed to replace the previous hierarchical network architecture which reduces the number of network layers and the signaling cost. In the light of clean-slate idea, Loc/ID separation is employed to relive the overriding IP address load. Bicasting scheme is used to reduce handoff latency and is limited locally. From the analysis, our DMM scheme is demonstrated superior to the previous ones which develop a new way to mobility management design.

References

[1] I. F. Akyildiz et al., "Mobility Management in Next-Generation Wireless Systems," Proc. IEEE, vol. 87, no. 8, Aug. 1999, pp. 1347-84.

[2] D. B. Johnson, C. E. Perkins, and J. Arkko, "Mobility support in IPv6," IEFT RFC 3775, June 2004.

[3] G. Koodli, "Fast handovers for mobile IPv6," IETF RFC 4068, July 2005.

[4] H. Soliman, C. Castelluccia, K. El-Malki, and L. Bellier, "Hierarchical mobile IPv6 mobility management (HMIPv6)," IETF RFC 4140, Aug. 2005.

[5] Christian Makaya and Samuel Pierre, "An Analytical Framework for Performance Evaluation of IPv6-Based Mobility Management Protocols", IEEE Trans. Wireless Commun., vol. 7, no. 3, Mar. 2008, pp. 972-983.

[6] J. Xie and I. F. Akyildiz, "A distributed dynamic regional location management scheme for mobile IP," in Proc. IEEE INFOCOM, 2002, pp. 1069-1078.

[7] Wenchao Ma and Yuguang Fang, "Dynamic Hierarchical Mobility Management Strategy for Mobile IP Networks", IEEE Journal on Selected Areas in Communications (JSAC), vol. 22, no. 4, May 2004, pp. 664-676.

[8] A.C. Snoeren and H. Balakrishnan, "An End-to-End Approach to Host Mobility," Proc. MobiCom, Aug. 2000.

[9] D. Maltz and P. BhaGWat, "MSOCKS: An Architecture for Transport Layer Mobility," Proc. INFOCOM, pp. 1037-1045, Mar. 1998.

[10] E. Wedlund and H. Schulzrinne, "Mobility Support Using SIP," Proc. Second ACM/IEEE Intl Conf. Wireless and Mobile Multimedia (WoWMoM99), Aug. 1999.

[11] SALTZER, J. H., REED, D. P., AND CLARK, D. D. "End-to-end arguments in system design". ACM TOCS 2, 4 (Nov. 1984), pp. 277 - 288.

[12] A. Misra et al., "IDMP-Based Fast Handoffs and Paging in IP-Based 4G Mobile Networks," IEEE Commun. Mag., Mar. 2002, pp. 138 - 45.

[13] A. T. Campbell et al.,"Design, Implementation, and Evaluation of Cellular IP," IEEE Pers. Commun., Aug. 2000, pp. 42 - 49.

[14] R. Ramjee et al., "HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-Area Wireless Networks," IEEE/ACM Trans. Net., vol. 10, no. 3, June 2002, pp. 396 - 410.

[15] Yun Mao, Bjorn Knutsson, Honghui Lu, and Jonathan Smith, "DHARMA: Distributed Home Agent for Robust Mobile Access," in Proceedings of the IEEE INFOCOM 2005 Conference, Miami, March 2005.

[16] Steven M. Bellovin, David D. Clark, Adrian Perrig, and Dawn Song, "A Clean-Slate Design for the Next-Generation Secure Internet," Report of an NSF workshop held at CMU, 2005.

[17] Seok Joo Koh and Wook Hyun,"mSIP: Extension of SIP for Soft Handover with Bicasting," IEEE Commun. Lett., vol. 12, no. 7, Jul. 2008, pp. 532 - 534.

[18] Yuguang Fang, Imrich Chlamtac and Yi-Bing Lin, "Portable Movement Modeling for PCS Networks," IEEE Trans. Veh. Technol., Jul.2008, vol.35, no. 4, pp. 1356 - 1363.

[19] F. V. Baumann and I. G. Niemegeers, "An evaluation of location management procedures," in Proc. 3rd Annual Int. Conf. Universal Personal Commun. (UPC94), Sept./Oct. 1994, pp. 359 - 364.

[20] Y. Xiao, Y. Pan, and J. Lie, "Design and analysis of location management for 3G cellular networks," IEEE Trans. Parallel Distrib. Syst., April 2004, vol. 15, no. 4, pp. 339 - 349.

Digital Object Identifier: 10.4108/ICST.WICON2010.8728
http://dx.doi.org/10.4108/ICST.WICON2010.8728