INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2012; **25**:185–204
Published online 28 April 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/dac.1263

An optimized QoS scheme for IMS-NEMO in heterogeneous networks

Jianxin Liao[1,2,*,†], Qi Qi[1,2], Tonghong Li[3], Yufei Cao[2], Xiaomin Zhu[1,2] and Jingyu Wang[1,2]

[1]*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China*
[2]*EBUPT Information Technology Co., Ltd., Beijing 100191, P.R. China*
[3]*Technical University of Madrid, Madrid 28660, Spain*

SUMMARY

The network mobility (NEMO) is proposed to support the mobility management when users move as a whole. In IP Multimedia Subsystem (IMS), the individual Quality of Service (QoS) control for NEMO results in excessive signaling cost. On the other hand, current QoS schemes have two drawbacks: unawareness of the heterogeneous wireless environment and inefficient utilization of the reserved bandwidth. To solve these problems, we present a novel heterogeneous bandwidth sharing (HBS) scheme for QoS provision under IMS-based NEMO (IMS-NEMO). The HBS scheme selects the most suitable access network for each session and enables the new coming non-real-time sessions to share bandwidth with the Variable Bit Rate (VBR) coded media flows. The modeling and simulation results demonstrate that the HBS can satisfy users' QoS requirement and obtain a more efficient use of the scarce wireless bandwidth.

Received 10 June 2010; Revised 27 November 2010; Accepted 2 February 2011

KEY WORDS: NEMO; QoS; heterogeneous networks; IMS

1. INTRODUCTION

Forthcoming fourth generation (4G) networks are expected to enable the users with portable devices to maintain the Internet connectivity through different and heterogeneous technologies, anytime and anywhere [1]. Most studies on mobility management have focused on the technologies to support host mobility in an individual manner [2]. In order to address the group mobility (for example, in a vehicle, there may exist tens or hundreds of users moving together), the concept of network mobility (NEMO) [3] is introduced by Internet Engineering Task Force (IETF). In NEMO schemes, all users in a vehicle are packed as a unit which is called mobile network (MN) and an inside gateway takes charge of the mobility management for all of them. Thus, Mobile Terminals (MTs) in the MN get rid of the individual process of mobility management such as location update and handover, and accordingly the signaling cost in the core network is reduced. For the QoS control in NEMO, however, the individual resource reservation has high signaling cost, because each active session transmits the same signaling from the inside gateway to the outside access network for QoS requesting and QoS state refreshing. Thus, several schemes [4–8] are proposed, in

*Correspondence to: Jianxin Liao, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China.
†E-mail: liaojx@bupt.edu.cn

which the inside gateway realizes bandwidth sharing and QoS request aggregation by maintaining a single QoS reservation state on behalf of all the MTs.

The IP Multimedia Subsystem (IMS) developed by Third Generation Partnership Project (3GPP) provides IP-based control plane architecture for the evolution of 4G networks. In IMS, Session Initiation Protocol (SIP) is the session control signaling, which provides QoS support through: (1) QoS description in Session Description Protocol (SDP); (2) QoS negotiation at session establishment; (3) QoS modification during the session [9]. SIP-based NEMO (SIP-NEMO) [10] has been proposed to improve the NEMO basic protocol on mobility management. Although IMS-based NEMO (IMS-NEMO) can take use of the SIP-NEMO, the QoS negotiation and resource reservation are still an open issue.

Considering the evolution of 4G networks, current QoS control schemes [4–8] have two drawbacks: (1) heterogeneous access networks are not considered, and all the sessions inside the MN can only be connected to a single network. Thus, the available resources of multiple access networks are not fully utilized; (2) the reserved bandwidth from the inside gateway to the access network is not utilized effectively. The Variable Bit Rate (VBR) coded media stream reserves bandwidth according to its peak bandwidth, thus it may have unused idle bandwidths during the session holding time. If the MN can make use of these idle bandwidths to serve the delay insensitive sessions, the scarce wireless resource can be saved.

Accordingly, this paper proposes a heterogeneous bandwidth sharing (HBS) scheme for the IMS-NEMO QoS control, in which the SIP signaling for QoS negotiation and resource reservation are presented for the first time. HBS includes two mechanisms to optimize the current QoS schemes: (1) the multi-access network selection policy, enabling the MN to select the optimal access network for each session; and (2) the Idle Bandwidth Sharing (IBS) algorithm, which can reduce the utilized bandwidth of the MN by making use of the idle bandwidth of VBR traffics. As HBS only needs several enhanced functionalities in the gateways of the MN and the access networks, its implementation is feasible.

The remainder of this paper is organized as follows. Section 2 surveys the related work. Section 3 proposes HBS scheme including IMS-NEMO QoS architecture, multi-access network selection policy and idle bandwidth sharing algorithm, with our original contribution presented in detail. Section 4 develops a k-D Markov model for analyzing the utilized bandwidth and the session blocking probability for HBS scheme. And the analytic model is validated against simulation experiments. Section 5 investigates the performance of HBS. Finally, Section 6 concludes the paper.

2. RELATED WORK

To support group mobility, the NEMO basic protocol (NEMO BSP) is proposed by IETF. In the NEMO BSP, there is a central node called Mobile Router (MR) that connects to the Internet and provides networking services for the attached MTs. But the NEMO BSP inherits the drawbacks of Mobile IPv6, such as inefficient routing path, single point of failure, long handover latency, high packet loss and high packet overhead. Most of the recent research efforts on MIP-NEMO have concentrated on solving these problems resulting in several optimization schemes [11–17].

Different from the MIP-based NEMO (MIP-NEMO), Huang *et al.* [10] first introduce SIP-NEMO, which uses SIP to provide mobility support for NEMO. In SIP-NEMO, the core element of the MN is a SIP NEMO Server (SIP-NMS), which is responsible for the mobility management of all the MTs in the MN. The SIP-NMS acts as a gateway and translates the SIP header of each incoming or outgoing message by using the mechanism of the Network Address Translation (NAT). In this way, SIP-NMS keeps all the attached MTs globally reachable and transparent to the movement of the MN. Pack *et al.* [18] point out that SIP-NEMO can easily be deployed and reduce the tunneling overhead incurred in the NEMO BSP; however, it increases the handoff latency due to the large message length. Therefore, some researchers improve SIP-NEMO on the session establishment and location management [19, 20]. Recently, Chiang *et al.* [10]

DOI: 10.1002/dac

integrate SIP-NEMO in IMS, and propose two interworking architectures: loosely coupled and tightly coupled.

For QoS control in NEMO, Resource Reservation Protocol (RSVP)-based resource reservation protocols cannot be applied directly due to the following two reasons: (1) the MTs inside the MN are unaware of MN's mobility; (2) excessive signaling overhead. To solve these problems, protocols that reserve resources for the active sessions inside the MN are proposed. The On-Board RSVP [4] extends the well-known RSVP protocol by adding a new object to the PATH and RESV messages to provide the best-effort treatment for the sessions inside the MN. The NEMOR [5] protocol uses a generic signaling protocol called Next Step In Signaling (NSIS) to exploit Differentiated Services (DiffServ) on the bi-directional tunnel between the MR and its home agent, and Integrated Services (IntServ) between MR's home agent and the corresponding nodes. The MR maintains a single QoS reservation state on behalf of the MTs that it serves.

Moreover, a Mobile Bandwidth-Aggregation (MBA) reservation scheme is proposed by Wang *et al.* [6] to support QoS guaranteed services for MIP-NEMO. In MBA, the MR works as the proxy of all MTs insides the MN, which aggregates and reserves their required bandwidths. Also, three bandwidth reservation policies, i.e. static reservation, dynamic reservation and hybrid reservation, are presented. The hybrid reservation in MBA requests a static increment of bandwidth dynamically whenever there is no spare bandwidth to accommodate new requests. The MBA scheme resolves the mobility unawareness and excessive signaling overhead problems in supporting QoS for MIP-NEMO.

Although the above bandwidth aggregation schemes improve the signaling efficiency by maintaining a single QoS state for the whole MN rather than for each individual session, its benefit may be undermined by the necessity of constantly maintaining this state when sessions are frequently created and terminated. Thus, Kamel *et al.* [7, 8] propose a dynamic cost-driven QoS aggregation policy, which aims at minimizing the cost per unit time of each aggregation cycle. The cost consists of a cost for holding QoS requests and a cost for sending a message to the access network to create or update a resource reservation. However, how to quantize the cost of holding a new session is not mentioned.

In conclusion, the current optimized QoS schemes for NEMO are based on RSVP [4–8]. However, they cannot totally realize the QoS control for IMS-NEMO, because QoS control signaling during the session establishment in IMS includes QoS negotiation and resource reservation which are based on SIP and Diameter [9]. Furthermore, both MBA scheme [6] and cost-optimal QoS aggregation method [7] have the following shortcomings: (1) the heterogeneous wireless environment provides more resources for mobile users than traditional single network, but both schemes cannot dynamically select the best network for each session according to its QoS requirements; (2) the policy of QoS adjustment for the arrival of new sessions does not effectively make use of the bandwidth that has been reserved for the whole MN.

3. A NEW QoS SCHEME FOR IMS-NEMO

The main idea of the new HBS scheme is to provide QoS support in the IMS-NEMO with SIP. Also, we include a heterogeneous network selection policy and an idle bandwidth sharing algorithm, in order to optimize the resource utilization without significantly sacrificing the QoS of each session within the MN.

3.1. IMS-NEMO QoS architecture

Figure 1 shows the IMS-NEMO architecture in the heterogeneous access environment, which uses SIP-NEMO [10] for mobility management. In IMS network, the Packet Data Network Gateway (PGW) provides an integrated access to the IMS core network for all the access networks belonging to its serving domain. The Policy and Charging Rule Function (PCRF) makes policy decisions and

DOI: 10.1002/dac

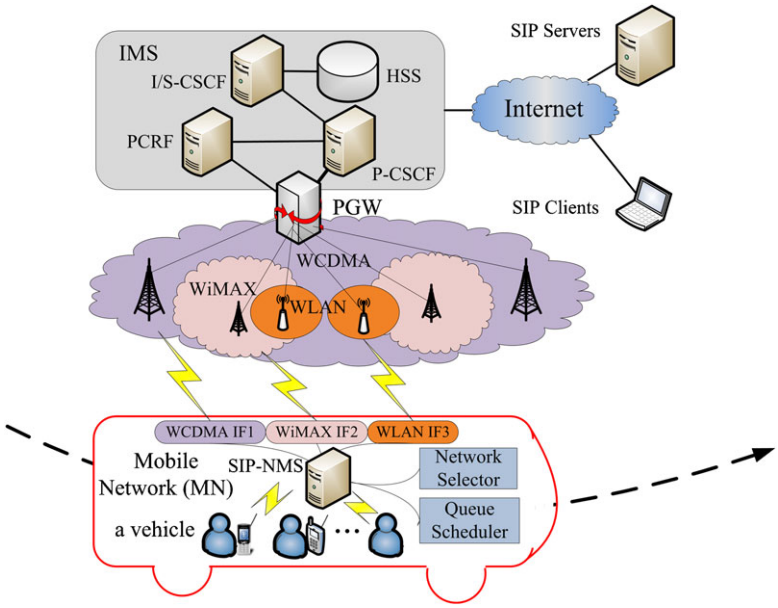


Figure 1. The architecture of the IMS-NEMO.

authorizes QoS resources. The Policy and Charging Enforcement Function (PCEF) is implemented in the PGW for resource reservation. The Proxy Call Session Control Function (P-CSCF) is the access point of IMS core network, which acts as the Application Function (AF) in the QoS policy decision [9, 21].

In the MN, SIP-NMS is the gateway with session border controller (SBC) mechanism, which aggregates the session control signaling and provides QoS support for IMS-NEMO by reserving an entire resource in the outside IMS network on behalf of all the sessions within the MN. The SIP-NMS connects MTs inside the MN by Wireless Local Area Network (WLAN) interfaces and attaches to the Internet through several external wireless interfaces, such as Wideband CDMA (WCDMA), World Interoperability for Microwave Access (WiMAX) and WLAN interfaces. For each wireless network, a bearer referring to a logical IP transmission path with specific QoS properties is established between the SIP-NMS and the PGW. Compared with the SIP-NMS in SIP-NEMO [10], two new components (Network Selector and Queue Scheduler) are added in our SIP-NMS. The Network Selector takes charge of selecting the suitable access network for each session according to its QoS requirements. The Queue Scheduler provides QoS guarantee for all the IP flows and makes use of the idle bandwidths of VBR flows in the established bearer to serve the delay insensitive sessions.

The SIP-NMS in SIP-NEMO provides a SIP URI-list service in order to avoid excessive signaling messages on wireless links when MN moves to a new network [10]. But the bearer establishment request (e.g. Activate Packet Data Protocol context request in Universal Mobile Telecommunications System) and the Diameter signaling do not support URI-list service, so the SIP URI-list service cannot totally realize the establishment of aggregated sessions in IMS-NEMO. Therefore, we provide an entire bearer establishment for IMS-NEMO, where the SIP-NMS aggregates the QoS requests and maintains an entire resource reservation state for the whole MN. The process of session aggregation can be invoked by the arrival of new sessions or the handover when the MN moves to a new IMS domain served by another P-CSCF. In Figure 2, the QoS control signaling for IMS-NEMO is presented, which consists of three phases: (1) QoS negotiation; (2) Bearer establishment and resource reservation and (3) QoS approval.

(1) QoS negotiation. A session's QoS profile is depicted in SDP, which includes media codec, bit-rate and bandwidth requirements. The SIP-NMS aggregates new session requests or initiates a handover request by sending a SIP INVITE (or re-INVITE when a handover happens) with

DOI: 10.1002/dac

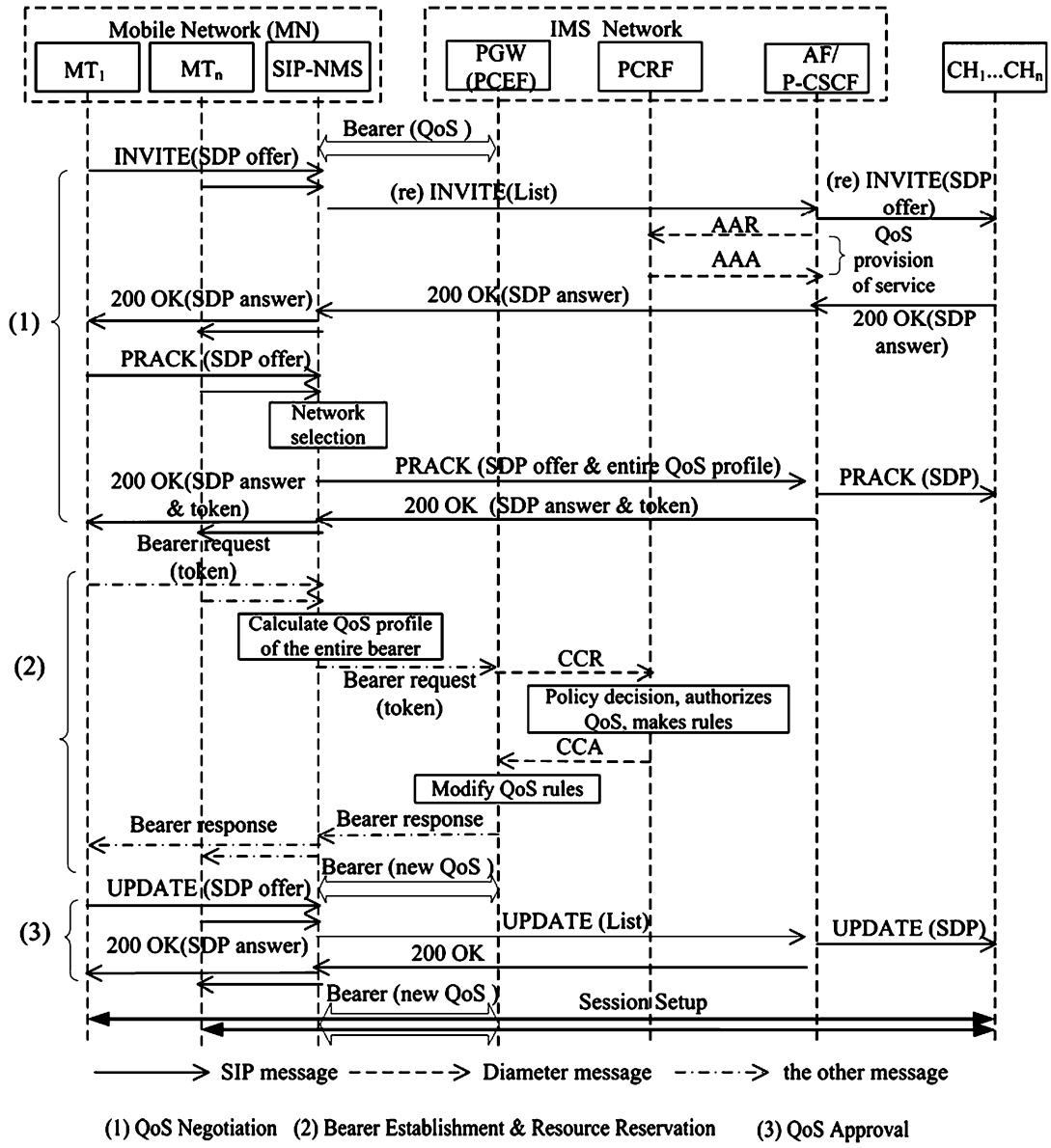


Figure 2. The QoS control signaling for IMS-NEMO.

a list of URIs and SDPs. The QoS negotiation is performed through the SDP Offer/Answer model. In this phase, the SIP-NMS selects suitable access network for each session (introduced in Section 3.2), and the P-CSCF/AF sends the service-based QoS information to the PCRF for policy decision.

(2) Bearer establishment and resource reservation. The SIP-NMS aggregates the bearer establishment requests from MTs, and sends an entire bearer establishment request to the PGW with each session's token after calculating the new QoS profile of the entire bearer. The entire resource reservation is finished in PGW after the policy decision, QoS profile authorization and QoS rule indication in PCRF. Afterwards, the bearer from the SIP-NMS to the PGW with the new QoS profiles is established.

(3) QoS approval. The SIP-NMS aggregates the UPDATE messages to notify the Corresponding Hosts (CHs) that the resource reservation is finished. After the sessions are set up, the media flows can be transmitted through the end-to-end entire bearer path.

DOI: 10.1002/dac

3.2. Multi-access network selection policy

The heterogeneous networks provide a variety of access technologies to the users. Thus, when the MN moves in the heterogeneous environment, the SIP-NMS should choose the most suitable access technology for each session. For example, when a vehicle travels in the city where the WLAN hotspots are overlaid with 3G network, a business person may prefer to connect the ongoing multimedia conference to the 3G network for service continuity; on the other hand, a student may connect the HTTP sessions to the WLAN for obtaining more bandwidth with the low price.

The selection of access network according to multiple criteria is a complicated optimization problem. For the normal MT, user preference is a good means to overcome the complexity of the multiple criteria [22]. However, it is not enough that SIP-NMS only considers user preference for the selection of access network. In our scheme, SIP-NMS takes into consideration the following three factors: (1) MN moving speed (V); (2) user preference (P) and (3) network load (L). Let $R_\theta^\sigma(s)$ represents access network θ's feasibility to serve session s when only factor σ is considered, where $\sigma \in \{V, P, L\}$, $\theta \in \{N_1, \ldots, N_n\}$ and $0 \leqslant R_\theta^\sigma(s) \leqslant 1$.

As all users in the MN move with the same velocity as MN, all sessions have the same $R_{N_i}^V(s)(1 \leqslant i \leqslant n)$. Assume that v is velocity of the MN, and v_{N_i} is the maximum velocity supported by access network N_i.$R_{N_i}^V(s)$ can be obtained according to v and v_{N_i}. For example, when the MN's velocity v is greater than 18 km/h, it cannot be served by WLAN because it crosses the WLAN cell in a short time (several hundreds of seconds). Similarly, if MN's velocity is greater than 72 km/h, WiMAX cannot provide good QoS to its sessions [22]. Thus, we define $R_{N_i}^V(s)$ as Equation (1), where 1 and 0 means that session s can and cannot be served by access network N_i, respectively.

$$R_{N_i}^V(s) = \begin{cases} 1, & 0 \leqslant v < v_{N_i} \\ 0, & v_{N_i} \leqslant v \end{cases} \tag{1}$$

$R_{N_i}^P(s)$ is computed based on the user preference. For simplicity, we define five levels (extremely suitable, suitable, medium, unsuitable, forbidden), which correspond to values (4, 3, 2, 1, 0). Based on the user preference, session s can decide its level of preference with respect to access network N_i, which is denoted as $U_{N_i}(s)$. For example, consider the charging policy defined in the user preference. If the charging policy is economical, we have: $U_{3G}(s) = 1$ (unsuitable); $U_{\text{WiMax}}(s) = 3$ (suitable); and $U_{\text{WLAN}}(s) = 3$ (suitable). We define $R_{N_i}^P(s)$ as Equation (2), which is the normalized $U_{N_i}(s)$.

$$R_{N_i}^P(s) = \frac{U_{N_i}(s)}{\sum_{j=1}^{n} U_{N_j}(s)} \tag{2}$$

To compute $R_{N_i}^L(s)$, we take into consideration the required bandwidth of session s and the load of each access network. Obviously, the access network with fewer loads is more suitable to serve the new session. Assume C_{N_i} is the bandwidth capacity of N_i, B_{N_i} is the unused bandwidth of N_i, and $b(s)$ is the required bandwidth of session s, we define $R_{N_i}^L(s)$ as:

$$R_{N_i}^L(s) = \begin{cases} 0, & b(s) \geqslant B_{N_i} \\ 1 - \dfrac{B_{N_i} - b(s)}{C_{N_i}}, & 0 \leqslant b(s) < B_{N_i} \end{cases} \tag{3}$$

In the process of selecting the best access network, the user preference (P) and network load (L) are considered together. Let α^L and α^P be the weight of L and P, respectively, where $\alpha^L, \alpha^P \in (0, 1)$ and $\alpha^L + \alpha^P = 1$. The weights can be defined by the network operator based on the requirements. For each access network N_i, we calculate its score as follows:

$$F_{N_i}(s) = a^L R_{N_i}^L(s) + a^P R_{N_i}^P(s) \tag{4}$$

If there is only one access network, $R_{N_i}^V(s)$, $R_{N_i}^P(s)$ and $R_{N_i}^L(s)$ do not need to be calculated for each new session. But when the MN moves to a new area where multiple access networks coexist,

DOI: 10.1002/dac

the multi-access network selection is triggered in the case of a new session arriving. The velocity of the MN is first considered. If only one access network is suitable for the MN, all the sessions are connected to that network, and the selection procedure is finished. Otherwise, more than one candidate access network can serve the MN in terms of its velocity. As a result, we select the best network by considering both user preference and network load: the SIP-NMS computes the weighted scores of all candidate access networks by Equation (4). Finally, the candidate access network with highest score is selected.

3.3. Idle bandwidth sharing (IBS) algorithm

The SIP-NMS maintains the bearer from SIP-NMS to the PGW according to the QoS profiles of all sessions. However, when a new session arrives or an active session terminates, the bandwidth of bearer may need to be adjusted. However, frequently adjusting the bandwidth can result in high signaling cost. We propose an idle bandwidth sharing algorithm, in order to reduce the QoS control signaling and save the bandwidth utilized by the MN.

Four QoS traffic classes are defined in UMTS [9, 23]: conversational, streaming, interactive and background. Conversational and streaming classes are mainly used to carry real-time traffic flows, which require sufficient bandwidth for media transmission. On the other hand, interactive and background classes are mainly adopted by non-real-time Internet applications such as WWW, Email, Telnet and FTP, due to their loose delay requirements. As for multimedia traffics, two encoding techniques can be used: (1) Constant Bit Rate (CBR) in which the frame size is fixed; (2) VBR in which the frame size is varied according to the media content. For the VBR flow, the resource is reserved in terms of its peak bandwidth during QoS negotiation. If the used bandwidth is less than the peak value, there exists idle bandwidth. As all sessions inside MN shares the bearer from SIP-NMS to PGW, the idle bandwidth reserved for VBR traffic can be exploited to serve the non-real-time flows of new sessions. Thus, if a new session only contains interactive or background traffics, there is no need to adjust the reserved bandwidth when the session is established. As a result, the QoS control signaling can be reduced and the bandwidth utilized by the bearer can be saved.

To simplify the explanation of IBS algorithm, we make the following assumptions:

(1) We only consider access network N_i, although the SIP-NMS may connect to multiple access networks simultaneously.
(2) We only discuss the process of uplink flows in SIP-NMS. The process of downlink flows is similar to that of uplink flows and it is performed in PGW.
(3) A session's bandwidth requirements are depicted in the SDP, which includes each media flow's transmitting rate.

The IBS algorithm works as follows. If a new session only contains non-real-time flows, it uses the idle bandwidth of VBR traffics, and is set up directly without the need of resource reservation shown in the phase 2 of Figure 2. A Non-Real-time Queue (NRQ) is allocated to buffer its packets. Otherwise, the session is held by setting a timer VT (the length of VT is a parameter) and n_s that denotes the number holding sessions is increased by one. If n_s exceeds the defined threshold or one of VT timers expires, the SIP-NMS calculates the required total bandwidth and then sends the aggregate QoS request to update the bearer. Note that if the session includes both real-time and non-real-time flows, only the bandwidth required by its real-time flows is considered. After that session is set up, a NRQ is allocated for each of its non-real-time flows. On the other hand, if a non-real-time flow in an established session waits for the idle bandwidth over a threshold time, its NRQ is removed and a new session is initiated to allocate bandwidth for it. To reduce the QoS control signaling, we also first hold this new session by setting a timer VT and increase n_s by one. After its required bandwidth is reserved, a Non-Real-time Flow (NRF) queue is allocated for it. Furthermore, if a session is initiated by a high-end user, each of its non-real-time flows is allocated an NRF queue in order to ensure its high priority.

Figure 3 shows the scheduling queues used by the IBS algorithm. Assume S_v is the set of active real-time VBR flows, S_{nr} is the set of non-real-time flows and T is the bandwidth aggregation

DOI: 10.1002/dac

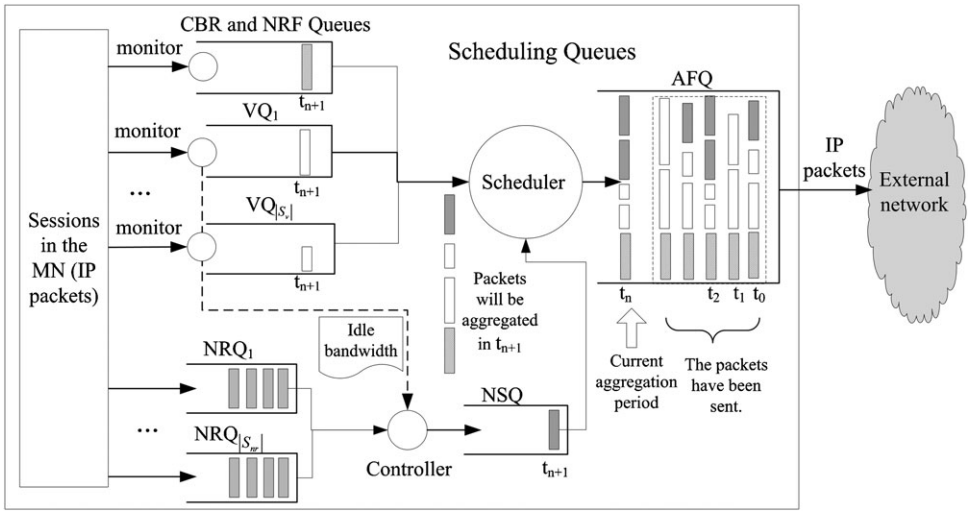


Figure 3. The scheduling queues in the SIP-NMS.

period. When a session is established, each of its VBR flow, CBR flow and non-real-time flow is allocated a VBR Queue (VQ), CBR queue and a NRQ to buffer its packets, respectively. Each VQ is associated with an event trigged monitor, which provides the information about the idle bandwidth of its corresponding VBR flow. Also, the monitor in each VQ, CBR and NFQ queue is responsible for inserting a tag at the end of each T period. During a bandwidth aggregation period (t_n), the VQs, CBRs and NFQs buffer their packets that will be transmitted in next T time (t_{n+1}). At the end of t_n, Controller checks every monitor and obtains the information about the total idle bandwidth. At the start of next T time (t_{n+1}), Controller takes out the packets from each NRQ in order of 'Round Robin' based on the information about total idle bandwidth, and puts them into the NSQ. But if there is no idle bandwidth, no packets are sent to the NSQ. Afterwards, Scheduler takes all packets from NSQ and all packets before the tags from VQs, CBR and NRF queues and sends them to the Aggregate Flow Queue (AFQ) by using General Processor Sharing (GPS) method [24]. In this way, the Scheduler can guarantee the QoS requirements of real-time flows, and enables the non-real-time flows to use the idle bandwidths of VBR flows when aggregating the flows of all the sessions inside the MN.

Assume b_p^i is the peak bandwidth required by VBR flow i and b_v^i is the bandwidth consumption of VBR flow i during period T. Assume b_{idle} and b_{nr} are the total idle bandwidth provided by all the VBR flows and the bandwidth utilized by the non-real-time flows during the last T time, respectively. Let ps_j be the packet size of non-real-time flow j. t_n is the nth period of bandwidth aggregation, and its length is T. The process of the Controller and the Scheduler at the end of each T time is shown as follows.

1: $b_{\text{idle}}=0$; $b_{\text{nr}}=0$; $j=1$
2: **For** $i=1$ to $|S_v|$//calculate the idle bandwidth for t_n period
3: **If** $b_v^i < T \times b_p^i$
4: Controller computes the idle bandwidth: $b_{\text{idle}}=b_{\text{idle}}+(T \times b_p^i - b_v^i)$
5: **End If**
6: **End For**
7: **While** $b_{\text{nr}} < b_{\text{idle}}$ **do**
8: Controller fetches a packet from NRQ_j and sends it to NSQ
9: $b_{\text{nr}}=b_{\text{nr}}+ps_j$
10: $j=(++j)\text{mod}|S_{\text{nr}}|$
11: **End While**
12: Scheduler fetches packets from VQs, NSQ, CBR and NRF queues to AFQ

DOI: 10.1002/dac

4. AN ANALYTICAL MODEL

In this section, we present an analytical model to obtain the bandwidth consumption and the session blocking probability for HBS. Also, the simulation is used to verify the analytical results. The notations are summarized in Table I.

4.1. k-D Markov model for bandwidth reservation without IBS

We assume that the maximum number of media flows owned by each session is k. The arrivals of sessions to the MN are Poisson distributed with rates λ, and the session holding time follows an exponential distribution with a mean of $1/\mu$. The media flows can be VBR coded, CBR coded or non-real-time flows. Considering that a media flow requesting multiple units of bandwidth can be separated into multiple flows with each requesting one unit of bandwidth, for the sake of simplicity and without loss of generality, we assume that each media flow requests a single unit of bandwidth, which is denoted as B_u

Let $p_l(1 \leqslant l \leqslant k)$ be the proportion of sessions requesting l media flows among the total arrival sessions in MN. Let λ_l be the arrival rate of sessions requesting l units of bandwidth. We have:

$$\lambda_l = p_l \lambda \quad \text{with} \quad \sum_{l=1}^{k} p_l = 1 \tag{5}$$

Assume that n_b is the number of bandwidth units that the outside networks provide for the MN, which can be calculated as:

$$n_b = \lfloor B_N / B_u \rfloor \tag{6}$$

Table I. Summary of key notations.

Notations	Description
k	The maximum number of media flows contained in a session
λ	Session arrival rate in MN
μ	Session holding time
B_u	The size of bandwidth requested by a media flow
B_N	The size of bandwidth provided by the outside access network
n_b	The number of bandwidth units provided by the outside network
x_l	The number of sessions requesting $l(1 \leqslant l \leqslant k)$ units of bandwidth
$\mathrm{S}(x_1, \cdots, x_l, \cdots, x_k)$	A feasible state means that the MN has accepted x_1 sessions requesting 1 unit of bandwidth, x_2 sessions requesting 2 units of bandwidth, and so on
p_l	The probability that a session has $l(1 \leqslant l \leqslant k)$ media flows
λ_l	The arrival rate of sessions requesting l units of bandwidth
$\delta(x_1, \ldots, x_l, \ldots, x_k)$	Indicator function guaranteeing that infeasible states are not considered in the k-D Markov model
$\mathbf{P}$	The transition probability matrix of k-D Markov model
$\boldsymbol{\pi}$	The vector representing the steady state probabilities of k-D Markov model
m_b	The total bandwidth consumption for the MN
p_b^r	The blocking probability for the session requiring $r(1 \leqslant r \leqslant k)$ units of bandwidth
p_b	The total blocking probability
Q	The transition probability matrix of the ON/OFF Markov model for VBR flows
$\boldsymbol{v}$	The vector represent the steady state probability of the ON/OFF Markov model
v_1	The steady state probability of the state 'OFF' in the ON/OFF Markov model
λ_l'	The arrival rate of the session needing to request l units of bandwidth
$s_{l,i}$	A service with l flows
$ps_{l,i}$	The proportion of service $s_{l,i}$ among sessions containing l flows
$\lambda_{s_{l,i}}$	The arrival rate of sessions belonging to service $s_{l,i}$
$n_{v,s_{i,l}}$	The number of VBR flows in a session belonging to service $s_{l,i}$
$n_{nr,s_{i,l}}$	The number of non-real-time flows in a session belonging to service $s_{l,i}$
λ_{idle}	The rate of idle bandwidth generated by all sessions

DOI: 10.1002/dac

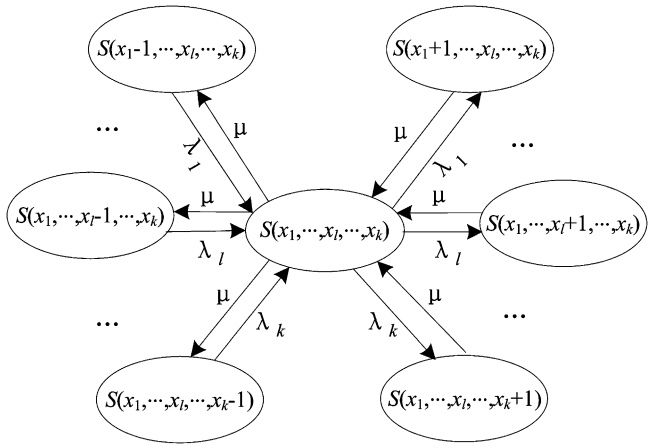


Figure 4. State transition diagram for a general state.

Here, B_N is the total bandwidth provided by the outside networks

To analyze the system behavior, a k dimension (k-D) Markov model is considered. Let x_l be the number of sessions requesting l units of bandwidth. In this model, a state $S(x_1, \ldots, x_l, \ldots x_k)$ means that the MN has accepted x_1 sessions requesting 1 unit of bandwidth, x_2 sessions requesting 2 units of bandwidth and so on. For example, when $k=3$, the state $S(4, 3, 5)$ represents the SIP-NMS has currently served four sessions with each requesting 1 unit of bandwidth, 3 sessions with each requesting 2 units of bandwidth and 5 sessions with each requesting 3 units of bandwidth.

Let S be the set of feasible states $S(x_1, \ldots, x_l, \ldots, x_k)$ satisfying $x_l \geqslant 0, (1 \leqslant l \leqslant k)$ and $\sum_{l=1}^{k} l x_l \leqslant n_b$. We define an indicator function $\delta(x_1, \ldots, x_l, \ldots, x_k)$ as follows:

$$\delta(x_1, \ldots, x_l, \ldots, x_k) = \begin{cases} 1 & \text{if } S(x_1, \ldots, x_l, \ldots, x_k) \in S, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The general state transition rates are deduced as follows. If a session requesting l media flows arrives when the state is $S(x_1, \ldots, x_l, \ldots, x_k)$, it transits to $S(x_1, \ldots, x_l+1, \ldots, x_k)$ with the rate of λ_l. If a session requesting l media flows terminates, the state $S(x_1, \ldots, x_l, \ldots, x_k)$ transits to $S(x_1, \ldots, x_l-1, \ldots, x_k)$ with the rate of μ. By the same method, the state transition rate from $S(x_1, \ldots, x_l-1, \ldots, x_k)$ to $S(x_1, \ldots, x_l, \ldots, x_k)$ is λ_l, and the state transition rate from $S(x_1, \ldots, x_l+1, \ldots, x_k)$ to $S(x_1, \ldots, x_l, \ldots, x_k)$ is μ.

Figure 4 shows the general state transition diagram at a non-boundary state $S(x_1, \ldots, x_l, \ldots, x_k)$. From this figure, we can deduce the Steady-State Balance Equation for $S(x_1, \ldots, x_l, \ldots, x_k)$ as:

$$\begin{aligned} & P_{(x_1,\ldots,x_l,\ldots,x_k)} \sum_{l=1}^{k} [\lambda_l \delta(x_1, \ldots, x_l+1, \ldots, x_k) + \mu \delta(x_1, \ldots, x_l-1, \ldots, x_k)] \\ & = \sum_{l=1}^{k} [\lambda_l P_{(x_1,\ldots,x_l-1,\ldots,x_k)} \delta(x_1, \ldots, x_l-1, \ldots, x_k) \\ & \quad + \mu P_{(x_1,\ldots,x_l+1,\ldots,x_k)} \delta(x_1, \ldots, x_l+1, \ldots, x_k)] \end{aligned} \tag{8}$$

where $P_{(x_1,\ldots,x_l,\ldots,x_k)}$ is the steady-state probability of $S(x_1, \ldots, x_l, \ldots, x_k)$. Similarly, we can obtain the Steady-State Balance Equations for all the other states. In addition, we have the following normalization constraint:

$$\sum P_{(x_1,\ldots,x_l,\ldots,x_k)} \delta(x_1, \ldots, x_l, \ldots, x_k) = 1.$$

DOI: 10.1002/dac

Here, we use the interactive method to solve the Steady-State Balance Equations [25]. First, we obtain the one-step transmission matrix from the Steady-State Equations. Second, the one-step transmission matrix is normalized to obtain the transition probability matrix **P**. Finally, the iterative power method is used to obtain the steady-state probability of **P**. Let $\boldsymbol{\pi}=[P_{(0,0,\ldots,0)}, P_{(1,0,\ldots,0)}, P_{(0,1,\ldots,0)}, \cdots, P_{(x_1,\ldots,x_l,\ldots,x_k)}, \cdots, P_{(n_b,0,\ldots,0)}]$ be the vector representing the steady-state probabilities. $\boldsymbol{\pi}$'s dimension is $|S|$ (i.e. the number of feasible states). Furthermore, let $\boldsymbol{\pi}(0)=[1, 0, \ldots, 0]$ be the vector representing the initial state probabilities, and $\boldsymbol{\pi}(n)$ be the state probabilities after n-step transition. By taking the limit of n, we can obtain $\boldsymbol{\pi}$ as follows:

$$\begin{aligned}\boldsymbol{\pi} &= \lim_{n\to\infty} \boldsymbol{\pi}(n) = \lim_{n\to\infty} \boldsymbol{\pi}(0) \times \underbrace{\mathbf{P}\times\ldots\times\mathbf{P}}_{n} \\ &= \boldsymbol{\pi}(0) \times \lim_{n\to\infty} \underbrace{\mathbf{P}\times\ldots\times\mathbf{P}}_{n} \end{aligned} \tag{9}$$

From the steady-state probabilities $\boldsymbol{\pi}$, we can derive the utilized bandwidth for the whole MN in total as:

$$m_b = B_u \left\{ \sum_{S(x_1,\ldots,x_l,\ldots,x_k)\in S} \left[P_{(x_1,\ldots,x_l,\ldots,x_k)} \sum_{l=1}^{k} l x_l \right] \right\} \tag{10}$$

For the session requesting $r(1\leqslant r\leqslant k)$ units of bandwidth, let S_b^r be the set of states $S(x_1, \ldots, x_l, \ldots, x_k)$ satisfying $r+\sum_{l=1}^{k} l x_l = n_b$, and its blocking probability can be computed as follows:

$$p_b^r = \sum_{S(x_1,\ldots,x_l,\ldots,x_k)\in S_b^r} P_{(x_1,\ldots,x_l,\ldots,x_k)} \tag{11}$$

Thus, the total blocking probability is computed as:

$$p_b = \sum_{r=1}^{k} p_r p_b^r = \sum_{r=1}^{k} p_r \left[\sum_{S(x_1,\ldots,x_l,\ldots,x_k)\in S_b^r} P_{(x_1,\ldots,x_l,\ldots,x_k)} \right] \tag{12}$$

The above analysis of k-D Markov model can also be applied to MBA for deducing its utilized bandwidth and session blocking probability.

4.2. Derivation for IBS algorithm

Shin *et al.* [26] model the VBR voice traffic as two-state Markov chains, shown in Figure 5. The state 'ON' means there are packets to be transmitted, while the state 'OFF' means there is no packet to be transmitted. For other types of VBR flows, we can still use the ON/OFF model to approximate it. For example in Figure 6, the irregular curve shows the traffic generated by the VBR flow every second. Discretizing the irregular curve, the ON/OFF model (shown in the grey

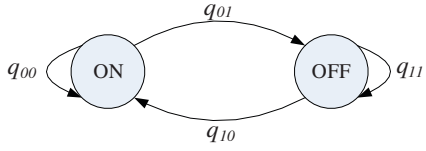


Figure 5. State transition diagram for the ON/OFF model.

DOI: 10.1002/dac

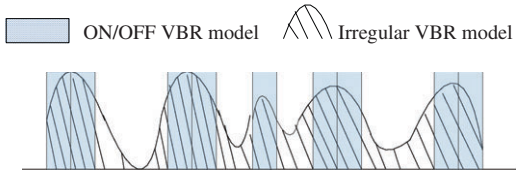


Figure 6. Bandwidth utilization of VBR flows.

pattern) can approximate the irregular curve such that the bandwidths used (i.e. the grey pattern and fringe pattern) are the same for a period of time. Let $\mathbf{Q}$ be the transition probability matrix of the ON/OFF Markov model shown in Figure 5, and $\mathbf{Q}=\begin{pmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{pmatrix}$.

Let $\boldsymbol{v}=[v_0, v_1]$ be the vector representing the steady-state probability of the state 'ON' and 'OFF'. Given the initial state $\boldsymbol{v}(0)=[1, 0]$, we have:

$$\boldsymbol{v}=\lim_{n\to\infty}\boldsymbol{v}(n)=\lim_{n\to\infty}\boldsymbol{v}(0)\mathbf{Q}^n. \tag{13}$$

Assume a session requesting l media flows can belong to one of m services, $s_{l,1}, \ldots, s_{l,m}$ and the proportion of $s_{l,i}$ is $p_{s_{l,i}}$. Let $\lambda_{s_{l,i}}$ be the arrival rate of sessions belonging to service $s_{l,i}$, which can be calculated as: $\lambda_l p_{s_{l,i}}$. For service $s_{l,i}$, the number of VBR flows is $n_{v,s_{l,i}}$ and the number of non-real-time flows is $n_{nr,s_{l,i}}$.

Let λ_{idle} be the rate of idle bandwidth generated by all sessions, which can be estimated as:

$$\lambda_{\text{idle}}=v_1\sum_{l=1}^{k}\sum_{i=1}^{m}\lambda_l p_{s_{l,i}} n_{v,s_{i,l}} \tag{14}$$

Let λ'_l be the arrival rate of sessions that need to request l units of bandwidth. We use the following algorithm to obtain all $\lambda'_l (1\leqslant l\leqslant k)$.

```
1: Set all the λ'_l to 0
2: Calculate λ_idle according to Equation (14)
3: For l = k to 1
4:   For each service s_{l,i} in sessions containing l flows
5:     If n_{nr,s_{l,i}} > 0 and λ_idle > 0
6:       If λ_{s_{l,i}} > λ_idle / n_{nr,s_{l,i}}
7:         λ'_l = λ'_l + (λ_{s_{l,i}} − λ_idle / n_{nr,s_{l,i}})
8:         λ'_{(l−n_{nr,s_{l,i}})} = λ_{(l−n_{nr,s_{l,i}})} + λ_idle / n_{nr,s_{l,i}}
9:         λ_idle = 0
10:        Else
11:          λ'_{(l−n_{nr,s_{l,i}})} = λ_{(l−n_{nr,s_{l,i}})} + λ_{s_{l,i}}
12:          λ_idle = λ_idle − n_{nr,s_{l,i}} λ_{l,s_i}
13:    End If
14:  Else λ'_l = λ'_l + λ_{s_{l,i}}
15:  End If
16:   End For
17: End For
```

This algorithm first initializes all λ'_l to be 0 (line 1) and computes the available idle bandwidth λ_{idle} (line 2). Then, it works by processing the sessions with the number of media flows from k to 1. For the session with the number of media flows as l, each of its service $s_{l,i}$ is processed as follows (lines 5–15): if its number of non-real-time flow (i.e. $n_{v,s_{i,l}}$) is 0 or λ_{idle} is 0 (line 14), it means that this service needs to request l units of bandwidth and thus λ'_l is increased by $\lambda_{s_{l,i}}$ (i.e. this service's arrival rate). Otherwise, it first checks whether current idle bandwidth λ_{idle} can serve

DOI: 10.1002/dac

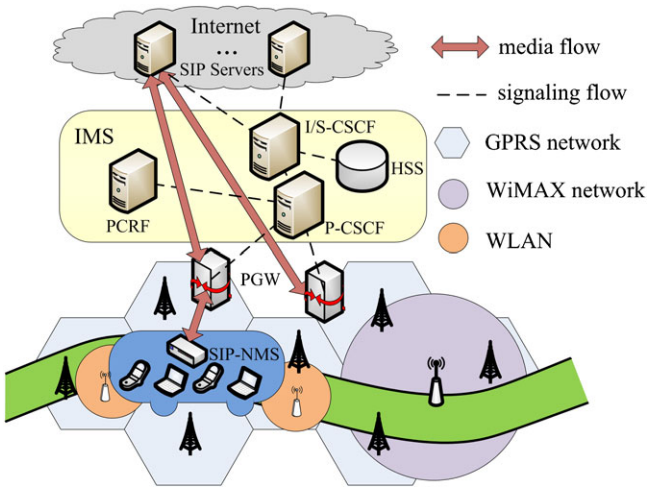


Figure 7. Simulation Scenario.

the non-real-time flows of all its sessions (line 6). If so (lines 11–12), it means that this service needs to request $l - n_{nr,s_{l,i}}$ units of bandwidth and thus $\lambda'_{(l-n_{nr,s_{l,i}})}$ is increased by $\lambda_{s_{l,i}}$ and λ_{idle} is decreased by $n_{nr,s_{l,i}}\lambda_{s_{l,i}}$; otherwise, the idle bandwidth can only serve the non-real-time flows of $\lambda_{\text{idle}}/n_{nr,s_{l,i}}$ sessions, and for the other $\lambda_{s_{l,i}} - \lambda_{\text{idle}}/n_{nr,s_{l,i}}$ sessions, each still needs to request l units of bandwidth, thus $\lambda_{s_{l,i}}$ is decreased by $\lambda_{\text{idle}}/n_{nr,s_{l,i}}$, $\lambda'_{(l-n_{nr,s_{l,i}})}$ is increased by $\lambda_{\text{idle}}/n_{nr,s_{l,i}}$ and λ_{idle} is set to 0 (lines 7–9).

After obtaining all $\lambda'_l (1 \leqslant l \leqslant k)$, we can update Steady-State Balance Equation (8) by replacing λ_l with λ'_l, and use the similar method as Section 4.1 to solve the Steady-State Balance Equations. In this way, m_b and p_b for HBS with IBS algorithm can be estimated from Equations (10) and (12), respectively.

4.3. Simulation validation

In order to validate the above analytical modeling, we have developed a Java-based test bed to simulate the IMS-NEMO, which includes several discrete-event simulators, IMS servers and the implementation of HBS. The IMS network entities, SIP clients and SIP-NMS are all developed based on the open source code SIP stack SIPp [27]. The discrete-event simulators simulate the session traffics and mobility behaviors of a vehicle moving along the road [28]. Figure 7 depicts the simulation scenario for IMS-NEMO. The MN (i.e. a vehicle) that is managed by a SIP-NMS can roam among three access networks, and connect to the Internet through IMS. Ten MTs are placed inside the vehicle. Four SIP servers are configured in the Internet. General Packet Radio Service (GPRS) network, WLAN and WiMAX network are connected to IMS with bandwidth set to be 240 kb/s, 1 Mb/s and 1.25 Mb/s, respectively The GPRS network covers the whole IMS domain. The radius of hotspot area provided by WLAN is set to be 1 km. The covered distance served by a wireless base station of WiMAX network is set to be 50 km. The vehicle's speed varies according to its type, e.g. a bus's speed can reach 100 km/h and a train's speed can reach 300 km/h. We will discuss the impact of different mobility parameters on the performance of HBS in the following tests.

Our simulation adopts a more realistic model to simulate the sessions and their bandwidth utilizations. Four services are developed: (1) multimedia conference with 2 VBR flows and 1 non-real-time flow; (2) VoIP service with 1 VBR flow and 1 CBR flow; (3) VoD service with 1 VBR flow and (4) FTP service with 1 non-real-time flow. Sessions arrive with an intensity of 0.133 sessions per second. The proportion of service (1), (2), (3) and (4) is 1/3, 1/3, 1/6 and 1/6, respectively. The holding time of all sessions except FTP ones follows an exponential distribution with the mean of 60 s. The single unit of bandwidth requested by each media flow

DOI: 10.1002/dac

Table II. Simulation parameters.

Parameters	Values
The number of SIP-NMS (MR) in the MN	1
The number of MTs in the MN	10
λ: session arrival rate	0.133 sessions per second
μ: session holding time	60 s
The bandwidth of GPRS network	240 kb/s
The bandwidth of WiMAX network	1.25 Mb/s
The bandwidth of WLAN	1 Mb/s
The covered distance of WiMAX network	50 km
The covered distance of WLAN	1 km
v: the vehicle speed	10 km/h, 70 km/h or 200 km/h
The packets size of non-real-time flows	1 kb
K: the parameter of Pareto distribution	10/9
B_u: the size of bandwidth a media flow requests	30 kb/s
T: the period for idle bandwidth aggregation in IBS algorithm	0.25 s

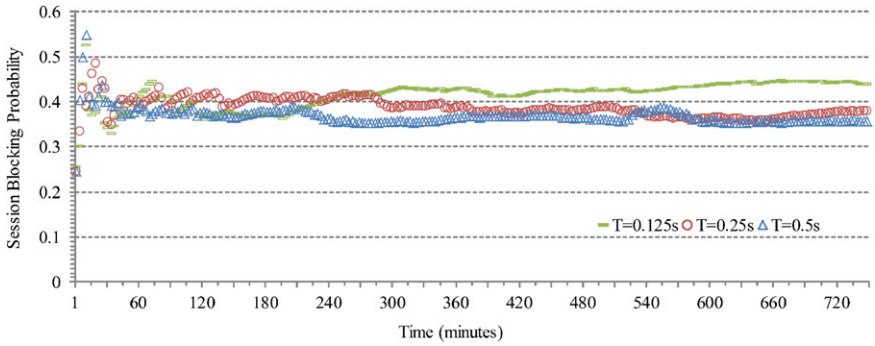


Figure 8. The impact of parameter T on the performance of IBS.

is 30 kbit/s. The packets size for FTP session with only non-real-time flow is 1000 bytes. The holding time of FTP session depends on the provided bandwidth and its traffic, which follows a Pareto distribution with the parameter K. Table II lists the values of parameters used in our simulations.

During the whole simulation period, a vehicle travels in the IMS domain which is covered by GPRS network. Along its traveling path, the vehicle may enter the overlapped area of GPRS and another access network such as WLAN or WiMAX. After SIP-NMS attaches to a new access network, part of the new coming sessions are allocated to the new access network by the Network Selector. If a new access network is assigned to serve a session, SIP-NMS sets up its bearer to PGW according to the calculated QoS profiles. When the vehicle leaves the overlapped area of GPRS and the other network, SIP-NMS re-establishes the bearer to PGW via GPRS network.

First, we discuss how to set the appropriate value of parameter T for IBS algorithm. We ran the simulations with different T values when $v = 200$ km/h (i.e. MN only accesses to GPRS network), and found that the value of T impacts the performance of IBS algorithm significantly, e.g. session blocking probabilities increase when T becomes smaller, as shown in Figure 8. Obviously, the idle bandwidth aggregated from the VBR flows in each T period should be large enough to transmit at least one non-real-time packet. With a small value of T, the non-real-time packet cannot be fit into the aggregated idle bandwidth. With the increase of T, the idle bandwidth aggregated from VBR flows increase. However, the packet delays for real-time flows are also increased. Thus, T cannot be set too large. In the following experiments, we set T as 0.25 s.

Figure 9 shows the session blocking probabilities during 600 min with two different speeds of the vehicle. The areas where the vehicle travels are as follows: (1) the GPRS network; (2) the overlapped area of WLAN and GPRS; (3) the GPRS network; (4) the overlapped area of WLAN and

DOI: 10.1002/dac

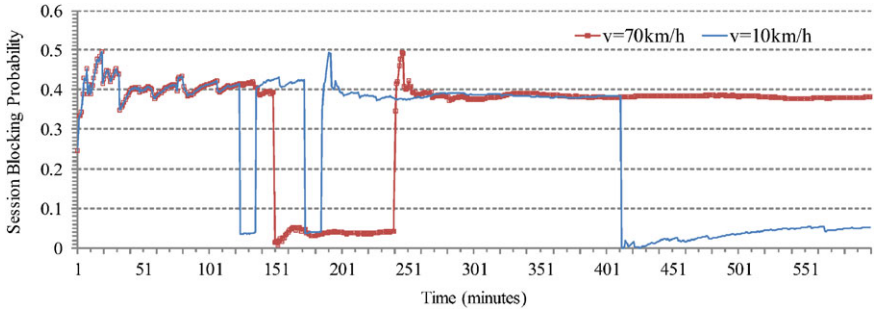


Figure 9. Simulation results under various speeds.

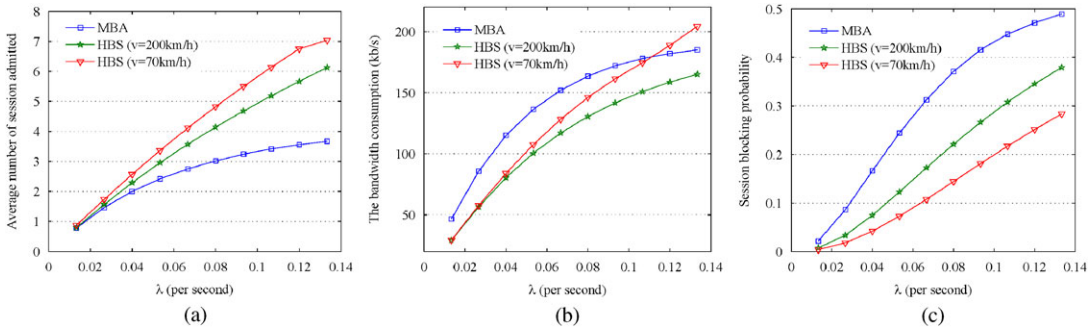


Figure 10. Performance metrics vs λ.

GPRS; (5) the GPRS network; (6) the overlapped area of WiMAX and GPRS (when $v = 10$ km/h, the vehicle terminates in this area at the 600th minute) and (7) the GPRS network. From Figure 9, we can see that the curve of $v = 70$ km/h changes acutely two times. When the vehicle can only access the GPRS network, the session blocking probability is about 0.40 because the provided bandwidth cannot serve all the incoming sessions. After 150 min, the session blocking probability decreases to 0.04, because the vehicle enters the overlapped area of WiMAX and GPRS and HBS's multi-access network selection enables the MN to utilize the bandwidth provided by the two access networks. After 250 min, the session blocking probability increases again because the vehicle can only access the GPRS network. Note that the vehicle cannot use the bandwidth of WLAN due to its high speed when it travels in the overlapped area of WLAN and GPRS network, which shows in Figure 9 that the session block probabilities of the first 150 min are similar to those after 250 min.

The MN's mobility has impact on the performance of HBS. For example, when the vehicle's speed is 10 km/h, we can see in Figure 9 that the session blocking probability decreases to about 0.04 in the period between the 123th and the 135th minute, and the period between the 167th and the 179th minute. During these periods, the vehicle is moving in the overlapped area of GPRS and WLAN. As HBS's multi-access network selection enables some of the sessions to access WLAN, the session blocking probability decreases.

Finally, we compare the simulation results with our analytical ones when $v = 70$ km/h. The session blocking probability obtained from the analytical model is 0.381 when the vehicle can only access the GPRS network. On the other hand, the analytical session blocking probability is 0.037 when the vehicle can access both the GPRS and WiMAX network, because in this case the maximum number of bandwidth units provided by GPRS and WiMAX network is 48. Therefore, Figure 9 shows that session blocking probability obtained from the simulation experiments is similar to that from the analytical model. The comparison results for the bandwidth consumption are similar and will not be presented in this paper.

DOI: 10.1002/dac

5. PERFORMANCE EVALUATION

In this section we evaluate the performance of HBS by comparing it with MBA in terms of: (1) the average number of admitted sessions; (2) the bandwidth consumption and (3) the session blocking probability. Also, we investigate the parameters of VBR flows on the performance of HBS. The vehicular traveling path is the same as Figure 7. We use 4 services described in Section 4.3, and change the proportion of each service according to different scenarios. The values of simulation parameters are similar to these in Table II, except the value of λ and v.

5.1. Comparison between HBS and MBA

To evaluate the multi-access network selection policy and IBS algorithm, we consider two different cases: one is that MN's speed is 200 km/h; and the other is that the MN's speed is 70 km/h. In the first case, the MN with HBS scheme can only utilize the bandwidth of a single access network, i.e. GPRS. In the second case, when the MN travels in the overlapped area of GPRS and WLAN, HBS scheme only allows the MN to access GPRS network. However, when it travels in the overlapped area of GPRS and WiMAX, HBS scheme allows the MN to access both networks at the same time. Let η be the ratio of resident time that the MN is inside and outside the overlapped area of GPRS and WiMAX network in the case of $v = 70$ km/h. Note that the performance of MBA is not impacted by the MN's speed because it can only use the GPRS network even in the overlapped area of GPRS and the other network.

In the first study, we vary the session arrival rates to compare the performance of MBA and HBS. Let $p_1 = p_2 = p_3 = 1/3$, $q_{00} = q_{11} = 0.5$, $\mu = 60$ s and $\eta = 1$. Figure 10(a) shows that the number of sessions admitted grows with the increase of λ. In case of $v = 200$ km/h, HBS performs better than MBA in terms of the number of sessions admitted, because HBS utilizes the idle bandwidth to serve non-real-time flows. Also, we can see that HBS at the speed of 70 km/h performs better than MBA and HBS at the speed of 200 km/h. The reason is that when the MN is traveling at 70 km/h in the overlapped area of GPRS and WiMAX, HBS can make use of both networks to serve the sessions. With more sessions arriving at the MN, the number of sessions admitted is increased.

From Figure 10(b), we observe that the bandwidth consumption increases as λ increases. In the case of $v = 70$ km/h, all the arrival sessions can be set up successfully by HBS when MN is in the overlapped area of GPRS and WiMAX, so the bandwidth consumption increases linearly with the session arrival rate of MN. When λ is less than 0.11, the bandwidth consumption of HBS with $v = 70$ km/h is smaller than that of MBA, because the non-real-time flows in HBS can take use of the idle bandwidth. But when λ is larger than 0.11, the bandwidth consumption of HBS with $v = 70$ km/h is higher than that of MBA, because HBS enables the MN to admit more sessions than MBA. In addition, when the MN can only access one network ($v = 200$ km/h), the bandwidth consumption of HBS is less than that of MBA due to its idle bandwidth sharing mechanism.

From Figure 10(c), we observe that HBS at $v = 70$ km/h has the lowest session blocking probability, because in the overlapped area of GPRS and WiMAX, the available bandwidth provided by these two networks can be used by HBS, which is sufficient to serve all the sessions. On the other hand, the blocking probability of HBS at $v = 200$ km/h is less than that of MBA. The reason is that when the MN accesses only one network, HBS can use the idle bandwidth to serve the non-real-time flows, and thus its blocking probability is smaller than that of MBA.

Now, we vary the proportion of a session containing three media flows (p_3) to further analyze the performance metrics. We set $p_1 = p_2 = (1 - p_3)/2$, $q_{00} = q_{11} = 0.5$, $\lambda = 0.067$ sessions per second, $\mu = 60$ s and $\eta = 1$. In Figure 11(a), we can see that when p_3 increases, the number of sessions admitted decreases, because more bandwidths are needed when the number of sessions with three flows is increased. On the average, HBS scheme can admit more sessions than MBA scheme. However, for HBS at the speed of 70 km/h, the increasing of p_3 has no effect on the number of sessions admitted, because in the overlapped area of GPRS and WiMAX, the available bandwidths provided by these two networks are sufficient to serve almost all arrival sessions.

DOI: 10.1002/dac

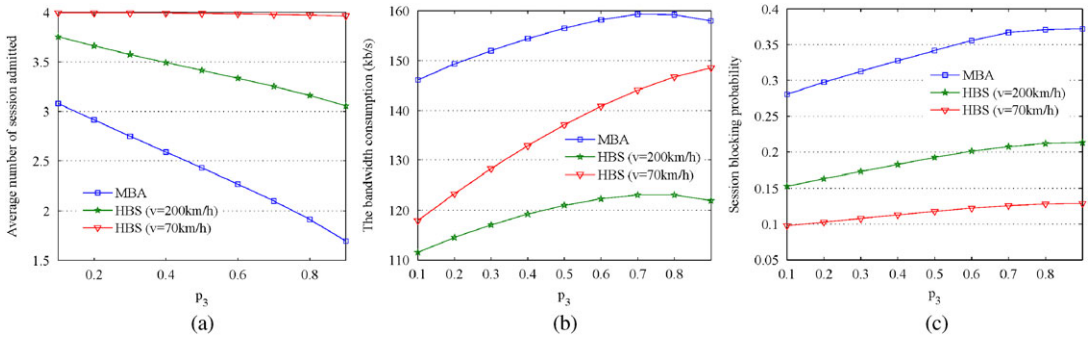


Figure 11. Performance metrics vs p_3.

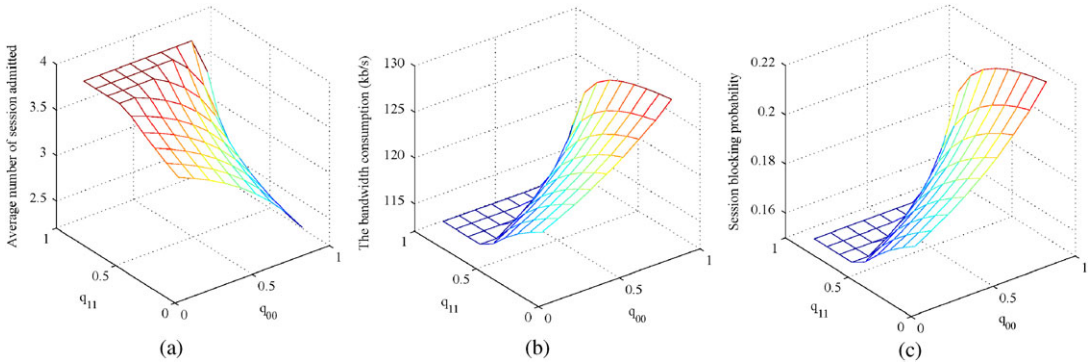


Figure 12. Performance metrics vs q_{00} and q_{11}.

In Figure 11(b), when the MN is moving at the speed of 70 km/h, the bandwidth consumption of HBS grows almost linearly with the increase of p_3. We can see that HBS utilizes less bandwidth than MBA due to its idle bandwidth sharing mechanism. But HBS at $v = 70$ km/h utilizes more bandwidth than HBS at $v = 200$ km/h, because in the overlapped area of GPRS and WiMAX, MN can admit more sessions, which results in the increase of bandwidth consumption. However, when p_3 is higher than 0.7, the bandwidth consumption of both MBA and HBS at $v = 200$ km/h decreases. As most of the sessions require three units of bandwidth, the number of blocked sessions increases, which results in the decrease of bandwidth consumption.

In Figure 11(c), we can see that MBA has the highest blocking probability. The higher p_3, the more bandwidth required by the MN. Thus, when the available bandwidths provided by the access networks are smaller, the session blocking probabilities are higher with the increase of p_3. The blocking probability is lowest for HBS at $v = 70$ km/h, because in the overlapped area of GPRS and WiMAX network, it can exploit the bandwidth of these two networks to serve all the new sessions.

5.2. *Impact of VBR parameters on HBS performance*

Finally, we study the impact of the transition probabilities in the ON/OFF model (q_{00} and q_{11}) on the performance of HBS. Let $p_1 = p_2 = p_3 = 1/3$, $\lambda = 0.067$ sessions per second, $\mu = 60$ s and $v = 200$ km/h. From Figure 12, we can see that the amount of bandwidth saved is especially evident when q_{00} is small and q_{11} is large, because VBR can provide higher idle bandwidth and thus more non-real-time session can be served. Along with the decrease of q_{00} and the increase of q_{11}, the average number of admitted sessions increases (shown in Figure 12(a)); the bandwidth consumption decreases (shown in Figure 12(b)) and the session blocking probability decreases (shown in Figure 12(c)).

DOI: 10.1002/dac

6. CONCLUSION

This paper proposes the HBS scheme to support the QoS guarantee for IMS-NEMO in heterogeneous 4G networks. The QoS control architecture and related QoS signaling for IMS-NEMO are presented for the first time. The establishment of entire bearer for the MN can reduce the cost of QoS signaling when the sessions arrive and terminate frequently. The multi-access network selection policy enables the MN to select the optimal access network for each session. The IBS algorithm reduces the bandwidth utilization by transmitting the non-real-time data via the idle bandwidth reserved for the VBR flows. The theoretical analysis and experimental simulation prove that the HBS scheme can satisfy users' QoS requirement and obtain a more efficient use of wireless bandwidth, compared with the current QoS schemes.

ACKNOWLEDGEMENTS

This work was jointly supported by: (1) National Science Fund for Distinguished Young Scholars (No. 60525110); (2) National 973 Program (Nos. 2007CB307100, 2007CB307103); (3) National Natural Science Foundation of China (No. 60902051); (4) the Fundamental Research Funds for the Central Universities (BUPT2009RC0505); (5) Development Fund Project for Electronic and Information Industry (Mobile Service and Application System Based on 3G); (6) On Open Multi-plan Framework and Resources Reconstitution Theory of Service Networks (No. 61072057); (7) MICINN (No. TIN2010-19077) and CAM (No. S2009TIC-1692).

REFERENCES

1. Hui SY, Yeung KH. Challenges in the migration to 4G mobile systems. *IEEE Communication Magazine* 2003; **41**:54–59.
2. Wang JY, Liao JX, Zhu XM. Latent handover: a flow-oriented progressive handover mechanism. *Computer Communications* 2008; **31**:2319–2340.
3. Johnson D, Perkins C, Arkko J. Mobility support in IPv6. *IETF, RFC 3775*, June 2004.
4. Malik MA, Kanhere SS, Hassan M, Benatallah B. On-board RSVP: an extension of RSVP to support real-time services in on-board IP networks. *Springer LNCS Series* 2004; **3326**:264–275.
5. Tlais M, Labiod H. Resource reservation for NEMO networks. *International Conference on Wireless Networks, Communications and Mobile Computing*, Kaanapali Beach, HI, 2005; 232–237.
6. Wang J-T, Hsu Y-Y, Tseng C-C. A mobile bandwidth-aggregation reservation scheme for NEMOs. *Wireless Personal Communications* 2008; **44**:383–401.
7. Kamel G, Mihailovic A, Hamid Aghvami A. Case analysis of a cost-optimal QoS aggregation policy for network mobility. *IEEE Communication Letter* 2008; **12**:130–132.
8. Kamel G, Mihailovic A, Pangalos P, Hamid Aghvami A. Cost-optimal QoS aggregation for network mobility. *IEEE Global Telecommunications Conference*, Washington, DC, U.S.A., November 2007; 5006–5010.
9. 3GPP TS 23.107 V8.0.0, *QoS Concept and Architecture*. December 2008.
10. Huang C-M, Lee C-H, Zheng J-R. A novel SIP-based route optimization for network mobility. *IEEE Journal on Selected Areas in Communications* 2006; **24**:1682–1691.
11. Chiang W-K, Ren A-N, Chung Y-C. Integrating SIP-based network mobility into IP multimedia subsystem. *IEEE Wireless Communications and Networking Conference*, Budapest, Hungary, April 2009; 1–6.
12. Chang I-C, Chou C-H. HCoP-B: a hierarchical care-of prefix with BUT scheme for nested mobile networks. *IEEE Transactions on Vehicular Technology* 2009; **58**:2942–2965.
13. Petander H, Perera E, Lan K-C, Seneviratne A. Measuring and improving the performance of network mobility management in IPv6 networks. *IEEE Journal on Selected Areas in Communications* 2006; **24**: 1671–1681.
14. Sazzadur Rahman Md, Bouidel O, Atiquzzaman M, Ivancic W. Performance comparison between NEMO BSP and SINEMO. *IEEE Global Telecommunications Conference*, Washington, DC, U.S.A., November 2007; 2398–2402.
15. Yousaf FZ, Tigyo A, Wietfeld C. NERON: a route optimization scheme for nested mobile networks. *IEEE Wireless Communications and Networking Conference*, Budapest, Hungary, April 2009; 1–6.
16. Lim H-J, Kim M, Lee J-H, Chung TM. Route optimization in nested NEMO classification, evaluation and analysis from NEMO fringe stub perspective. *IEEE Transactions on Mobile Computing* 2009; **8**: 1554–1572.
17. Zafar A, Shahriar M, Atiquzzaman M, William I. Route optimization in network mobility: solutions, classification, comparison, and future research directions. *IEEE Communications Surveys and Tutorials* 2010; **12**: 24–38.

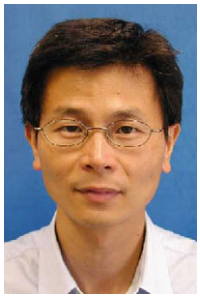
DOI: 10.1002/dac

18. Pack S, (Sherman) Shen X, Mark JW, Pan J. Mobility management in mobile hotspots with heterogeneous multihop wireless links. *IEEE Communication Magazine* 2007; **45**:106–112.
19. Tseng Y-C, Chen J-J, Cheng Y-L. Design and implementation of a SIP-based mobile and vehicular wireless network with push mechanism. *IEEE Transactions on Vehicular Technology* 2007; **56**:3408–3420.
20. YieLeu F. A novel network mobility handover scheme using SIP and SCTP for multimedia applications. *Journal of Network and Computer Applications* 2009; **32**:1073–1091.
21. Ali I, Casati A, Chowdhury K, Nishida K, Schmid S, Vaidya R. Network-based mobility management in the evolved 3GPP core network. *IEEE Communication Magazine* 2009; **47**:58–66.
22. Nguyen-Vuong Q-T, Agoulmine N, Ghamri-Doudane Y. A user-centric and context-aware solution to interface management and access network selection in heterogeneous wireless environments. *Computer Networks* 2008; **52**:3358–3372.
23. Jamalipoura A, Lorenz P. End-to-end QoS support for IP and multimedia traffic in heterogeneous mobile networks. *Computer Communications* 2006; **29**:671–682.
24. Parekh A, Gallager R. A generalized processor sharing approach to flow control in integrated services networks: the single code case. *IEEE/ACM Transactions on Networking* 1993; **1**:344–357.
25. Gelabert X, Pérez-Romero J, Sallent O, Agustí R. A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks. *IEEE Transactions on Mobile Computing* 2008; **7**:1257–1270.
26. Shin S, Schulzrinne H. Measurement and analysis of the VoIP capacity in IEEE 802.11 WLAN. *IEEE Transactions on Mobile Computing* 2009; **8**:1265–1279.
27. SIPp, (Available from: http://sipp.sourceforge.net/).
28. Yang S-R, Chen W-T. SIP multicast-based mobile quality-of-service support over heterogeneous IP multimedia subsystems. *IEEE Transactions on Mobile Computing* 2008; **7**:1297–1310.

AUTHORS' BIOGRAPHIES

Jianxin Liao was born in 1965, obtained his SB, SM and PhD degree at the University of Electronics Science and Technology of China in 1985, 1991 and 1996, respectively. He is currently the dean of Network Intelligence Research Center and the full professor of State Key laboratory of Networking and Switching Technology in Beijing University of Posts and Telecommunications. He has published hundreds of research papers and several books. He has won a number of prizes in China for his research achievements, which include the Premier's Award of Distinguished Young Scientists from National Natural Science Foundation of China in 2005, and the Specially invited Professor of the 'Yangtse River Scholar Award Program' by the Ministry of Education in 2009. His main research interests include mobile intelligent network, service network intelligent, networking architectures and protocols and multimedia communication.

Qi Qi was born in 1982, obtained her PhD degree from Beijing University of Posts and Telecommunications in 2010. Now she is an assistant professor in Beijing University of Posts and Telecommunications, China. Her research interests include performance evaluation for mobility management and future Internet, IP multimedia subsystem, Ubiquitous services, QoS and multimedia communication.

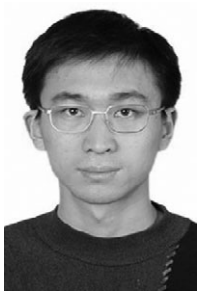


Tonghong Li was born in 1968, obtained his PhD degree from Beijing University of Posts and Telecommunications in 1999. He is currently an assistant professor with the Department of Computer Science, Technical University of Madrid, Spain. His main research interests include resource management, distributed system, middleware, wireless networks and sensor networks.

DOI: 10.1002/dac

Yufei Cao was born in 1974, obtained his PhD degree from Beijing University of Posts and Telecommunications in 2008. He joined the EBUPT Information Technology Company, China, in 2008, and is currently working as a Research Engineer. His research interests include SIP protocol, communications software, Next Generation Network and IP multimedia subsystem.

Xiaomin Zhu was born in 1974, obtained his PhD degree from Beijing University of Posts and Telecommunications in 2001. Now he is an associate professor in Beijing University of Posts and Telecommunications. His major is Telecommunications and Information Systems. His research interests span the area of intelligent networks and next-generation networks with a focus on 3G core network and protocol conversion. He has published over 110 papers, among which there are more than 30 first-authored ones, in different journals and conferences. His personal home page is http://zhuxm.ik8.com.



Jingyu Wang was born in 1978, obtained his PhD degree from Beijing University of Posts and Telecommunications in 2008. Now he is an assistant professor in Beijing University of Posts and Telecommunications, China. His research interests span broad aspects of performance evaluation for Internet and overlay network, traffic engineering, image/video coding, multimedia communication over wireless network.

DOI: 10.1002/dac