DEMAPS: A Load-Transition-Based Mobility Management Scheme for an Efficient Selection of MAP in Mobile IPv6 Networks

Tarik Taleb, Member, IEEE, Abbas Jamalipour, Fellow, IEEE, Yoshiaki Nemoto, Senior Member, IEEE, and Nei Kato, Senior Member, IEEE

Abstract—One major concern for mobile networks consists of finding efficient ways of handling user mobility so that the handover process has minimum impact on users’ ongoing sessions. Given the dominance of Internet-based applications in next-generation mobile networks, Mobile IP (MIP) has become an important protocol for accommodating the IP mobility. To overcome the excessive delay and signaling involved in the first version of MIP, the Hierarchical Mobile IPv6 (HMIPv6) protocol has been introduced. The key concept behind HMIPv6 is to locally handle handovers by the usage of an entity called mobility anchor point (MAP). Although the new protocol provides a more efficient way for mobility management in IP networks, it does not control traffic among multiple MAPs in the network. Thus, in many cases, the selected MAP is overloaded, and extensive delays are experienced during the routing process. To tackle this problem, this paper proposes a new technique, i.e., dynamic efficient MAP selection (DEMAPS). The scheme works similar to HMIPv6 when the network is not overloaded. When the network gets heavy loads, the selection of MAPs is based on an estimation of MAP load transition using the exponential moving average method. Simulation results demonstrate that DEMAPS can efficiently balance the signaling traffic load among MAPs and provides a superior network performance compared with traditional HMIPv6 schemes.

Index Terms—Exponential moving average (EMA), Hierarchical Mobile IPv6 (HMIPv6), Mobile Internet Protocol (MIP), Mobile IPv6 (MIPv6), mobility anchor point (MAP), mobility management.

I. INTRODUCTION

Along with the exponential growth of the Internet and the continuous success of wireless communication networks, the telecommunications industry has entered a new era where communication needs are no longer limited to wired/wireless networks and have moved toward a new paradigm, i.e., pervasive access to Internet services anywhere at any time. Communication over mobile systems has, thus, been gaining ground at a tremendous pace during the last few years. Internet-based applications and the data traffic load that was generated from these applications have changed the path for the mobile network into an all–Internet Protocol (IP) configuration. Therefore, finding efficient optimum solutions for handling IP mobility has become an important topic of research.

As we have originally specified, the IP protocol does not support mobility due to issues with the protocol syntax and semantics. To support global mobility in IP networks, the Mobile IP Working Group within the Internet Engineering Task Force (IETF) proposed a packet-based mobility management protocol, i.e., the Mobile IPv6 Protocol (MIPv6) [3]. It has subsequently been modified, in line with the new version of IP, toward the so-called MIPv6 [4]. In MIPv6, mobile nodes (MNs) are identified with two different IP addresses: 1) the home address (HoA) and 2) the care of address (CoA). HoA indicates a unique permanent name of the MN. It remains constant, even when the node moves from subnet to subnet. CoA specifies a temporary address for the MN based on the current position of the node in the network and changes the address as the MN roams to a network other than its home network. The MN can obtain the CoA of the visiting subnet by issuing a router solicitation (RS) message to its foreign agent. The standard MIPv6 consists of two procedures: 1) binding update (BU) and 2) data delivery. The BU operation aims at associating the HoA and CoA of each MN.

It is well understood that the MIPv6 protocol does not represent an effective solution for environments in which MNs frequently change their points of attachment to the network. Applying MIPv6 to a large population of users with relatively high mobility features will (most likely) generate a large number of BU requests in a single burst. To process such bursts of BU requests, an important amount of network bandwidth and computational load is required. Thus, this BU cost becomes huge, and the system ultimately becomes unscalable to operate. In case of mobile users that roam far from their home networks, the system performance further gets aggravated, and the signaling delay for the BU increases. This results in the loss of a significant amount of in-flight packets, which eventually affects the overall quality of service (QoS) of the system. It is impossible to avoid such bursting occurrences of handovers in mobile networks, so the mitigation of BU frequency can only be done by developing new mobility management strategies.

To make the MIP scalable for a large highly mobile network and, at the same time, reduce the amount of signaling and the length of signaling paths, the signaling points can be arranged.
in a hierarchical pattern. Therefore, the IETF has proposed the Hierarchical Mobile IPv6 (HMIPv6) protocol [5], [6]. Fig. 1 depicts the key components of the HMIPv6 protocol. The key idea behind the HMIPv6 protocol is to separate local mobility from global mobility. HMIPv6 is based on the deployment of a number of local agents called mobility anchor points (MAPs). MAPs can be located at any level in a hierarchical network of routers. Each MAP administrates a set of access routers (ARs) that form a single network domain.

One important issue, which is highly missed in the design of HMIPv6, is on the traffic load and processing power distribution over multiple MAPs in a large mobile network. Indeed, in the case where a single MAP administers a large domain, the distance between MNs and the MAP can add significant delays to packets and affect route optimization as MNs move away from the MAP. Operators of large mobile networks may, thus, choose to deploy several MAPs in one domain. In such large networks with multiple MAPs, it is easily possible that some MAPs become congested, whereas other MAPs are underutilized. To cope with such an issue, which affects the overall network performance, a dynamic MAP management strategy is required. In this paper, a dynamic efficient method for selecting the most appropriate MAP is proposed. The operation of the proposed scheme consists of two steps. Initially, the proposed scheme behaves similar to the traditional distance-based MAP selection scheme of HMIPv6. In a given domain, MAPs with loads that are less than a predefined threshold are sorted. The furthest MAP (among these MAPs) is selected first as in the distance-based selection scheme of HMIPv6. When all MAPs have loads that exceed the threshold, the selection of MAPs becomes based on an estimation of MAP load transition using the exponential moving average (EMA) method. Distant MAPs with load decrease tendency are selected first. The proposed selection scheme is called dynamic efficient MAP selection (DEMAPS).

Extensive simulations are conducted to evaluate the performance of the proposed scheme. Simulation results demonstrate that the scheme reduces the number of packet drops, guarantees shorter service delays, makes better utilization of the network resources, and maintains a fair efficient distribution of the network load.

The remainder of this paper is structured as follows. Section II highlights the relevance of this paper to the state of the art in mobility management. The key design philosophy and distinct features that were incorporated in the proposed scheme are described in Section III. Section IV portrays the simulation environment and reports the simulation results. Section V discusses the applicability of the proposed scheme. This paper concludes in Section VI with a summary that recaps the main advantages and achievements of the proposed scheme.

II. RELATED WORK

Since its standardization, the MIP protocol has been the focus of extensive research work. This section describes the main poststandard improvements that have been devised in recent literature to improve the performance of MIP in mobile networks.

Mobility management techniques, in general, may be classified into two categories: 1) micro mobility and 2) macro mobility [7]. In micro mobility, handoffs are locally handled without involving any home agents (HAs). Notable examples are cellular IP [8], [9] and handoff-aware wireless access Internet infrastructure (HAWAII) [10]. Cellular IP is specifically designed to support handoff for frequently moving hosts. It is applied on a local level and can interwork with MIP to support wide area mobility, i.e., mobility among cellular IP networks. The HAWAII protocol divides the network into hierarchies based on domains. The functioning of HAWAII hinges on the assumption that users’ mobility is local to domains. For each host, the HA and any correspondent node (CN) are unaware of the node’s mobility within the host domain. Each domain has a gateway called the domain router, and each host has an IP address and a home domain. In HAWAII, host-based forwarding
entries are installed in gateways using a set of specialized path setup schemes. These entries help reduce both the data path disruptions and the number of BUs. One major credit for micromobility management techniques consists of their reduction in handoff signaling delays. A set of IP micromobility protocols and a comparison of their performances is presented in [11]. It also contains more detailed descriptions of the two mechanisms that we have presented.

In macro mobility, when an MN roams to a different network area, the node solicits for a new CoA. A BU message is, then, sent to the HA. The major issue with macro mobility pertains to the significant handoff signaling delays for users that roam far from their home networks. These delays disrupt active connections each time a handoff to a new attachment point of the network is performed. Therefore, the time that is required to establish a new connection between MNs and their CNs becomes remarkably long, and the loss of in-flight packets may become significant.

To cope with packet losses that may occur during handoffs due to the broken data path from the source to the destination, a set of mobility management techniques has been proposed in recent literature. They can be classified, in turn, into two categories: 1) caching-based techniques and 2) smooth handoff techniques. In the first category, when a handoff occurs, the old AR caches and forwards the packets to the new AR based on a request to forward the packets. The most pioneering example that uses this technique is the Fast Handovers for Mobile IP [12]. In the second category, packets are routed to multiple nearby ARs around the MN to ensure that the packets are delivered to the node. In addition to the recently proposed multipath smooth handoff scheme [13], multicast mobility support [14] and bicast that is used in Cellular IP [15] use this technique. One combination of the smooth handoff and buffering techniques is proposed in [16]. One detailed description of most of these techniques and comparison among their performances is given in [7].

To reduce handoff-signaling delays in macro mobility, a large body of prior work has been proposed. The central theme in these pioneering studies pertains to adopting hierarchical management strategies by using local agents. HMIPv6 and TeleMIP for Cellular IP [17] are notable examples. Most proposed protocols employ hierarchies to localize the binding traffic. Determining the optimal size of local networks is one of the most challenging tasks in hierarchical management procedures. To deal with this task, Xie et al. propose an analytic model based on the average total location update and packet delivery cost [18]. In [19] and [20], the decision on the optimal size of regional networks is based on mobility patterns, registration delays, and the central processing unit (CPU) processing overhead that is loaded on the local mobility agents. Although most hierarchical techniques are intended to reduce the BU traffic by localizing handoff signaling, they cause additional issues on network traffic management. Effectively, some local agents get congested with traffic, whereas other agents are not efficiently utilized. To overcome this deficiency, the choice of network hierarchies should be performed in a dynamic manner. In this regard, Pyo et al. propose a dynamic distributed domain-based mobility management scheme [21]. In this scheme, a group of ARs forms a domain. A domain list that indicates the ARs that belong to the same domain is stored at each AR. MNs that reside in a given domain maintain that domain list. If an MN changes its point of attachment to a new AR within a different domain, the node will, then, update its domain list to that of the new AR, and the new AR will serve as a MAP for the node. Ma et al. propose another dynamic hierarchical mobility management scheme for MIP networks [22]. In this scheme, when a mobile host connects to a new subnet via a new AR, the new AR notifies the new CoA of the host to the previous AR. The new AR then serves as a new location-management hierarchical level for the node. One major drawback of these two schemes is that they both deliver packets to users via multiple levels of ARs—a fact that leads to long packet delivery delay and congestion of the selected ARs with redundant traffic. One possible solution to this issue is to reduce the size of the subnet domains. However, this would lead to frequent interdomain handoffs and, consequently, excessive BU cost.

Another approach to solve the issue of traffic distribution in HMIPv6 is possible by referring to the mobility pattern of users [23], [24]. In [23], for instance, users are classified based on their velocity. Users receive thresholds from the network and compare their velocity to those thresholds. Users with velocities that exceed the propagated thresholds simply register with higher levels of the MAP hierarchies. Although this idea is straightforward, it does not solve the issues of traffic distribution among MAPs. Indeed, in the case where all users have the same feature of mobility, they end up registering with the same MAPs. This will intuitively overload the selected MAPs with traffic, whereas other MAPs remain underutilized. In addition, the velocity range for each MAP is fixed. To cope with this issue, Chung et al. [25] considered a dynamic setting of the velocity range of each MAP, depending on the actual velocities of MNs that are currently serviced by the MAP. However, a general requirement for mobility management schemes, which are based on the velocity of MNs, consists of guaranteeing high accuracy in estimating the velocity of MNs. Such a task is not always simple, which results, more frequently, in selecting inappropriate MAPs.

In [26], the moving range of an MN is the main factor in the MAP selection. In this scheme, MNs are assumed to keep track of their moving area. The lowest MAP that covers the entire moving area is considered the most appropriate MAP for registration. In this scheme, issues related to how the moving range of each MN can be defined, in addition to how the scheme can be applied to MNs that keep changing their moving areas, have yet to be solved. In [27], a distributed location-management scheme is proposed. The key idea behind this scheme is to allow an MN to roam over an area that is covered by a number of MAPs (among which the farthest MAP is distant from the MN by a threshold of hops) without triggering regional BUs. Although this operation achieves load balancing, to some extent, and reduces the BU cost, it comes at the price of longer delivery delays. In [28], a newly defined factor, i.e., the session-to-mobility ratio (SMR), is used as a factor for selecting the serving MAP. SMR is defined as the ratio of the session arrival rate to the handover frequency. In the SMR-based scheme, the highest MAP is selected for MNs with small values of SMR.
One interesting analysis among the aforementioned approaches can be found in [29].

III. DEMAPS Scheme

This section gives a detailed description of the proposed scheme, i.e., the DEMAPS scheme. As we have discussed, operators of large mobile networks may need to deploy several MAPs in one domain. In such networks, although most dynamic hierarchical mobility schemes reduce the frequency of BU messages to HAs, they do not efficiently balance the load over these multiple MAPs. Such a deficiency can overload some MAPs with traffic and cause higher packet delivery delays, whereas it leaves other MAPs underutilized. To tackle this issue, an efficient management strategy of MAP load is required. Based on this strategy, mobile users that reside in hierarchical mobile networks could select the most appropriate MAP for communication based on their current resources utilization. Fig. 2 depicts the major steps in the proposed MAP selection method.

The proposed scheme operates as follows. Similar to HMIPv6, DEMAPS adopts the dynamic MAP discovery approach. Each AR receives MAP option messages from high-layer MAPs every Δ period of time (see Step 1 in Fig. 2). Unless otherwise specified, Δ is set to 1 s. Using the information that is included in MAP option messages and based on a given computational model, each AR selects the optimum MAP (OMAP) for communication (see Step 2 in Fig. 2). Details on the computational method used will be given later in this section.

Upon performing handoff, an MN sends an RS message to the new AR (see Step 3 in Fig. 2). MNs can be designed not to submit RS messages in case ARs multicast router advertisement (RA) messages on a regular basis. In response to the RS message, the AR notifies the MN of the previously selected OMAP (see Step 4 in Fig. 2). It should be emphasized that an MN is notified of OMAP only when it changes its point of attachment to a new AR. This does not incur additional energy consumption (compared with HMIPv6) and does not have an effect on the critical battery life of MNs.

After receiving information on the OMAP, the MN makes comparisons between the selected OMAP and the previous MAP (PMAP) that is being used prior to handoff. If the OMAP is the same as the PMAP, the MN judges the handoff as an intradomain movement and sends a BU message to only the OMAP (see Step 5 in Fig. 2). This step aims at minimizing the handoff-signaling delay and reducing the signaling traffic for users that roam far from their home networks. In case the OMAP is different from the PMAP (e.g., interdomain handoff), the MN sends three BU messages to the OMAP, the HA, and its CN, respectively. In response to the BU message, the OMAP acknowledges the MN of a successful BU via a binding acknowledgment message (see Step 6 in Fig. 2).

As stated earlier, each access point receives MAP option messages from high-layer MAPs every Δ period of time. In DEMAPS, we consider the inclusion of information on instant loads of MAPs in the MAP option messages. Examples of parameters that can define a MAP load are memory size, CPU processing power, and used bandwidth. For simplicity, we denote the load of the ith MAP, as shown in its kth downstream node, at the nth time slot as $M_{i-k}[n]$ and define it as the integer part of the ratio of the number of packets that is processed on the link that connects the ith MAP to its kth downstream node to the total number of packets that the MAP can process on the same link during the computation period of time $\left[\Delta, (n + 1) \cdot \Delta\right]$, i.e.,

$$M_{i-k}[n] = \left\lfloor \frac{p_{i-k}[n] + W \cdot q_{i-k}[n]}{C_{i-k} \cdot \Delta} \cdot 100 \right\rfloor,$$  (1)

where $p_{i-k}[n]$ denotes the data processing speed of the ith MAP on the link to its kth downstream node. Note that a network element among the communication path can function as either a MAP or a mere router. The former case concerns packets that are destined to MNs and are registering with the network element as their MAP, whereas the latter case relates to packets that are destined to nodes with other network elements as MAPs. Similarly, $p_{i-k}[n]$ and $q_{i-k}[n]$ denote the total number of data packets that are forwarded by the ith MAP to its kth downstream node as a mere router and the number of data packets that are destined to MNs and are registered with the ith MAP, respectively, at the nth time slot. Intuitively, the computational load that is required by a mere router to forward a data packet and that required by a MAP to transmit a data packet to a node registered with it are different. $W$ is a weight factor for reflecting the difference in these two computational loads. It is assumed that access points have prior knowledge on the two parameters $C_{i-k}$ and $W$ for each ith MAP and for each respective kth downstream node. Upon computing their loads, MAPs notify access points of this information via the 7 bits of the reserved (RES) field carried in the packet header of MAP option messages, as we will explain later.

Having a potential number of MNs that is connected to the same MAP for communication may likely lead to congesting the MAP in question and result in an inefficient distribution of

---

1For each MAP option message that is destined to a downstream router, the included MAP load information refers to the load of the MAP, as shown in the downstream router.
To avoid congesting MAPs with traffic, ARs should advise newly arriving MNs with the most appropriate MAP. MAPs should, thus, be aware of the ongoing dynamics in network conditions and should reflect these dynamics in the signaling messages that they send to ARs. Based on these dynamics, ARs sort MAPs according to their availability, i.e., their loads with respect to the ARs, and advise newly arriving MNs with the most optimum MAP.

To notify ARs of possible changes in network conditions, MAPs use the EMA method to predict possible future transitions in their loads. The underlying reason beneath the choice of EMA consists of the EMA being a cut-and-dry approach for analyzing and predicting performance. The EMA is also easy to implement and requires minimal computational load. As we have mentioned earlier, the traffic load is periodically measured every $\Delta$ period of time in the proposed scheme. Let $M_{i \rightarrow k}[n]$ and $E_{i \rightarrow k}[n]$ denote the measured load value and the EMA value of the $i$th MAP load with respect to its $k$th downstream node at the $n$th time slot, respectively. By definition, $E_{i \rightarrow k}[n]$ is expressed as follows:

$$E_{i \rightarrow k}[n] = \frac{\sum_{k=0}^{\infty} ((1 - r)^k M_{i \rightarrow k}[n - k])}{\sum_{k=0}^{\infty} (1 - r)^k}$$

where $r$ is the exponential smoothing constant ($0 < r < 1$). Considering that $(\sum_{k=0}^{\infty} \theta^k = 1/(1 - \theta))$, $E_{i \rightarrow k}[n]$ can easily be computed in a recursive manner as follows:

$$E_{i \rightarrow k}[n] = r M_{i \rightarrow k}[n] + (1 - r)E_{i \rightarrow k}[n - 1]. \quad (2)$$

To give more weight to the latest data, $r$ is set to 0.9 throughout this paper.

The key idea behind the proposed method is to use the EMA value to predict the transition tendency of the MAP load on each available link. This prediction is based on the comparison between the two values $E_{i \rightarrow k}[n]$ and $M_{i \rightarrow k}[n]$ for each $i$th MAP and each respective $k$th downstream node. In case ($E_{i \rightarrow k}[n] < M_{i \rightarrow k}[n]$), the load of the $i$th MAP on the link to its $k$th downstream node has more tendency to increase [i.e., the load increase (LI) tendency], whereas in the case of ($E_{i \rightarrow k}[n] > M_{i \rightarrow k}[n]$), the MAP load on the same link may likely decrease [i.e., the load decrease (LD) tendency]. Upon predicting their load transitions, MAPs notify ARs of this information via the 7 bits of the RES field carried in the packet header of MAP option messages. We used 1 bit of the RES field as a flag to indicate the load transition tendency of the MAPs: 1 for LI and 0 for LD. The remaining 6 bits are used to indicate the MAP loads. A MAP $i$ sets the 6 bits of the RES field (of the MAP option message that is destined to its $k$th downstream node) to null if its load is smaller than 36%.\(^2\) Otherwise, it sets the 6 bits of the RES field to the integer part of the difference between the load and 36% as follows:

$$\text{RES}_{\text{val}} = \text{Max} \left(0, \lfloor M_{i \rightarrow k}[n] - 36 \rfloor \right). \quad (3)$$

Based on this information, ARs decide the most appropriate MAPs for future visiting mobile users. This operation is performed following two stages. When the network is not overloaded, the selection of MAPs is conducted based on distance as in HMIPv6. In a particular domain, MAPs with loads that are less than a predefined threshold $\beta$ are sorted. The parameter $\beta$ indicates the level of congestion that a network operator can tolerate. Unless otherwise specified, $\beta$ is set to 80% throughout this paper. The fastest MAP among the sorted MAPs is selected first, similar to the traditional distance-based selection scheme of HMIPv6. This operation is repeated until the loads of all MAPs exceed the threshold. At this stage, the selection of MAPs is based on the estimation of the MAP load transition using the EMA method. High-hierarchy MAPs with LD tendencies are preferably selected as MAPs for communications. In case of multiple MAPs with LD tendencies, the MAP router at the highest hierarchy is chosen. This aims at creating large MAP domains for MNs so that their future handoffs can locally be handled. This process ultimately minimizes the handoff signaling cost. In the case where all high-hierarchy routers have LI tendencies, ARs select the high-hierarchy MAP router with the minimum traffic load, i.e., the lowest value of $\text{RES}_{\text{val}}$.

IV. PERFORMANCE EVALUATION

A. Simulation Setup

Having described the details of the proposed scheme, the focus is now directed to its performance evaluation through extensive simulations using the QualNet 4.0 Simulator [30]. Particular attention is paid to the design of an accurate realistic simulation setup, as described in the following, which justifies the choices that were made along the way. Unless otherwise

\(^2\)Note that $36 = 100 - 2^6$. 

---

Fig. 3. Simulation environment.
noted, the parameters that are specified in the following are those used in all the experiments throughout this paper.

The abstract configuration of the considered network is depicted in Fig. 3. The wireless part of the network consists of four neighboring wireless cells. The coverage radius of each wireless cell is set to 75 m. The distance between two neighboring ARs is fixed to 100 m. These parameters are chosen without any specific purpose in mind and do not change any of the fundamental observations about the simulation results. The four ARs are connected to the wired network through a two-layer network that is made of four MAPs. To form cross links among the MAPs, MAPs 1 and 2 are both connected to MAPs 3 and 4. The choice of such a two-layer MAP (three-tier) network with cross links represents a general simple case [31]. Furthermore, in [32], a mathematical model demonstrated that a three-tier architecture represents the optimum hierarchy level for hierarchical mobile environments. In the considered topology, MAP 3 serves ARs 1 and 2, whereas MAP 4 serves ARs 3 and 4. The MAP network is connected to an HA and a server (i.e., a CN) via a wired network. The one-way propagation delay over the wired network is set to 10 ms. As for other links, the delay of each is set to 2 ms. In general scenarios, wireless links have smaller bandwidth compared to their wireline counterparts. In the simulations, the capacity of the wired network is set to 100 Mb/s, and the capacity of inter-MAP links, “MAP-to-AR” links, and wireless links is set to 20 Mb/s. Note that setting the bandwidth of wireless and wireline links to different rates should have no effect on the fundamental observations about the proposed scheme. In addition, simulating similar or disparate bandwidths for inter-MAP or “MAP-to-AR” links should not affect the system performance either, as the MAP selection in DEMAPS is based on the loads of MAPs rather than the links’ capacity.

A population of 50 nodes is simulated and randomly scattered over the wireless communication area. To avoid forming bottlenecks at the wireless network, each MN receives user datagram protocol (UDP) packets from the CN at a rate that is approximately equal to 390 kbps. The UDP packet size is set to 1 kB. Due mostly to its simplicity and its wide usage in today’s switches and routers, all routers use Drop-Tail as their packet-discarding policy. All MAPs are assumed to have buffers of 250 kB, i.e., an amount of data that is worth the bandwidth-delay product. The buffer size is chosen without specific commercialized standards in mind. For simplicity, the computational load that is required by a mere router to forward a data packet and that required by a MAP to transmit a data packet to a node that is registered with it are assumed to be the same. The parameter $W$ is, thus, set to one. To better investigate the interactions of the proposed scheme with different levels of the network congestion, we run some background traffic over inter-MAP links. Over each inter-MAP link, the rate of the background traffic is randomly chosen from within 40% to 70% of the link capacity. All simulations are run for a duration of 600 s, which is long enough to ensure that the system has reached a consistent behavior. The first 60 s are used to initialize the simulations, and the last 60 s are used to stabilize the results. All results are an average of multiple simulation runs. Table I shows a complete list of the simulation parameters.

| TABLE I |

<table>
<thead>
<tr>
<th>Factor</th>
<th>Simulation Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of mobile nodes</td>
<td>50</td>
</tr>
<tr>
<td>UDP traffic rate</td>
<td>390 kbps</td>
</tr>
<tr>
<td>Background traffic rate</td>
<td>8 Mbps - 14 Mbps</td>
</tr>
<tr>
<td>Simulation time</td>
<td>600 s</td>
</tr>
<tr>
<td>Wired network delay</td>
<td>10 ms</td>
</tr>
<tr>
<td>Wired network capacity</td>
<td>100 Mbps</td>
</tr>
<tr>
<td>Inter-MAP and MAP-AR link delay</td>
<td>2 ms</td>
</tr>
<tr>
<td>Wireless link capacity</td>
<td>20 Mbps</td>
</tr>
<tr>
<td>Buffer size</td>
<td>250 kB</td>
</tr>
<tr>
<td>Cell Coverage Radius</td>
<td>75 m</td>
</tr>
<tr>
<td>Distance between neighboring ARs</td>
<td>100 m</td>
</tr>
<tr>
<td>MAP option transmission period $\Delta$</td>
<td>1 s</td>
</tr>
<tr>
<td>Exponential smoothing constant $\alpha$</td>
<td>0.9</td>
</tr>
<tr>
<td>Packet size</td>
<td>1 kB</td>
</tr>
<tr>
<td>MAP load threshold $\beta$</td>
<td>80%</td>
</tr>
<tr>
<td>Parameter $W$</td>
<td>1</td>
</tr>
</tbody>
</table>

In the performance evaluation, we consider two mobility models that are available in QualNet: 1) the random waypoint model and 2) the group mobility model. In both models, the MN speed is set between the range [0 m/s, 2 m/s], and the pause time is set to null. In the group mobility model, four groups are simulated, with each consisting of 12 or 13 MNs. In the performance evaluation, HMIPv6 and the velocity-based MAP selection scheme that was proposed in [24], which is referred to as HMIPv6-UP, are used as comparison terms. In HMIPv6-UP, the velocity threshold is set to 0.5 m/s. In this scheme, if the velocity of an MN exceeds the velocity threshold, the MN registers with an upper MAP; otherwise, it registers with a lower MAP.

The following quantifying parameters are used for comparison.

- **Individual packet delivery delay.** This measure shows the average packet delivery delay that each simulated MN experiences.
- **Individual MN throughput.** This metric indicates the throughput that each MN achieves.
- **Processed load.** This measure involves the total bytes, including signaling packets and UDP data packets, that a specific network element processes during the simulation time.
- **Processed data packets.** This measure refers to the count of data packets that a specific network element processes during the simulation time.
- **Packet drops.** This indicates the total number of packets that a specific network element drops during the simulation time.

### B. Simulation Results

First, we investigate the load transitions of the four inter-MAP links in case of the two mobility models and when the three schemes are in use. Fig. 4 graphs the load transitions of the four links. It demonstrates that the proposed scheme enables a better distribution of traffic among all links, i.e., in the case of the two simulated mobility models. We next investigate the behavior of individual MNs. Given that the three schemes exhibit the same behavior, regardless of the simulated
Fig. 4. Load transitions of the four inter-MAP links in case of (upper row) the random waypoint mobility model and (lower row) the group mobility model. (a) HMIPv6. (b) HMIPv6-UP. (c) DEMAPS. (d) HMIPv6. (e) HMIPv6-UP. (f) DEMAPS.

Fig. 5. Behavior of individual MNs in terms of packet delivery delay and throughput. (a) Individual pocket delivery delay. (b) Individual throughout.

mobility model, and due to space limitations, we present the results of only the group mobility model. Fig. 5(a) and (b) graphs the individual delay and throughput that the simulated 50 nodes achieved. It shows that the improvements, in terms of increasing the throughput and reducing the end-to-end delay, that DEMAPS achieved are highly encouraging. The packet delivery delay was reduced for MNs in the case of DEMAPS. This performance is mostly attributed to the selection of the most appropriate MAP for communication. Effectively, in the case of HMIPv6, when MNs roam from one AR to another AR, for instance, each MN always selects the same MAP that was previously used, even if the MAP has heavy packet processing load. This event also occurs in the case of HMIPv6-UP when the MN’s mobility pattern does not change. This ultimately causes higher values of packet delivery delay. Fig. 5(b) indicates another achievement of the proposed scheme: DEMAPS significantly increases the throughput of each MN. The reason behind this performance lies beneath the fact that, in DEMAPS, MAPs with LD tendencies are preferably selected, whereas in HMIPv6 and HMIPv6-UP, in-flight packets travel over congested links, which overloads some MAPs, causes packet drops, and ultimately affects the individual MNs throughput.

The aggregate performance is sometimes more interesting, as it is often more useful for network provisioning. We, therefore, direct our focus to the aggregated behavior of MNs and the overall network performance. Fig. 6(a) and (b) plots the processed traffic load (in bytes) and the number of packet drops at each inter-MAP link. It is clearly shown that, compared with HMIPv6 and HMIPv6-UP, DEMAPS achieves better distribution of traffic load among the MAPs. Packet drops are almost null when DEMAPS is used. This is mostly due to the transmission of packets over congested links upon handoff occurrences. Given the limited buffer size of routers, an important amount of these packets are dropped. This justifies the high values of packet drops that were experienced in HMIPv6 and HMIPv6-UP, as shown in Fig. 6(b). Note that the plotted DEMAPS traffic includes both UDP data packets and signaling packets. Nevertheless, Fig. 6(a) shows that the overall bandwidth consumption
in the case of DEMAPS (including signaling packets) is almost the same as that of HMIPv6 and HMIPv6-UP, which confirms that the additional cost due to signaling packets is minimal. In addition, the obtained performance gains are worthwhile and can be used to advocate the small overhead that may be incurred by frequent transmissions of MAP option messages.

We next evaluate the performance of the three schemes in terms of BU traffic and BU latency. Fig. 7(a) and (b) graphs the number of BU messages at each network element and the BU latency that was experienced by each MN and averaged by the number of BUs. Fig. 7(a) indicates that HMIPv6 and HMIPv6-UP reduce the frequency of BU messages to HAs. However, these schemes do not consider the load distribution among multiple MAPs, i.e., many data packets drop at congested links, as we have described earlier. Although DEMAPS certainly increases the frequency of BU messages to HAs, it significantly reduces the packet drops and achieves the highest throughput for MNs compared with the other two schemes. Fig. 7(b) shows that the average BU latency for some MNs in the case of HMIPv6 and HMIPv6-UP is higher than in the case of DEMAPS. This is due to the selection of heavily loaded MAPs, which ultimately leads to high queuing delays and results in high BU latencies.

In the proposed scheme, information on load transition is sent to ARs every Δ time interval. In the simulations that were conducted so far, Δ was set to 1 s. To investigate the effect of Δ on the DEMAPS performance, we plot the number of packets processed by inter-MAP links for different values of Δ in Fig. 8(a). The figure demonstrates that setting Δ to higher values results in a poor distribution of network traffic among MAPs. The choice of Δ is a compromise between enhancing the traffic distribution and reducing the frequency of MAP option messages. Indeed, small values of Δ would efficiently distribute the data traffic on the network, whereas large values of Δ would reduce the number of MAP option messages that are sent over the communication time.

For more clarity, the following index is used [33]:

$$\Phi = 1 - \frac{\sum_{i=1}^{N} |\alpha_i - \bar{\alpha}|}{2\bar{\alpha}(N - 1)}$$  \hspace{1cm} (4)

where \(\alpha_i\) is the number of packets that were processed by the \(i\)th link, and \(N\) is the number of inter-MAP links. \(\bar{\alpha}\) is the average value of \(\{\alpha_i, i = 1, \ldots, N\}\). Φ captures the efficiency of traffic distribution over the network and ranges from zero to one. Low values of Φ represent a poor distribution of network traffic and lead to significant packet drops. Fig. 8(b) graphs the value of Φ for different values of Δ and demonstrates that setting Δ to values that are larger than 30 s degrades the traffic distribution over the network. On the other hand, results based on Fig. 6(a) show that the system overhead remains minimal when we set Δ to 1 s. Note that similar
Fig. 8. Impact of $\Delta$ on the performance of DEMAPS. (a) Number of data packets that were processed at each inter-MAP link for different values of $\Delta$. (b) Traffic distribution index $\Phi$ for the values of $\Delta$.

Fig. 9. Prediction accuracy of EMA for two different values of $\Delta$. (a) $\Delta = 1$ s. (b) $\Delta = 10$ s.

experiments were conducted, considering different traffic mobility patterns, and identical results were obtained. To conclude, ($\Delta = 1$ s) represents a good tradeoff between an efficient distribution of data traffic and a reduced frequency of MAP option packets.

Another credit of small values of $\Delta$ consists of the guarantee of high prediction accuracy of the EMA method. To illustrate this idea, we investigated the variation in the actual data traffic and the estimated value of EMA ($E[\alpha]$) during a period of the simulation time. Fig. 9(a) and (b) graphs the variation in the measured load and the EMA value for two different values of $\Delta$, i.e., 1 s and 10 s, respectively, and demonstrates that setting $\Delta$ to high values results in a deviation of the EMA estimation from the actual data load. This step obviously affects the prediction accuracy of the load transition. However, setting $\Delta$ to small values provides more accurate prediction of the load transition. Note that carefully chosen values of $\Delta$ maintain an efficient distribution of the network traffic, guarantee higher accuracy of the load transition prediction, and minimize the system overhead by reducing the transmissions of MAP option packets.

So far, the exponential smoothing constant $r$ has been set to 0.9. This setting aims at giving more weight to the latest data in the load transition prediction. Admittedly, the constant $r$ plays a major role in predicting the load transition, and it may, therefore, be thought as largely affecting the distribution of the network traffic. To investigate such a possible impact, the variation of the traffic distribution index $\Phi$ is plotted for different values of $r$. Fig. 10 demonstrates that the system guarantees an efficient traffic distribution for all the values of $r$ and reaches its optimum when $r$ takes large values (e.g., in the vicinity of 1.0). This efficient distribution is manifested in the form of high values of $\Phi$ ($\Phi > 0.93$).

V. DISCUSSION

As we have previously discussed, the DEMAPS scheme notifies MNs of the most optimal MAP to use upon handoffs. For this purpose, the DEMAPS scheme neither generates any
new signaling packets, modifies the HMIPv6 protocol itself, nor require any modifications at the mobile terminals. Thus, the implementation of the proposed scheme is practical and can widely be accepted.

One scenario that may put limitations on the performance of DEMAPS is when an MN is asked (by an AR) to register with a new MAP, rather than the old MAP that it was previously using, whereas it could have kept using the same old MAP without fearing any congestion of the latter. Such a scenario will ultimately oblige the MN to register again with its HA and CN. All of these steps are admittedly unnecessary and may defeat the purpose of having HMIPv6 in the first place. As a remedy to this issue, MNs should be given freedom in choosing the MAPs to which they register. For example, an AR can list all the available MAPs with which it is connected, along with the information on their resources utilization, in the RAs. Upon receiving the RA, the MN first verifies if the old MAP that it has been using is among the list. If so, it refers to the resource-utilization and load-transition tendency of the MAP. It then decides whether it should keep the registration with the same MAP or change it to another MAP, obviously with a lighter traffic load. If the load of the old MAP is not at a critical point, the MN can keep registering with it. It accordingly saves itself from sending BU messages to the HA and CN. If the PMAP is not available at the list or is handling a relatively higher load compared with other MAPs, the MN selects another MAP from the list and updates its binding following the update procedures of HMIPv6.

Selecting the MAP may require some energy at the MN. This operation is, however, performed only upon handoff, and the consumed energy shall be minimal. The MAP selection may slightly affect the battery life of the MN only in the case of fast moving nodes that perform frequent handoffs or keep flip-flopping over the overlapped area between the coverage areas of multiple ARs. Although this scenario is unlikely to happen, the energy issue can be mitigated by implementing the MAP decision-making mechanism at ARs. In such a case, the MN should notify the AR of the old MAP that it was using prior to handoff. This notification can be done via the RS message.

The working of DEMAPS can further be enhanced by anticipating the occurrence of MN handoffs. This anticipation is possible by using cross-layer protocols [35], along with predicting the user movement based on the user mobility patterns [36]. On the other hand, the proposed scheme, although it has been tested in MIPv6 with fixed routers, can easily be applied with minor modifications to the networks with mobile routers, as used for seamless Internet access in public transportation and in wireless metropolitan networks (e.g., Worldwide Interoperability for Microwave Access and IEEE 802.16 [34]). In such environments, the network topology of mobile routers can be considered dynamic through dynamic virtual topology [37]. This way, the network can be modeled as a set of time-discrete snapshots of network topologies of fixed routers over short slots of time (e.g., a few seconds). This idea is similar to most routing concepts that were proposed for ad hoc and mobile sensor networks in the recent literature [38], [39]. Given that, in the proposed scheme, the load-transition prediction is performed over short periods of time (Δ = 1 s), the proposed scheme can efficiently be applied to these snapshots of the mobile network topology and, ultimately, to the entire network while it is on the move. The performance evaluation of the proposed scheme for such scenarios can be an interesting topic for future investigation.

VI. CONCLUDING REMARKS

In this paper, we have proposed a method that significantly improves the performance of HMIPv6 in large mobile networks. In such large networks, operators may need to deploy a set of MAPs over a given domain. The proposed scheme has provided a reliable balanced traffic load among these MAPs. Although most of the strategies that were proposed earlier in the literature attempt to solve the macromobility issues and provide fast transition performance, they create a complex landscape for network traffic management. Some routers are overly congested with packets, whereas others are underutilized.

The proposed DEMAPS is a dynamic efficient technique for selecting the most appropriate MAP for registration. It functions as HMIPv6 when the network is not overloaded. When the network is running under heavy loads, the MAP selection is based on an estimation of MAP load transition using the EMA method. Information on load transition is notified to ARs via the transmission of MAP option messages. The proposed scheme is easy to implement, and the additional cost that signaling packets requires is proven to be minimal.

Extensive simulation results have demonstrated that the proposed scheme has the potential of substantially improving the average communication delay, reducing the number of losses, and making better utilization of the network resources. All of these achievements are highly important for implementing integrated or differentiated services (DiffServ) architectures to support QoS over mobile IP networks. This achievement represents a major goal of the research that was carried out in most of the IETF Working Groups. The actual enhancements that the proposed DEMAPS scheme can bring to DiffServ deserve further study and form the basis for our future research.

REFERENCES


Tarik Taleb (S’04–M’05) received the B.E. degree (with honors) in information engineering in 2001 and the M.E. and Ph.D. degrees in computer sciences in 2003 and 2005, respectively, from Tohoku University, Sendai, Japan.

From October 2005 to March 2006, he was a Research Fellow with the Intelligent Cosmos Research Institute, Sendai. He is currently an Assistant Professor with the Graduate School of Information Sciences, Tohoku University. He is serving on the editorial boards of Wiley journals. His research interests include wireless networking, satellite and space communications, congestion control protocols, mobility and handoff management, multimedia streaming, and network security.

Dr. Taleb is on the editorial board of the IEEE WIRELESS COMMUNICATIONS and the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS journal. He is currently the Vice Chair of the Satellite and Space Communications Technical Committee of the IEEE Communications Society. He has been the Chair of several symposia on wireless networking in international conferences such as the 2008 IEEE International Wireless Communications and Mobile Computing Conference (IWCMC ‘08) and the 2009 IEEE International Conference on Telecommunications (Globecom ’09). He has also served on the technical program committees of other IEEE conferences, including the International Conference on Communications (ICC) and the Wireless Communications and Networking Conference (WCNC). He received the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2008, the Funai Foundation Science Promotion Award in April 2007, the IEEE Computer Society Japan Chapter Young Author Award in December 2006, the Niwa Yasujiro Memorial Award in February 2005, and the Young Researcher’s Encouragement Award from the IEEE Vehicular Technology Society (VTS) Japan Chapter in October 2003.

Abbas Jamalpour (S’86–M’91–SM’00–F’07) received the Ph.D. degree from Nagoya University, Nagoya, Japan. He is currently with the School of Electrical and Information Engineering, University of Sydney, Sydney, Australia. He is a Technical Editor of Wiley’s International Journal of Communication Systems and several other journals. He is the author of the first book on wireless IP and two other books, is a coauthor of seven books and more than 190 journal articles and conference proceedings, and is the holder of two patents on wireless telecommunications.

Dr. Jamalpour is an IEEE Distinguished Lecturer and a Fellow of the Institution of Engineers—Australia. He has been very active within the IEEE Communications Society, where he was the Chair of the Satellite and Space Communications Technical Committee from 2004 to 2006, and is currently the Vice Chair of the Communications Switching and Routing Technical Committee. He is also the Chapter Chair of the Asia-Pacific Board Coordinating Committee and is the Editor in Chief of IEEE Wireless Communications, and is a Technical Editor of IEEE Communications Magazine. He is a voting member of IEEE GLOBECOM/ICC Technical Content Committee (GICT) and was the Vice Chair of the 2003–2006 IEEE Wireless Communications and Networking Conference (WCNC), the Chair of the 2005 IEEE International Conference on Telecommunications (Globecom) Wireless Communications Symposium, and a Symposium Cochair of the IEEE International Conference on Communications (ICC) from 2005 to 2008 and the IEEE GLOBECOM from 2006 to 2007. He has received several international awards, including the Best Tutorial Paper Award and the Distinguished Contribution to the Satellite Communications Award from the IEEE Communications Society in 2006.
Yoshiaki Nemoto (S’72–M’73–SM’05) received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1968, 1970, and 1973, respectively.

He was the Director of the Information Synergy Center, Tohoku University, where he is currently a Full Professor with the Graduate School of Information Sciences. His research interests include microwave networks, communication systems, computer network systems, image processing, and handwritten character recognition.

Dr. Nemoto is a Fellow of the Institute of Electronics, Information, and Communication Engineers and the Information Processing Society of Japan. He is a recipient of the 2005 Distinguished Contributions to Satellite Communications Award from the IEEE Communications Society and a corecipient of the 1982 Microwave Prize from the IEEE Microwave Theory and Techniques Society.

Nei Kato (M’03–A’04–SM’05) received the M.S. and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1988 and 1991, respectively.

He is currently a Full Professor with the Graduate School of Information Sciences, Tohoku University. His research interests include computer networking, wireless mobile communications, image processing, and neural networks.

Dr. Kato is a member of the Institute of Electronics, Information, and Communication Engineers of Japan. He has served on several Technical Program and Organizing Committees of international conferences. Since 2006, he has been a Technical Editor of *IEEE Wireless Communications*. 