Abstract: Intra-Domain Mobility Management Protocol (IDMP) has been recently proposed as a protocol for managing IP mobility within a cellular access network. This paper investigates the scalability performance of IDMP’s Quality of Service (QoS) framework, which uses a modified form of the Differentiated Services architecture, with a centralized Bandwidth Broker (BB) performing admission control and resource provisioning for different traffic classes. The theoretical analysis show that requests for bandwidth reservation due to intra-domain mobility should be controlled to alleviate the processing burden at the BB. Accordingly, we propose a scalable bandwidth reservation scheme. By reserving the bandwidth in trunk instead of on per-host basis, QoS-related signaling load and handoff latency can be reduced significantly.

I. INTRODUCTION

With the advancement of wireless access technologies and emergence of numerous multimedia applications over Internet, mobile users require not only the seamless mobility support but also certain level of Quality of Service (QoS) guarantee. To improve the performance of Mobile IP [1] on handling the fast moving mobile hosts within a local administrative domain, various micro-mobility (or intra-domain) protocols (e.g., Cellular IP [2], HAWAII [3]) have been proposed. However, their objective is mainly on reducing the handoff latency and location update cost by hiding the mobility of mobile hosts within the hierarchical local access domain, while there is no any integrated form of QoS support.

In this paper, we describe the development of such an integrated QoS support for the Intra-Domain Mobility Management Protocol (IDMP), which has recently been proposed [4] as alternative two-level hierarchical mobility management solution. IDMP was developed as a standalone intra-domain mobility protocol, which could support multiple global mobility protocols, such as Mobile IP [1] or Session Initiation Protocol (SIP) [5]. In IDMP, each mobile node (MN) is assigned two distinct transient care-of addresses. The global care-of address (GCoA) is associated with a specialized node called the Mobility Agent (MA), which provides a stable gateway to the Internet. Subnet Agents (SA) located at the wireless access edge provides the MN an additional local care-of address (LCoA) to handle intra-domain mobility. The fundamental architectural design of the IDMP’s QoS framework was presented in [6], which uses a centralized Bandwidth Broker (BB) to perform admission control and resource (bandwidth) provisioning for different traffic classes within the access network.

To convey the extra information for QoS negotiation between different entities, the QoS messaging under the IDMP-based framework increases the signaling cost compared with other micro-mobility protocols without QoS support [7], although it is nominal in the bandwidth-limited wireless links. This paper analyzes the factors contributing to the handoff latency and signaling load due to QoS-aware intra-domain mobility under the IDMP-based QoS framework. It is observed that the request sent to BB for bandwidth reservation should be controlled to alleviate the processing burden at BB. Accordingly, we propose a scalable bandwidth reservation scheme. By reserving the bandwidth in trunk instead of on per-host basis, QoS-related signaling load and handoff latency can be reduced significantly.

The rest of the paper is organized as follows. In the next section, an overview of the IDMP-based QoS framework is presented. In section III we analyze the factors contributing to the handoff latency and signaling overhead due to QoS-aware intra-domain mobility. Accordingly, in section IV, a scalable trunk resource reservation scheme is proposed. The scalability performance of the IDMP-based QoS framework has been investigated, which validates the effectiveness of the proposed trunk resource reservation scheme. Finally, section V concludes the paper.

II. IDMP-BASED QoS FRAMEWORK

The functional layout of the IDMP-based QoS framework is illustrated in Fig. 1. The architecture is based on the Differentiated Services (DiffServ) framework, with all traffic being classified into a relatively small number of QoS classes. From the DiffServ standpoint, the entire access network is treated as a separate domain, which peers with additional DiffServ (or non-DiffServ) domains in the core network or over the wireless link. Appropriate traffic filtering and conditioning is applied at the SA and MA nodes, which effectively serve as domain ingress and egress points for upstream (mobile hosts to network) traffic; for downstream traffic, the SA and MA essentially switch roles. This framework introduces two new QoS-specific entities, the Mobility Server (MS), which implements the load balancing algorithm that dynamically assigns one or more mobility agents to an MN, and the Bandwidth Broker (BB), which dynamically modifies network resources reserved for different traffic classes at each intermediate domain node.
In simple terms, the QoS functionality involves the following fundamental steps:

1. The MN must first specify its QoS requirements during its initial registration with an SA in the domain.
2. The QoS requirements are relayed by the SA to the MS, which uses a load-balancing algorithm and other policies to dynamically assign the MN to a specific MA. When the MN registers with the MA (to obtain a valid GCoA), the MA may subsequently modify the QoS-related parameters.
3. On every subsequent movement of the MN within the domain, the MA is responsible for communicating the MN’s outbound traffic descriptor and classification policies to the new SA—this obviates the need for additional QoS-signaling from the MN at every move.
4. Whenever the MN changes its point of attachment, and the current resources on the new path prove inadequate to support its assured QoS levels, the MA requests the BB to provision additional resources on this new path. (If the BB fails to provision the required additional resources, the network cannot guarantee the MN’s QoS requirements at the new SA, and consequently, the handoff is aborted).

III. SIGNALING ANALYSIS OF QoS-AWARE HANDOFF

In this section, we perform a signaling analysis on the QoS-aware handoff, which means not only that a MN reconnects with the network but also that the QoS negotiation between the MN and the network has completed if necessary. In IDMP, if the reserved resources in the subnet can accommodate the QoS requirements of the entering MN, there is no need for the MA to communicate with the BB for bandwidth re-allocation. The MA can forward directly the QoS profile of the MN to its new attached SA. If using the proactive fast handoff mechanism [8], MA also signals the QoS profile to neighboring SAs (those in the shadow cluster), to minimize the latency associated with intra-domain handoff. However, if the BB has to be consulted, fast handoff does not complete until the BB has done the bandwidth re-allocation. Thus, in this case the proactive fast handoff scheme has no benefit on reducing the handoff latency. Fig. 2 shows the message flows for the QoS-aware handoff in IDMP.

In general, the QoS-aware handoff latency without the proactive fast handoff procedure (except for the initial movement into the domain) always consists of the following components:

1. Subnet registration time $T_{sa}$: the time taken to obtain a new valid LCoA;
2. MA-registration time $T_{ma}$: the time taken to update the MA of the new LCoA;
3. QoS profile update time $T_{ma-sa}$: the time taken by the MA to signal the new SA of the MN’s traffic descriptor and QoS profile.

In addition, if the MA needs to provision additional resources on the path from it to the new SA, the handoff latency will include $T_{hb}$, which represents the additional time needed to complete a new reservation request between the MA and the BB. On the other hand, if the MA uses the fast proactive update procedure to keep neighboring SAs informed about the MN’s traffic descriptor, the MA-registration time and the QoS-profile update time can be eliminated, but only if the MA does not need to issue a new reservation request to the BB for any of the paths to the neighboring SAs. Accordingly, we can express the QoS-aware handoff latency denoted by $T_{ho}$ through Equations (1) and (2), where $P_{0}$ denotes the probability of an MA having to consult the BB.

**Normal Handoff**: $T_{ho} = T_{sa} + T_{ma} + T_{ma-sa} + P_{0}T_{hb}$

**Fast Handoff**: $T_{ho} = T_{sa} + P_{0}(T_{ma} + T_{ma-sa} + T_{hb})$

From Equation (1) and (2), we can observe that the QoS-aware handoff latency, especially for the proactive fast handoff, depends greatly on the probability that the BB has to be consulted for the bandwidth re-allocation. In addition, the BB response delay $T_{hb}$ is a major factor contributing to the handoff delay because the BB needs to query all the MA’s to get the system resource information and its response delay also depends on the response delay of the MAs. In the case of high rate of resource reservation request, the delay could be significant, which has been verified by the experiment results based on our prototype implementation [7].

---

1 We are currently investigating extensions to the IDMP’s QoS framework where the lack of adequate resources on the new path does not necessarily result in handoff failure. Instead, the MN’s QoS profile is re-negotiated to be compatible with the currently available resources.
Fig. 3 shows the response delay of the MA request to the BB as a function of the number of requests/sec. As we can see, the response delay of the BB is fairly large when the number of requests is small, but rises fairly sharply to about 1 sec, as the number of requests increases to about 1/sec. While the exact numbers are obviously implementation-dependent, the figure illustrates an important performance bottleneck that the BB lookup can lead to a significant increase in the handoff delay (which is otherwise around ~150 msecs [7]). Therefore, we would like to keep $P_h$, the probability of BB lookup, fairly small, so that a significant fraction of the handoffs occur with very low latency.

IV. SCALABILITY ANALYSIS OF BB-RELATED SIGNALLING COSTS

In the previous section, we notice that the QoS-aware handoff delay could be large due to the possible BB processing delay. While several studies (e.g., [9,10]) have demonstrated that the BB-based centralized QoS management solution can be effective in smaller domains, we need to understand the performance implications of the IDMP-based QoS architecture in an environment where a significant fraction of the resource reservation requests are generated due to node movement (rather than due to new traffic flows). In the following scalability analysis, we represent the mobility pattern of individual users by a fluid flow model [11,12]. It is assumed the cell area is square-shaped, and the perimeter of a cell is denoted by $L$. MNs are uniformly distributed with a density $\rho$ and move at an average velocity of $v$ in directions that are uniformly distributed over $[0,2\pi]$. The cell-crossing rate denoted by $r_c$ is then

$$r_c = \frac{\rho v L}{\pi} \quad (3)$$

Let the portion of the active MNs be $\alpha$, the active handoff arrival rate denoted by $\lambda_{ha}$ to a subnet is then

$$\lambda_{ha} = \alpha \rho v L \quad (4)$$

Table I represents the various parameters used in our analysis, and the typical values that we have used for performance figures. Note that our parameter choices imply that the average number of mobile users in a single domain (which equals the number of cells per domain $\times$ the area of a single cell $\times$ the density of user, or $N_d(L/4)^2\times\rho) = 5000$, indicating a relatively large size for our candidate domain.

In the IDMP’s QoS framework, the frequency of MA-BB signaling would equal the frequency of cell change over all the active MNs currently located in the access domain if the MA were to reserve resources from the BB on a per-host basis. Since each domain would typically have in the order of tens of thousands of mobile nodes, this frequency can be sufficiently high, leading to significant performance bottlenecks at the BB. To provide the appropriate level of scalability, we propose the trunk resource reservation approach, where an MA requests aggregate bandwidth from the BB, rather than issue individual reservation requests. Issuing trunk requests serves to reduce the handoff latency in two distinct ways:

1. Since Mobility Agents now issue aggregate reservation requests to the BB, the load on the BB exhibits a dramatic drop-off. Fig. 3 then shows that the BB processing latency ($T_{bb}$) is fairly small.
2. By issuing requests for chunks of bandwidth, the MA also reduces the probability ($P_h$) that an individual node movement results in a new request to the BB. An MA would need to issue a new reservation request only when the sum of the current requested resources to a specific SA exceeds the currently provisioned resources.

To analyze the performance implications of trunk reservation, let us assume that the MA makes reservations on a per-class basis, i.e., a MA is only responsible for handling a single traffic class (more details on MA allocation strategies can be found in [6]), and that each member of a particular service class is allocated a fixed quantum of resources (say X Kbps of bandwidth). Accordingly, the trunk size is specified in $S$, the number of users that may be accommodated in the trunk $^2$.

Since our focus is on the signaling load at the BB, without loss of generality, we consider only one class of users and a single MA in the domain. Multiple MAs and classes merely result in the division of the load across the various MAs and classes, but do not significantly affect the overall request rate to the BB. Let the arrival rate of sessions for nodes currently associated with a specific MA in a subnet/cell, covered by a single SA, be modeled as a Poisson process with arrival rate $\lambda_n$, the handoff arrival rate be $\lambda_{ha}$, and the handoff departure rate (per active node) be $\lambda_{hd}$. Let the length of the service time (duration of each individual single session) be exponentially distributed with mean time $1/\mu$. Then, the overall number of users in the SA can be modeled as a birth-death process, with an increase rate in the $i^{th}$ state of $(\lambda_{ha} + \lambda_n)$ and a decrease rate of $i^*(\lambda_{hd}+\mu)$.

A. BB Consultation Probability

To compute the probability, $P_h$, that an individual node movement or call arrival results in a call to the BB, we assume that both bandwidth allocation and bandwidth de-allocation requests have size $S$. Accordingly, $P_h$ equals the probability that the number of users of that class (for a specific MA) is $S$, 2$S$, 3$S$, and so on, up to a maximum $\left\lfloor \frac{m}{S} \right\rfloor *S$, where $m$ denotes the maximum number of individual users that the BB can accommodate for that class. In general, $m$ will depend on the bandwidth allocation and admission control algorithm.

---

2 It is trivial to extend this analysis to the more general case, where the trunk size is specified explicitly in terms of bandwidth, and users are allowed to specify different levels of bandwidth requirements.
Thus, the corresponding M/M/m/m queue is in one of the states \( S, 2S, \cdots \) over the whole domain by multiplying \( \lambda \). We can then compute the total signaling frequency on the BB for this specific SA, by re-provisioning. It is then easy to see that the reservation request rate of size \( S \) to the BB for this specific SA, is then obtained by multiplying \( \frac{\lambda}{\mu} \). QoS extensions) without having to request the BB for resource re-provisioning.

### B. Request Rate of Resource Reservation

The average rate, \( \rho_a \), at which the MA issues new trunk reservation requests of size \( S \) to the BB for this specific SA, is then obtained by multiplying \( \rho_a \) with the average arrival rate of a new call (\( \hat{\lambda}_{ha} + \hat{\lambda}_{hd} \)). Thus,

\[
P_b = \frac{\left[ \sum_{i=0}^{\infty} \left( \frac{\hat{\lambda}_{ha} + \hat{\lambda}_{hd}}{\mu + \hat{\lambda}_{ha}} \right)^i \cdot \frac{1}{k!} \right]}{\sum_{i=0}^{\infty} \left( \frac{\hat{\lambda}_{ha} + \hat{\lambda}_{hd}}{\mu + \hat{\lambda}_{hd}} \right)^i} \cdot \frac{1}{(i-S)!} \tag{5}
\]

Fig. 4 plots the variation of \( P_b \) with \( S \), the size of the trunk reservation request. We can easily see that an increase in the “chunks” of reserved bandwidth allows an MA to respond to a significant fraction of the Intra-domain Update messages (with QoS extensions) without having to request the BB for resource re-provisioning.

#### C. BB-Related Signaling Load

To compute the total BB-related signaling load, we need to multiply this average BB-signaling rate by the various BB related messages that are generated when an MA-BB request is made. This results in the expression in Equation (7), where:

\[
C_{BB} = N_s \cdot (2 \cdot H_{MA-BB} \cdot L_{MA-BB} + 4 \cdot N_i \cdot H_{BB-IR} \cdot L_{BB-IR}) \tag{7}
\]

The multiplicative constant \( N_s \) corresponds to the fact that \( \rho_a \) is specified only on a per-SA basis, the first term corresponds to the round-trip messaging between the MA and the BB, while the second term corresponds to two separate round-trip signaling messages between the BB and all the \( N_i \) intermediate routers, the first one to query the bandwidth consumption status, and the second one to modify the existing reservations. Fig. 6 plots the total signaling overhead related to MA-BB interactions for various values of the trunk reservation size \( S \). As we can see, even a relatively small amount of aggregation in the request sizes results in a reasonably low level of signaling overhead (O(Kbps) on each hop), thus proving that our BB-based architecture can be deployed in IP-based cellular access domains.

#### D. Bandwidth Utilization

Previous analysis confirms that the proposed trunk resource reservation scheme can greatly reduce the frequency of the BB bandwidth re-allocation for a new entering MN, while the portion of over-reserved bandwidth will be wasted unless there are new arriving MNs. The average bandwidth utilization denoted by \( BU \) can be estimated as follows for the evaluation of the efficiency of the trunk reservation:

\[
BU = \sum_{i=1}^{N_s} U_i \cdot P_i \tag{8}
\]

### Table I: Parameters for Analysis of BB Signaling Cost

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_{MA-BB} )</td>
<td>Hops between MA and BB</td>
<td>2</td>
</tr>
<tr>
<td>( H_{BB-IR} )</td>
<td>Average hops between BB and intra-domain router</td>
<td>2</td>
</tr>
<tr>
<td>( N_i )</td>
<td>Number of Subnet Agents in the domain</td>
<td>256</td>
</tr>
<tr>
<td>( N_s )</td>
<td>Number of intermediate routers queried by a BB</td>
<td>2</td>
</tr>
<tr>
<td>( \rho )</td>
<td>Population density</td>
<td>2000/km^2</td>
</tr>
<tr>
<td>( \nu )</td>
<td>Average velocity</td>
<td>20 m/s</td>
</tr>
<tr>
<td>( L )</td>
<td>Perimeter of a cell</td>
<td>400 m</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Portion of active MNs</td>
<td>5%</td>
</tr>
<tr>
<td>( \lambda_i )</td>
<td>Rate of new sessions for idle MNs</td>
<td>6/hour</td>
</tr>
<tr>
<td>( I/\mu )</td>
<td>Average exponentially distributed session duration</td>
<td>5 mins</td>
</tr>
<tr>
<td>( \lambda_{ha} )</td>
<td>Handoff arrival rate</td>
<td>( \alpha \cdot \rho \cdot \nu \cdot L / \pi )</td>
</tr>
<tr>
<td>( \lambda_{hd} )</td>
<td>Handoff departure rate for individual node</td>
<td>( \nu \cdot 16 \pi )</td>
</tr>
<tr>
<td>( L_{MA-BB} )</td>
<td>Length of request/response message between MA and BB</td>
<td>200 byte</td>
</tr>
<tr>
<td>( L_{BB-IR} )</td>
<td>Length of QoS signaling messages between BB and intra-domain routers</td>
<td>250 bytes</td>
</tr>
</tbody>
</table>

**Figure 4. BB Consultation Probability vs. Trunk Reservation Size**

**Figure 5. Reservation Request Arrival Rate vs. Trunk Reservation Size**
where $P_i$ denotes the probability that there are $i$ number of MNs in the system, and $U_i$ denotes the bandwidth utilization when there are $i$ number of MNs in the system, which can be expressed as follows:

$$U_i \begin{cases} 
\frac{i}{(\lfloor \frac{i}{s} \rfloor + 1)s} & \text{for } (i \mod s) \neq 0 \\
1 & \text{for } (i \mod s) = 0
\end{cases}$$  \hspace{1cm} (9)

With Equation (8) and (9) and applying the same set of parameters in Table I, Figure 7 shows the numerical results of bandwidth utilization as well as the BB consultation probability as the function of trunk reservation size. It is as expected that the bandwidth utilization decreases as the trunk reservation size increases because more bandwidth is over-reserved. However, compared with the sharp decrease of the BB consultation probability, the bandwidth utilization decreases smoothly. With relatively little cost of bandwidth wastage, resource reservation with medium trunk size, e.g., $s=6$, can achieve the ultimate objective, i.e., scalable QoS support. More importantly, system bandwidth is saved with the reduced QoS-related signaling. Alternatively, an adaptive resource reservation scheme can be used to comprise the tradeoff between the bandwidth utilization and the system scalability. The reservation trunk size can be adjusted dynamically by the BB based on the request rate of the resource reservation, i.e., the less the rate, the smaller the trunk size.

**E. Consideration of Different Traffic Types**

In the above analysis, we assume the homogeneous traffic type in the system. However, multimedia traffic is common in the future 3G/4G cellular networks. Suppose there are multiple traffic classes, each of which requires the use of different amount bandwidth, the reservation trunk size should be determined not only by the traffic load but also the traffic types to control the QoS signaling for BB consultation. Extending the analysis and results in section IV.A of the case to the multiple traffic classes, we can obtain the optimal configuration at BB for dynamically adjusting the trunk size.

**V. CONCLUSION**

In this paper, the scalability performance of the IDMP-based QoS framework has been investigated. Our performance studies showed that the frequency of resource reservation requests to the BB must be controlled to avoid sharp increases in the response time of BB-related queries. Analysis of the request generation events showed that this frequency (and the

![Figure 6. BB-Related Signaling Load vs. Trunk Reservation Size](image)

Figure 6. BB-Related Signaling Cost vs. BB Consultation Probability

![Figure 7. Tradeoff of Bandwidth Utilization and BB Consultation Probability](image)

Figure 7. Tradeoff of Bandwidth Utilization and BB Consultation Probability

handoff latency) could be unacceptably high, in medium-sized domains, if resource reservations are performed on a per-host basis. Scalability and satisfactory performance can be assured by making resource reservations in terms of “trunks”: even the use of moderately small trunk sizes leads to a sharp decrease in the signaling load on the BB. In our future work, the trunk reservation strategies need to be studied further as discussed in section IV.E. Since different classes have different bandwidth requirements, it is likely that the appropriate trunk size will be not only traffic load-dependent, but also class-dependent, with a smaller size for classes with larger bandwidth requirements.

**REFERENCES**