Adaptive QoS Mechanism for Wireless Mobile Network

KwangSik Kim
Department of Information Sciences,
Tokyo University of Information Sciences, Chiba, Japan
h08001kk@edu.tuis.ac.jp

Shintaro Uno
Department of Graduate Program in System for Intellectual Creation,
Graduate School of Engineering, Kanazawa Institute of Technology, Tokyo, Japan
suno@y4.dion.ne.jp

MooWan Kim
Department of Information Sciences
Tokyo University of Information Sciences, Chiba, Japan
mwkim@rsch.tuis.ac.jp

Received 22 March 2010; Revised 6 May 2010; Accepted 7 June 2010

Wireless mobile multimedia communications have been greatly increased in the number of users, diversity of applications and interface technologies. Wireless mobile networks are being evolved and integrated into IP based core network, so it is necessary to provide sufficient QoS (Quality of Service) mechanism to provide enhanced user's satisfaction. In this paper, we propose a new adaptive QoS mechanism based on utility function borrowed from the field of microeconomics, call setup and handover signaling mechanism integrating QoS and mobility management. Through a simulation, we show that adaptive resource allocation based on user preferences can be realized in the wireless mobile network with some considerations.

Categories and Subject Descriptors: System and Architecture **[Network and Communications]**

General Terms: Wireless Mobile Network, Mobility, Call Setup, Handover

Additional Key Words and Phrases: Adaptive QoS, Utility Function, Resource Allocation

---

1. INTRODUCTION

Wireless mobile networks are being evolved from 3G to 3.9G/4G, integrated into IP core network based on IMS (IP Multimedia Sub-system). In this environment, it is essential to provide sufficient QoS (Quality of Service) mechanism which enables different services to provide enhanced user satisfaction. Currently QoS is being studied in 3GPP [3GPP TS 29.212. 2008-09], but it still deals with only class oriented mechanism. For example, QoS class is defined according to service types, and mapping between QCI (QoS Class Identifier) and DSCP (DiffServ Code Point) is studied [Ludwig et al. 2006]. In the class oriented QoS mechanism, the QoS class is defined in the network operator's view point, and then, QoS mechanism allocates the fixed bandwidths to the user through the mapping between the amount of resource and QoS class belonged to user's service.

On the other hand, according to the user's various requirement, it is a trend that many services are increasingly becoming adaptive and they can adjust their level of quality based on the amount of the available resources, such as MPEG or layered multicast [McCanne et al. 1996; Yoshimura et al. 2001]. From the user's perspective, a new QoS mechanism is needed to comply with this trend, which enables it to allocate network resources according to the user's preference. Namely, it needs adaptive resource allocation based on the differentiated user's preference regarding the available network resources. Also, even in the operator's perspectives, the needs of user differentiation by their rates are being increased to maximize effective use of the limited resources.

In the current class oriented QoS mechanism, however, it is unable to differentiate the user's preference regarding the available network resources. For example, when there are two users: one user places a call to a stock broker for a stock sale, and the other is browsing the web, the user with the financial transaction would be willing to pay a higher price to obtain the same quantity of resource from the network than the user who is browsing the web if available bandwidth resources are preemptive by users. The existing class based QoS mechanism is not able to differentiate the heterogeneity of user's preference to a certain QoS, and treat these two users alike.

In this paper, we propose a new adaptive QoS mechanism providing adaptive resource allocation based on each user's preferences by using the utility function borrowed from the field of microeconomics. The utility function qualifies the value that a user perceives for all possible amount of resources allocated. Above shortcoming in the class based QoS mechanism is the motivation to introduce the utility function to our proposed QoS mechanism. The reason we prefer this approach is the utility function can be utilized as a mean to differentiate the user's preference regarding the available network resources, and it provides a mean to manage an adaptive resource allocation for each user. There are already some studies in which the utility function is effectively adopted in the QoS field such as [Nomura et al. 2001; Yamori and Tanaka 2004] and [Gonzalez and Michael Needham 2004]. [Nomura et al. 2001] studied the relationship between the waiting time and the user's utility through the experiment of subjective evaluation of use's utility. Through the experiment, this paper addresses how to differentiate user's utility

subjectively for the service waiting time which is one of the QoS metrics, but does not deal how to differentiate user's preference to the available network resources. In [Yamori and Tanaka 2004], the quantitative relationship between the guaranteed minimum bandwidth and the user's willingness to pay through the experiment of streaming contents is investigated. This paper addresses how to differentiate user's utility subjectively to the guaranteed bandwidth, which is referred in our approach. It can be referred to define detailed pricing policy on the differentiated user's preferences to the available network resources. Especially in [Gonzalez and Michael Needham 2004], the authors propose a resource allocation algorithm for RAN (Radio Access Network), and priority control method for CN (Core Network) based on utility function. However, this paper does not deal with the mobility management function in the CN, and because of this shortcoming, it does not provide the sufficient adaptive QoS mechanism on the following points.

1) No detailed resource adaptation algorithm considered the flow (i.e. consider both caller and callee side together).

2) No signaling mechanism to integrate QoS and mobility management.

In the proposed mechanism, we have developed a resource adaptation algorithm based on utility function which considers the caller and callee side together. Also, we have developed signaling mechanism to integrate the QoS and mobility management based on the detailed protocols.

The rest of the paper is organized as follows: Chapter 2 describes the basic technologies of the proposal. Chapter 3 presents a new adaptive QoS mechanism with mobility, and a signaling mechanism to integrate QoS and mobility management. Chapter 4 describes the simulation results for the call setup and handover with some considerations. Finally, conclusion is described in Chapter 5.

2. BASIC TECHNOLOGIES

This chapter describes the basic technologies utilized to develop an adaptive QoS mechanism.

2.1 Utility Function

Utility is defined in the field of microeconomics that the level of satisfaction acquired from the consumption of properties such as services or commodities [Varian 1992].

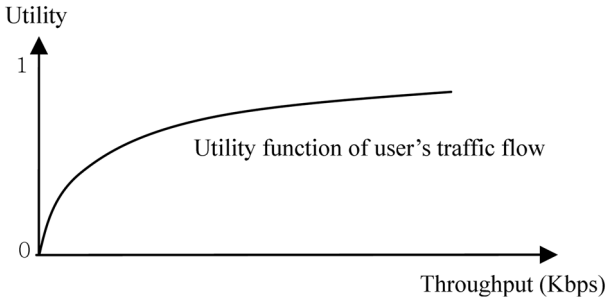

Figure 1. Example of the Utility Function.

The user's total utility obtained from a network service will depend on several QoS metrics, such as throughput, delay, and jitter. In the throughput perspective, the user's utility depends on the bandwidth availability in the network to satisfy the resource requirement of service. Figure 1 shows the example of utility function regarding the throughput allocated to the user [Shenker 1995].

The user's utility, however, shows non-linear characteristics regarding the variation of physical quantity. Let us consider the network situation in which we have M users in the system. We let *Ui* (*ri*) denote the utility derived by user's flow *i* for a bandwidth allocation *ri* and C is the total link capacity. User *i* is allocated with *ri* units of resource that is the i-th component of the solution $r = [r1, r2, \ldots, rM]$ of the following optimization problem [Gonzalez and Michael Needham 2004]:

$$\max_{[r1,r2,\ldots r_M]} \sum_{i=1}^{M} U_i(r_i) \tag{1}$$

$$\text{subject to } \sum_{i=1}^{M} r_i \le C$$

$$r_i \ge 0 \quad \text{for } i = 1, 2, \ldots, M$$

In this network situation, in order to adopt utility function to the network QoS management, following requirements should be satisfied.

1) Continuous utility functions should be represented in discrete functions in order to utilize the utility values as a main parameter of network QoS management.

2) Resource allocation should be done based on discrete utility function satisfying above Equation (1).

Figure 2 shows the example of discrete utility functions.

In the above example, we let $U(r)$ denote the discrete utility functions where r is the amount of resource. It is the utility obtained from that allocation. Followings are the properties of discrete utility functions:

1) Non-negativity: $U(r) \ge 0$ for all $r \ge 0$. Obviously the user cannot associate a negative utility with a positive resource allocation.

2) Non-decreasing nature: U(r) has to be a non-decreasing function. Clearly also is the fact that users cannot associate a higher utility with a smaller allocation.

Also, resource allocation should be done in a way that maximizes aggregation of

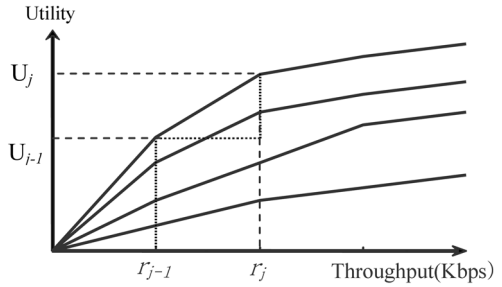

Figure 2. Example of Discrete Utility Function.

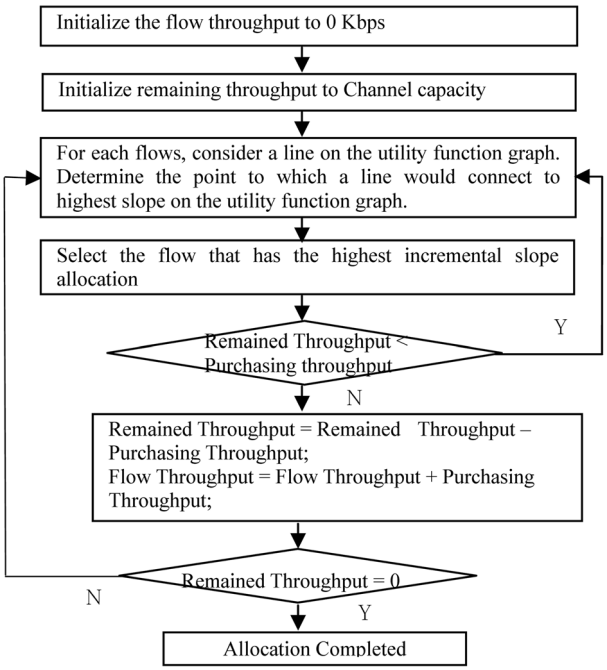

Figure 3. Procedure of the MSS Algorithm.

user's utility based on the discrete utility function which satisfies Equation (1).

By using the discrete utility function to the network QoS management, we can realize that the adaptive resource allocation comply with user's preferences. Details of the resource allocation algorithm will be discussed in the next section.

2.2 MSS Resource Allocation Algorithm

Regarding the adaptive services such as the delay non-sensitive services (e.g. Ftp, Web, E-mail) or delay sensitive real-time services (e.g. VoIP, streaming service), the user's satisfaction will be increased if network resources are enabled to be allocated to the flow adaptively according to the amount of the available resources.

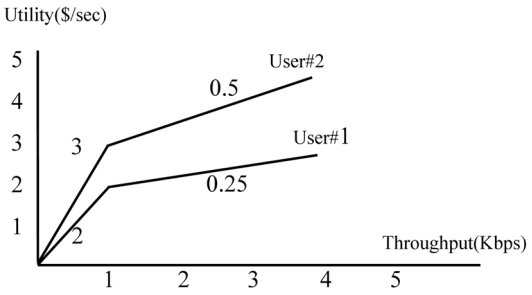

Figure 4. Example of Discrete Utility Function.

Table I. Example of the Resource Allocation by MSS Algorithm.

User	Slope	User's Throughput	Total Allocated	Capacity Left
U2	3	1	1	7
U1	2	1	2	6
U2	0.5	3	5	3
U1	0.25	3	8	0

The MSS (Maximum Segmental Slope) resource allocation algorithm to satisfy the Equation (1) in the previous section has been developed by [Gonzalez and Michael Needham 2004]. The algorithm depicted in Figure 3 is devised to manage the bandwidth resource efficiently among the competing users. The main characteristic of this algorithm is that it allows resource allocation to maximize the user's satisfaction by allocating a unit of resource firstly to the flow that has the highest segmental slope.

A simple numerical example is as follows. Let us assume that we have a channel capacity "C" of 8 units and we have users with utility functions as shown in Figure 4. A channel capacity "C"of 8 units of resources is allocated to each flow based on the MSS algorithm as Table I.

3. ADAPTIVE QOS MECHANISM WITH MOBILITY

In this chapter, we propose an adaptive QoS mechanism with mobility based on the assumed wireless mobile network architectural model.

3.1 Wireless Mobile Network Architectural Model

We assume the conceptual wireless mobile network architectural model as depicted in Figure 5 which comprises RAN (Radio Access Network) with MNs (Mobile Nodes), and CN (Core Network) with two main players: the QS (QoS Server) and the MM (Mobility Manager) [Politis et al. 2004].

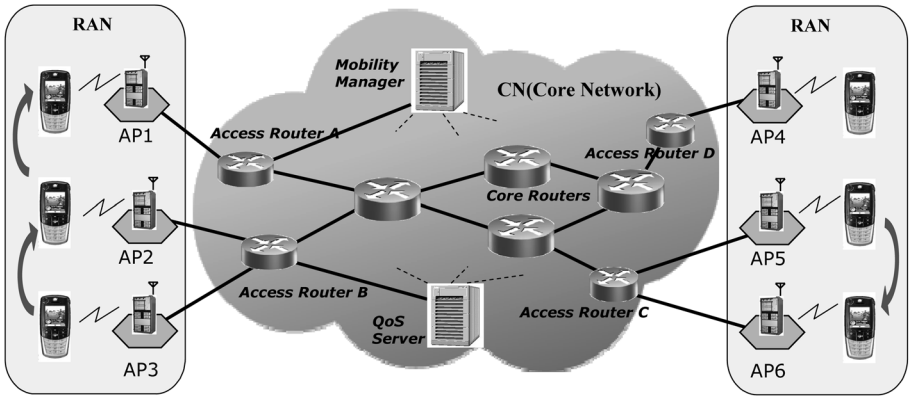

Figure 5. Assumed Architectural Model for Wireless Mobile Network.

The ARs (Access Router) in CN are key control points in the network. They are IP routers that have one IP hop distant from the mobile node via AP (Access Point). All data packets to and from the mobile node and signaling messages between the mobile node and various servers in the network pass through the ARs. The ARs are integrating various radio access networks (e.g. 3.5G, 3.9G). The CRs (Core Router) are high speed routers that lie in the core network.

The main function of QS is Admission Control based on the adaptive QoS mechanism proposed in the next section. The MM manages MN's location information where the MN is located in the wireless mobile network and performs handover based on the handover policies provided by the network operator. The MM interacts with QS during the call setup, handover and termination.

3.2 Adaptive QoS Mechanism

There were some existing studies in terms of resource allocation such as [Ludwig et al. 2006] and [Kaneda 2008]. These studies, however, developed static resource allocation for the flow based on the class oriented fixed QoS so that the resources once

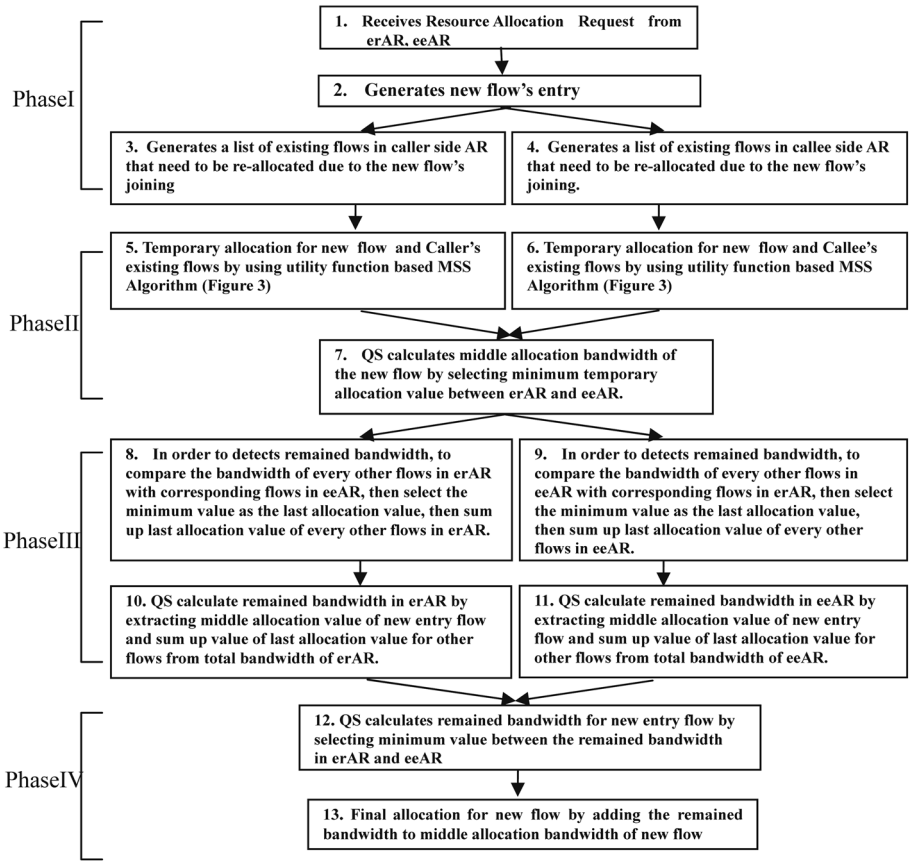

Figure 6. Adaptive QoS mechanism Based on the Utility Function.

allocated to the flow can not be adjusted during the service. Hence it contains the limitation not to satisfy the user's rate sensitive adaptive applications.

In order to solve such problem, we propose a new adaptive QoS Mechanism, under the assumption that network has the knowledge of utility function for all users, and we adopt MSS algorithm from caller to callee side in order to adjust the level of service quality based on the amount of the available resources. Also, in the mechanism, we extend resource adaptation algorithm for the flows by adding the procedure eliminating the differences of allocated resource between caller and callee side. The differences of allocated resources may be caused by different capacities of ARs and variance of returned resources from call terminations. Figure 6 shows a new adaptive QoS mechanism. Details of the mechanism are as follows.

1) Phase I (1-4): QS provides AR's flow list for the caller and callee side. Details of action for each step are as below:
   1. QS receives resource allocation request from caller or caller's AR called erAR.
   2. QS generates this new flow's entry called $\boldsymbol{f}_i$ which has the information of its caller's access router ($\boldsymbol{f}_i$-erAR), and callee's access router ($\boldsymbol{f}_i$-eeAR).
   3&4. From previous step 2, QS generates a list of the existing flows that need to be re-allocated due to $\boldsymbol{f}_i$'s joining. It is called $\boldsymbol{f}_i$-re-flow-list. This list is made of flows that are passing through the caller and callee ARs and that are affected by $\boldsymbol{f}_i$ joining. For example, $\boldsymbol{f}_i$-erAR-re-flow-list (for example, it includes flow-a, flow-b, flow-c, ...) and $\boldsymbol{f}_i$-eeAR-re-flow-list (for example, it includes flow-x, flow-y, flow-z, ...).
2) Phase II (5-7): Whenever new flow is added, temporary bandwidth for each existing flows and new flow are calculated according to the MSS algorithm for AR's existed flows for both of caller and callee side. For the new flow, QS selects the minimum value between caller's temporary bandwidth and callee's temporary bandwidth and set it as middle allocation value. Details of action for each step are as below.
   5. By using the utility function based on MSS algorithm described in 2.2., a temporary allocation is made to $\boldsymbol{f}_i$ and the flows in the caller AR list.
      So, $\boldsymbol{f}_i$ gets $\boldsymbol{f}_i$-er-alloc(bandwidth/sec), and the flows in $\boldsymbol{f}_i$-erAR-re-flow-list gets $\Sigma \boldsymbol{f}_i$-*erAR-re-flow-list-temp-alloc* = [temp-alloc-a + temp-alloc-b + temp-alloc-c + ...]).
   6. Same procedures are done for the flow $\boldsymbol{f}_i$ at the callee AR.
      Flow $\boldsymbol{f}_i$ gets $\boldsymbol{f}_i$-ee-alloc (bandwidth/sec), and the flows in the list for callee AR ($\boldsymbol{f}_i$-eeAR-re-flow-list) get $\Sigma \boldsymbol{f}_i$-*eeAR-re-flow-list-temp-alloc* = [temp-alloc-x + temp-alloc-y + temp-alloc-z + ...].
   7. QS calculates middle allocation for $\boldsymbol{f}_i$, this value is called $\boldsymbol{f}_i$-mid-alloc which is the minimum of [$\boldsymbol{f}_i$-er-alloc, $\boldsymbol{f}_i$-ee-alloc].
3) Phase III (8-11) : In order to detect remained bandwidth for both caller side AR and callee side AR, QS compares with the temporary value for each existed flow's in caller side AR and one of callee side, QS selects minimum value set as the last allocation for each existed flows, and sum up all as the last allocation for each caller side AR and callee side AR. Details of action for each

step are as below.

8. QS will detect if there is remaining un-utilized bandwidth in the ARs. This is done by re-calculating the bandwidth of the other flows in f_i-erAR-re-flow-list similarly to the process above. The calculation is done for the corresponding callee-AR of every other flow in the list. And again a minimum is selected for each flow. This is the last allocation for the other flows in the caller AR, resulting in last-alloc-a, last-alloc-b, last-alloc-c, ... Finally, QS sums up these values [last-alloc-a + last-alloc-b + last-alloc-c + ...] which can be expressed in Σf_i-*erAR-re-flow-list-last-alloc.*
9. In the similar manner, QS re-allocates each flow in f_i-eeAR-re-flow-list and results in Σf_i-*eeAR-re-flow-list-last-alloc* which is the sum of [last-alloc-x + last-alloc-y + last-alloc-z + ...].
10. Then the remaining bandwidth in caller side AR is f_i-erAR-remain-bw = (f_i -erAR-total-bw-f_i-mid-alloc-Σf_i-*erAR-re-flow-list-last-alloc*).
11. Similarly, the remaining bandwidth in the callee side AR is calculated in the way as above and results in f_i-eeAR-remain-bw = (f_i-eeAR-total-bw-f_i-mid-alloc-Σf_i-*eeAR-re-flow-list-last-alloc*).

4) Phase IV (12-13), QS compares the remained bandwidth for caller side AR with the one for callee side AR, then select minimum value and set as a network remained value. Finally, QS allocate the middle allocation bandwidth and remained bandwidth to the new flow. Details of action for each step are as below.

12. The remaining bandwidth for caller AR and callee AR for the flow f_i is remain-bw = min [f_i-erAR-remain-bw, f_i-eeAR-remain-bw].
13. At last, as the result, the final allocation for f_i=[f_i-mid-alloc + remain-bw].

3.3 Mobility Management

For mobility management, MM uses information (e.g. signaling strength) from MN and mobility profile (e.g. user location, frequency of user movement) as mobility triggers for handover. Such mobility triggers are used to initiate the resource allocation at the AR.

When an MN moves from one AR to another due to user's mobility, resources must be allocated for the MN in the new AR and the allocated resources in the old AR should be released. Moreover, in case if sufficient resources are not available at the new AR, the QS should be able to adaptively adjust the level of service offered.

Thus, MM needs to interact with QS to ensure the satisfied QoS for the MN located in the new AR. In this perspective, integration of mobility management and QoS mechanism is essential to provide the mobility to the user roaming from to ARs.

3.4 Signaling Mechanism

In this paper, we classify 4 classes of signaling messages that are used to the assumed network architectural model.

1) Path Signaling: These messages, which traverse the data path of the flows, will be processed by routers in the path from source to destination. We use RSVP signaling [Braden et al. 1997].

2) Access Signaling: These messages travel between the MN and the AR. As these signals use last hop air-link, the frequency and size of messages must be kept to a minimum. Specific signaling depends on wireless access technologies.

3) Server Signaling: Many entities in a network may need to communicate with each other. MM and QS should communicate in handover. In this paper, we use COPS (Common Open Policy Service) signaling [Durham et al. 2002].

4) Control Signaling: This signaling refers to Layer 3 messages supporting call control and mobility management. Proxy MIP (Mobile IP) [Gundavelli et al. 2008] and SIP [Rosenberg et al. 2002] signaling may be considered as candidate signaling. In this paper, we use SIP in order to verify the feasibility of integration of QoS and mobility management. We do not extend it to MN as usual SIP operation does; this is done to allow the use of non-SIP capable legacy MN.

3.4.1 *Call Setup.* Figure 7 shows the call setup signaling procedure integrating mobility and QoS management among the entities in the mobile network.

In the call procedure, the notification of AP, between MN and AR is abbreviated to simplify the simulation. Also, all the algorithms in the adaptive QoS mechanism in Figure 6 are implemented in QS entity. When MN1 calls MN2, MM proceeds to setup a session using SIP INVITE method. This procedure comprises of the 3 phases as follows:

1) Phase I: Call initiation procedure is performed between caller and callee's AR by the initiative of MM. After MM receives Call Request from caller, MM confirms the location of callee, and checks the availability of the path in the core network. Details of action for messages in sequence are as below.
   1. MN1 sends Call Request to AR1 with application ID. This Call Request message is functionally similar to SIP INVITE.
   2. AR1 responds with Wait Signal and forwards MN1's request to MM.
   3. MM finds the callee's location (i.e., under AR2) and sends SIP INVITE to callee AR2.
   4. With the recognition of QoS call request and confirming that the callee is attached to AR2, AR2 answers with a reliable provisional response: 183 Session Progress. At this time, AR2 does not alert the callee until all the mandatory QoS preconditions in the answer are met later.
   5. MM confirms the callee's existence, and sends MM SIP INVITE to caller's AR1 with callee' location information (e.g. under AR2).
   6. AR1 answers with a reliable provisional response: 183 Session Progress.
   7. MM generates PRACK response to 183 to caller's AR1 and callee's AR2
   8. Caller's AR1 and callee's AR2 send back 200OK to answer PRACK.
   9. At the same time, with the callee's location information written in received INVITE message, caller's AR1 sends RSVP PATH message to callee's AR2 to probe core network's availability.
   10. AR2 receives RSVP message and sends RSVP RESP message back to AR1.
2) Phase II: QS calculates the available bandwidth according to the algorithm (Figure 6) for the resource allocation for both caller and callee side. Details of action for messages in sequence are as below.

11. AR1 sends bandwidth request to QS by via COPS REQ message. Caller, callee's location information, requested bandwidth, and available bandwidth are listed in this message.
12. QS checks its database for the utility function associated with the caller and application ID, and QS calculates bandwidth allocation for this session and then sends this allocation to caller's AR1 and callee's AR2 with COPS DEC message.

3) Phase III: MM continues call proceeding procedure after the confirmation of the resource allocation from both AR of caller and callee side, then the data path is opened. Details of action for messages in sequence are as below.

13. With this QoS notification, AR1 can answer the early INVITE request with 200 OK to MM. Although the QoS preconditions are met, AR1 doesn't alert caller to send data packet at this time, since the session has not been established yet.
14. At the callee's side, AR2 alerts callee MN2 with proprietary Call Request message and sends a provisional response (180 Ringing) to answer the early

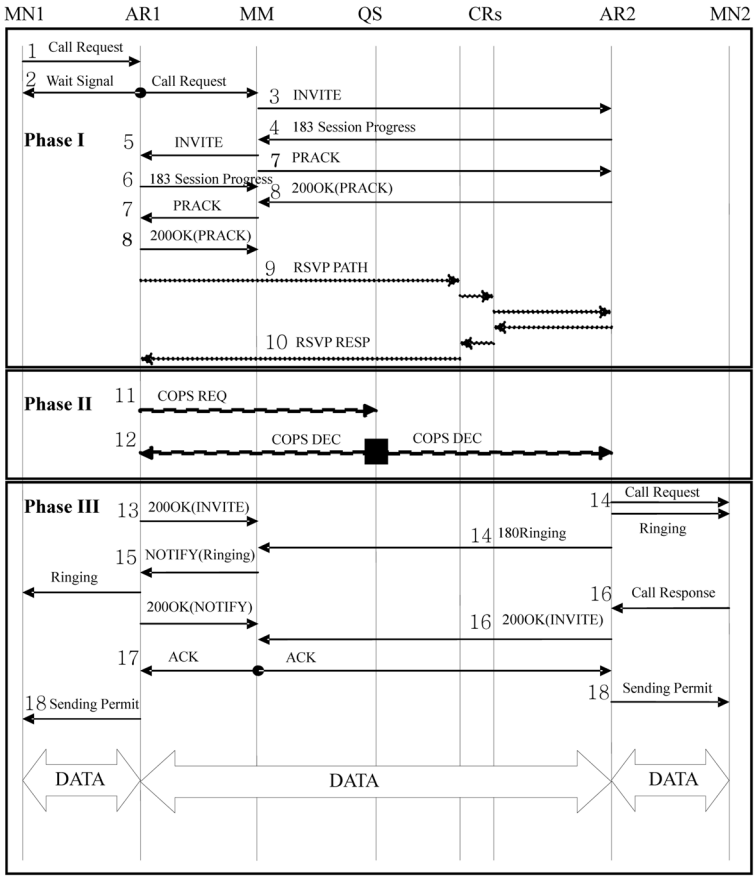

Figure 7. Call Setup Signaling.

INVITE request from MM.

15. MM may pass this information to AR1 and MN1 by SIP NOTIFY message.
16. When MN2 picks up the phone, AR2 receives Call Response message from MN2 and answers INVITE with the definitive response 200 OK to MM.
17. At this time, MM receives two final responses for the early two INVITE requests. It sends ACK to both caller and callee.
18. After receiving Sending Permit, the caller and callee begin to exchange data packets using notified bandwidth for the communication.

In Figure 7 Call Setup Signaling, the allocation procedure at QoS server is added to the existing setup signaling so that the startup delay may become longer by this procedure processing time delay.

We consider the case where QoS server can manage utility functions related to wireless resources of 3 base stations (Node B), which are equivalent to 3 APs, and the number of the simultaneous connection is 500 for each Node B. When the new call is generated and the new bandwidth should be allocated based on utility functions, the new inserting procedure should be conducted at QoS server. In the case where one inserting is assumed to be 200 micro sec in average, and a new call is generated at each Node B in which the simultaneous connections are 500, the total inserting time at maximum, will be 300 msec (200×500×3/1000) in the condition that the total inserting time can be proportional to the connection number. The transmission delay between QoS server and Node B is assumed to be 20msec so that the total delay is 320msec. In comparison with this delay and the startup delay of HSPA (High Speed Packet Access), the 3.5 Generation in 3GPP standard, which is usually 2 or 3 sec, the delay of this allocation procedure is not so long.

3.4.2 *Call Handover.* When a MN moves from one AR to another, resources must be reserved for the MN in the new AR and resources that were reserved in the old AR should be released. When a handover is about to occur, the MM initiates handover QoS signaling based on inputs from the MN. Figure 8 shows the handover signaling procedure and the entities involved integrating mobility and QoS management.

1) Phase I: When a callee moves from one AR to another AR, MM proceeds to setup another new session with the moving callee's new AR, and to check if the path is available in the core network. Details of action for messages in sequence are as below.
   1. MN2 Re-REGISTER to MM from AR2 to AR3 with location information for AR3. It depends on the specific handover triggering mobility scheme.
   2. MM sends SIP INVITE message to new AR3, and Re-INVITE to AR1.
   3. AR1 and AR2 answers with a reliable provisional response: 183 Session Progress. At this time, AR1 and AR3 do not alert the MN1, MN2 until all the mandatory QoS preconditions in the answer are met later.
   4. MM generates PRACK response to AR1 and AR3 in response of 183 Session Progress.
   5. R1 and AR3 respond back 200OK to MM to answer PRACK.
   6. AR1 sends RSVP PATH message AR3 to probe core network's availability.

7. AR3 receives RSVP message and sends RSVP RESP message back to AR1 if core network is available.

2) Phase II: QS calculates available bandwidth according to the algorithm (Figure 6) for resource allocation for the new session. Details of action for messages in sequence are as below.

8. AR1 sends bandwidth request to QS by via COPS REQ message. MN1, MN2's location information, requested bandwidth, and available bandwidth are listed in this message.
9. QS checks its database for the utility function associated with the caller and application ID, and QS calculates bandwidth allocation for this session, and then sends bandwidth allocation information to MN1's AR1 and MN2's AR3 with COPS DEC message.

3) Phase III: MM continues session call proceeding procedure after confirmation of the resource allocation from both AR of caller and callee side, and then the data path is opened. Details of action for messages in sequence are as below.

10. With the QoS notification, AR1 answers with 200 OK for the early Re-INVITE request to MM.

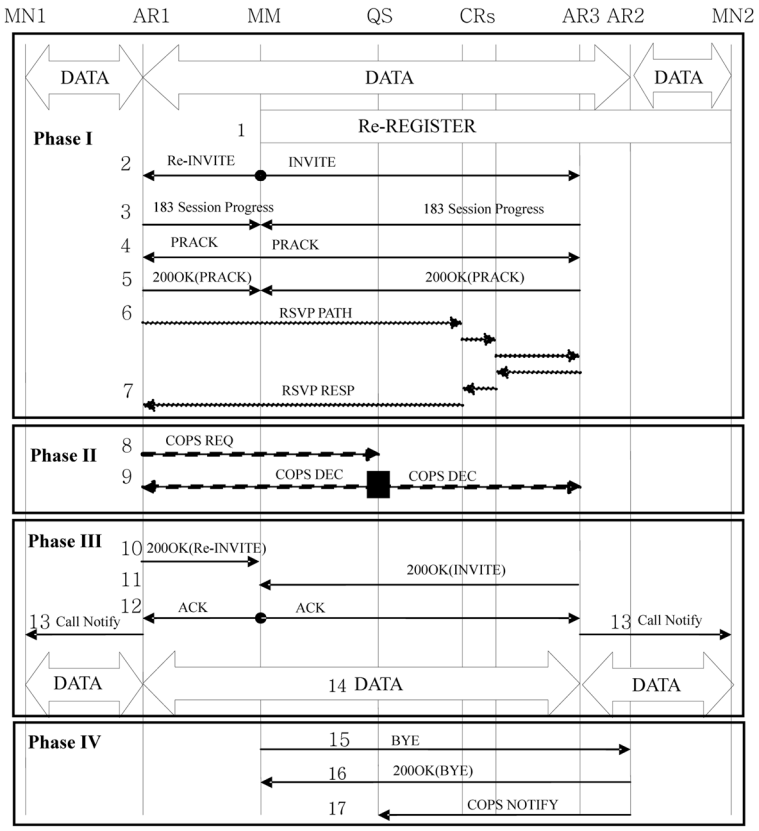

Figure 8. Handover Call Signaling.

11. AR3 answers with 200 OK for the early INVITE request to MM.
12. At this point, MM sends ACK to both AR1 and AR3.
13. After receiving this ACK, both AR1 and AR3 notice Call Notify to each MN1 and MN2.
14. After receiving Call Notify, MN1 and MN2 begin to exchange data packets using notified bandwidth for the communication.

4) Phase IV: After the new session opens, MM initiates to close the old session. Details of action for messages in sequence are as below.

15. MM sends SIP BYE to old AR2 to close the old session.
16. AR2 200 OK as the response for SIP BYE.
17. AR2 notifies the sessions' setup and modification to the QS by sending COPS NOTIYFY.

4. EVALUATION

4.1 Network Topology and Conditions

In order to evaluate the feasibility of our proposed QoS mechanism, simulations have been performed by using the NS2 simulator on the Solaris OS in the multi purpose workstation system. All network entity nodes such as MN, AR, CR, MM, and QS are implemented in C++ as the agents on the NS2 simulator compliant with the topology shown in Figure 9.

In the simulated network topology, we assumed the network has 5ARs, 7 CRs, 20 MNs, 1 MM, and 1 QS. MM and QS have linked to each AR. Also, each AR has 4 units of bandwidth to administrate (1 unit = 40 kbps). The data flows are using RTP packets. Our proposed QoS mechanism is implemented in QS, the proposed signaling procedures are implemented in the event scheduler according to the

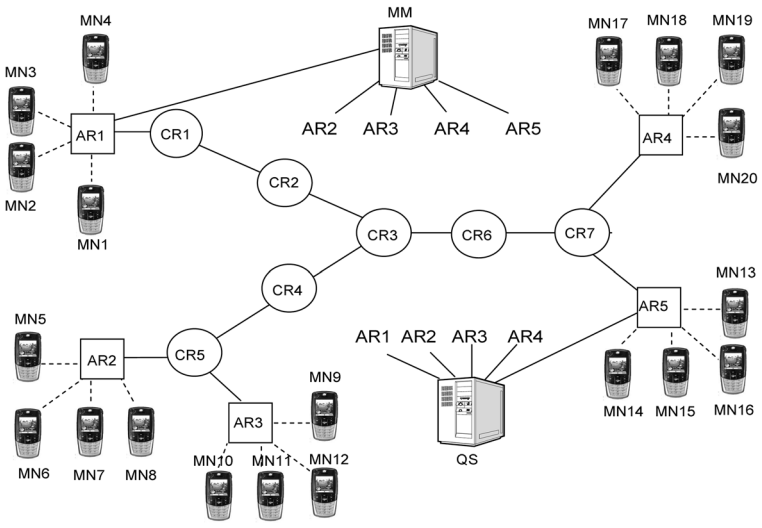

Figure 9. Simulated Network Topology.

following scenarios.

1) Scenario for simple 4 flows not including H.O call.
   MN1, 2, 3, and 4 under AR1 have sessions with MN5 under AR2, MN9 under AR3, MN13 under AR5, and MN17 under AR4, respectively.
2) Scenario for simple 3 flows including H.O. call.
   MN2 under AR1, MN6 and MN8 under AR2 have sessions with MN9, MN10 under AR3 and MN14 under AR5, respectively. MN8 moves to area under AR3. Figure 11 shows the scenario for the simple flows including H.O call
3) Scenario for more complicated 8 flows including H.O. call.
   MN1, 2, 3, and 4 under AR1 have sessions with MN5 under AR2, MN9 under AR3, MN13 under AR5, and MN17 under AR4, respectively. Additionally, MN6 under AR2 has session with MN10 under AR3, MN8 under AR2 has session with MN14 under AR5, MN8 moves to area under AR3, MN11 under AR3 has session with MN18 under AR4, and MN15 under AR5 has session with MN19 under AR4, respectively.

For the simulation, it is assumed that the location registration for MN is conducted before the call setup. When an MN turns on or enters into an area covered by AR, MN sends Registration Request message to its AR, and the MN's location is recorded in the Location server (i.e. MM) and AR if necessary. Also, AP is abbreviated to simplify the simulation. The proprietarily defined messages are used instead of using a specific wireless protocol in signaling between MN and AR to be independent of the specific wireless access technology.

4.2 Simulation Results

As the result of the first scenario for non handover call simulation, Figure 10 shows the bandwidth allocation for 4 flows. The utility value of Flow1 from MN1 to MN5,

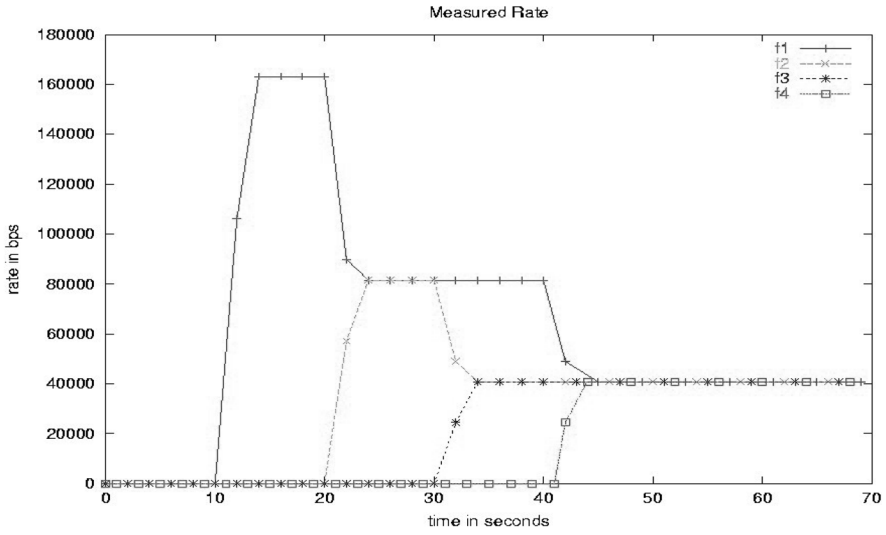

Figure 10. Simple Flows Evaluation for Non-Handover.

Flow2 from MN2 to MN9, Flow3 from MN3 to MN13, and Flow4 from MN4 to MN17 are listed in Table II. Flow1 enters the network at time 10s. As it is the first flow in the system, it is allocated the entire AR1's bandwidth. At time 20s, Flow2 enters the network. At time 30s, with Flow3's join, the allocation becomes f1:f2:f3 = 2:1:1. Finally, after Flow4 enters, each flow has same bandwidth and this is equal to the allocation that expected.

As the result of the second scenario for 3 flows including handover call simulation, Figure 11 shows bandwidth allocation for each flow when a handover takes place. The utility value of Flow-big, between MN2 and MN9, Flow-small between MN6 and MN10, and Flow-h (handover), between MN8 and MN14 are listed in Table III. Flow-h starts its call at time 30s and MN8 starts to move toward to new AR (AR3) at time 40s. Before Flow-h enters to AR3, there are two flows in AR3. Our algorithm therefore will allocate 2 units of bandwidth to each flow (Flow-big and Flow-small). Once a new MN having an additional utility function joins to an AR, there is a need to re-allocate the existing resources among the users in the AR according to the utility function that the users have contracted. Hence, when Flow-h enters the area of AR3, the allocation becomes Flow-big : Flow-h : Flow-small = 2 : 1 : 1 units. The reason for this allocation is that Flow-big has higher utility than Flow-h and

Table II. Utility Values for Flows.

Flow	U1	U2	U3	U4
Flow1	0.9	0.3	0.15	0.1
Flow2	0.8	0.25	0.14	0.1
Flow3	0.7	0.2	0.2	0.1
Flow4	0.4	0.4	0.1	0.1

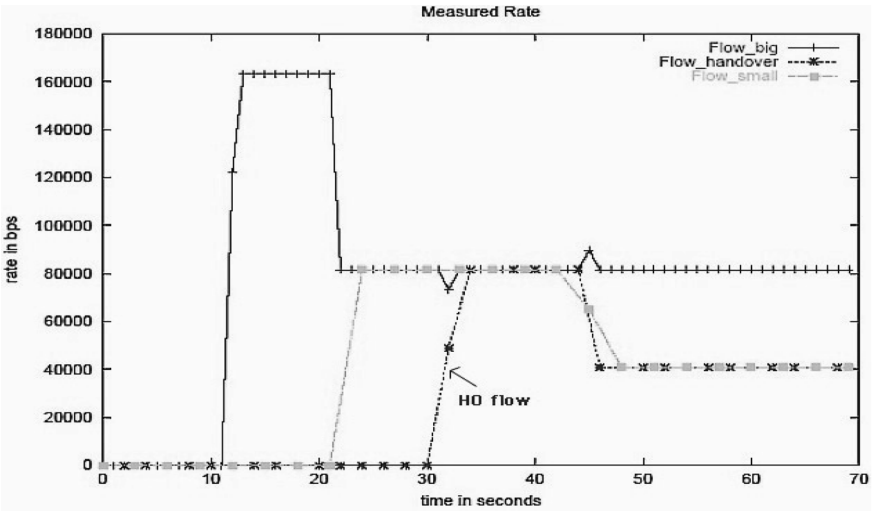

Figure 11. Simple Flows Evaluation for Handover.

Table III. Utility Values for Flows.

Flow	U1	U2	U3	U4
Flow_big	0.7	0.5	0.14	0.1
Flow_small	0.45	0.3	0.14	0.1
Flow_handover	0.5	0.4	0.14	0.1

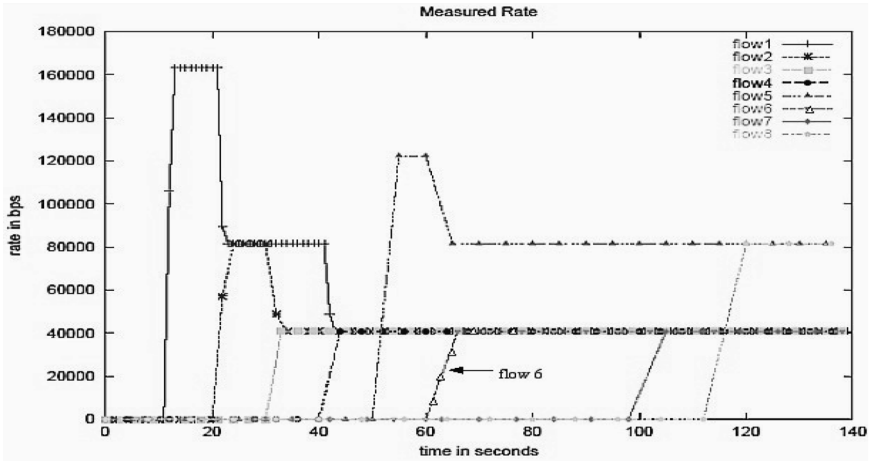

Figure 12. Multiple Flows Evaluation for Handover.

therefore it is not affected by Flow-h joining the AR. On the other side, Flow-small has lower utility than Flow-h and therefore one unit is re-allocated to Flow-h.

As the result of third scenario for more complicated 8 flows including handover call simulation, Figure 12 shows the bandwidth allocation for each flow when a handover takes place (Flow 6*). The utility value of each flow is listed in Table IV. Regarding Flow 1 to Flow 4, they are the same as Figure 10 and Flow5, Flow 6*, Flow 7 and Flow8 are from MN6 to MN10, from MN8 to MN14, from MN11 to MN18 and from MN15 to MN19, respectively. In this simulation, MN8 in Flow 6

Table IV. Utility Values for Flows.

Flow	U1	U2	U3	U4
Flow1	0.9	0.3	0.15	0.1
Flow2	0.8	0.25	0.14	0.1
Flow3	0.7	0.2	0.2	0.1
Flow4	0.4	0.4	0.1	0.1
Flow5	0.6	0.52	0.14	0.1
Flow6*	0.45	0.3	0.14	0.1
Flow7	0.4	0.25	0.14	0.1
Flow8	0.24	0.15	0.14	0.1

moves from AR2 to AR3 (handover) so that at AR3 bandwidth is reallocated. From this result, we can confirm that the final allocation done by our algorithm shown in Figure 12 is equal to the expected allocation.

4.3 Considerations

Through the simulation, we have confirmed the followings.

1) From the user's viewpoint, the users can obtain the more satisfied service by this mechanism. Normally when the bandwidth is fully utilized (i.e. congested), the new service request is rejected. In case of handover, the flow is suspended (i.e. mobility service is suspended). The proposed mechanism, however, can solve these issues. By this mechanism, since the priority control can be managed by handling the value of the utility functions, the new service can be prioritized even in the congested situation including handover cases. That is, by this mechanism, the users' satisfaction will be increased even in the congested or handover situation.

2) We can expect that the situation for some heavy users to dominate the resources will be avoided by managing utility functions. In the light traffic environment, heavy traffic users can use the resources without restriction, but in the congested traffic environment, the heavy traffic users are affected by users whose incremental value of utility functions is higher so that it can avoid for heavy users, whose incremental value of utility functions is lower. That is, we can control this mechanism to be applied to some limited users, not to all users.

3) In the wireless network, bandwidth will be allocated based on the utility function for the effective management of wireless resources. While, in the wired network, the existing priority control for the high-priority real-time communication services can be adopted to realize the low delay. On the other hand, by using our method, the total network traffic in the wired network may be increased, and the delay may be increased as well. However, it can be controlled by network operation. For example, when the most of whole network bandwidths in wired network are used (e.g. 90%), the queue delay will be increased even though the priority control is adopted. But, when the usage of bandwidths are controlled under the limited ratio (e.g. 70%), the priority control can work correctly so the delay will not become long.

4) In this paper, we assume that the network has knowledge of utility function for all users. However, there are two issues, the one is how to formulate utility function information for each user, and the other is how to maintain information for all users.

On the first issue, the utility value that users may obtain from a certain quantity of bandwidth resource can be thought as the price that they would be willing to pay to obtain a specific quantity of resource. In this view point, in order to let the network to have knowledge of utility function, a network operator should provide ways how to formulate incremental values of utility function, which means setting the prices that user would be willing to pay for each unit of bandwidth resources. In regard to this, the network operator may provide typical sets of incremental values of utility function when the user subscribes to the network, then the network operator sends a set of incremental values selected by the user to the QS. After

user's subscription, the user may update incremental values/prices of each unit of bandwidth resources in the utility function according to the user's preference through the service provided by network operator, and send it to the QS in the network by using the signaling. Detailed operation is depended on the operation policy of the network operator.

On the second issue, as it is written in Section 4.2 Consideration, as the utility function can be applied to the limited users, incremental values/prices of utility function for all or part of users (based on the operator's policy) can be sent to the QS in the network system through the ways explained as above, and maintained in QS under the management by network operator. Network operators may employ a centralized QS system in the network. Advantage of this approach is that it provides simple operation, but it may not scale to large network due to the processing over head. On the other hand, network operators may employ distributed QS system. This approach is more scalable and may reduce latency of processing time. But, it involves more signaling overhead compared with centralized QS. Selection of the specific approach is depended on the operation policy of network operator.

5. CONCLUSION

In this paper, we have proposed a new adaptive QoS mechanism based on the utility function and its integration with an SIP based generic mobility management protocol. By the proposed mechanism, the adaptive resource allocation based on user preference can be realized in the mobile network. Also, the QoS mechanism can be performed in handover without any restrictions. As a next step, we can expand this mechanism on the following items.

1) In case that some of the existing user's utility values are lower and new users having higher utility value are joining continually, some of the existing user's flow may be suspended based on our proposed mechanism. However, even in this case, the service can be continued with slight modification of our mechanism so as to the whole existing services may guarantee some minimum bandwidth.

2) Pricing policy of the utility function for each user and the topology of QS system such as centralized and distributed approach can be studied in the next step.

3) We have focused on only bandwidth allocation, but other QoS metrics such as the delay or jitter can be studied in the next step with detailed pricing policy on the differentiated user's preferences.

REFERENCES

3GPP TS 29.212. 2008-09. 3rd Generation Partnership Project: Technical Specification Group Core Network and Terminals; Policy and Charging Control over Gx reference point (Release 8).

BRADEN, R., JHANG, L., BERSON, S. ET AL. 1997. Resource Reservation Protocol. RFC2205, IETF.

DURHAM, D., BOYLE, J., COHEN, R. ET AL. 2002. Common Open Policy Service. RFC2748, IETF.

GONZALEZ, O., MICHAEL NEEDHAM. 2004. QoS Provisioning Architecture for Next Generation Mobile Networks, *IEICE Transactions on Communications. E87-B*, 5, May.

GUNDAVELLI, S., LEUNG, K., DEVARAPALLI, V. ET AL. 2008. Proxy Mobile IPv6. draft-ietf-netlmm-proxymip6-18.txt.IETF.

KANEDA, Y. 2008. Policy-based End-to-End QoS Guarantee Using On-Path Signaling for Both QoS Requests and Feedback. ICOIN I-1. Jan.

LUDWIG, R., EKSROM, H., WILLARS, P. ET AL. 2006. AN Evolved 3GPP QoS Concept. VTC Spring. May.

MCCANNE, S., JACOBSON, V. AND VETTERLI, M. 1996. Receiver-driven layered multicast. Proc. ACM SIGCOMM 96, 117–130, Standford, CA. Aug.

NOMURA, T., YAMORI, K., TAKAHASHI, E. ET AL. 2001. Waiting time versus utility to download images. In APSITT2001 session 5. Nov. 128–132.

POLITIS, C. ET AL. 2004. Cooperative Networks for the Future Wireless World. IEEE Commu. Mag., Sept. 70–79.

ROSENBERG, J., SCHULZRINNE, H., CAMARILLO, G., SPARKS, R., HANDLEY, M. ET AL. 2002. SIP: Session Initiation Protocol. RFC 3261. Proposed Standard. IETF.

SHENKER, S. 1995. Fundamental Design Issues for the Future Internet. *IEEE Journal on Selected area in Telecommunication. 13*, 7, Sept.

VARIAN, H. R. 1992. Microeconomic Analysis, W. W. Norton and Co.

YAMORI, K. AND TANAKA, Y. 2004. Relation between willingness to pay and guaranteed minimum bandwidth in multiple-priority services. In APCC2004. no.MA06-1. Aug. 113–117.

YOSHIMURA, T., OHYA, T., KAWAHARA, T., AND ETOH, M. 2001. A QoS Cotrol Method of MPEG Video with an RTP Monitoring Agent for Mobile Streaming Service, *IEICE J85-B*, 8 (Japanese) Aug. 1243–1253.

Kwang Sik Kim received B.E. and ME. degree in Computer Engineering from HongIk University and MyungJi University, Seoul Korea in 1985 and 1987 respectively.
He joined ETRI (Electonics and Telecommunications Research Institute), Taejon Korea in 1989, and had engaged in research and development of ISDN and ATM switching system. He joined Motorola Japan in 1998, and had engaged in research and development on CDMA2000 BSC System. In 2000 he joined Lucent Japan and had engaged in research and development on IMS and Parlay. He is a candidate of PhD in Tokyo University of Information Sciences and is engaged in research on Ubiquitous Network.

Shintaro UNO received BS in mathematics from Keio University and MS and Ph.D in electrical engineering from Keio University. He joined Toshiba R&D center in 1983 to research control methods for satellite communication. He joined Nokia Research Center Tokyo to study VoIP in 1999 and he joined Motorola Japan Research Lab in 2000 to study ITS, QoS and next generation wireless networks. He is now visiting professor of Graduate School of Engineering, Kanazawa Institute of Technology, Tokyo Japan.
His interest includes wireless networks, QoS and mobility management.

Moo Wan Kim received B.E., M.E. and Ph.D degree in electronic engineering from Osaka University, Osaka, Japan in 1974, 1977 and 1980, respectively. He joined Fujitsu Lab. in 1980 and had been engaged in research and development on multimedia communication systems, Intelligent Network, ATM switching system and operating system. In 1998 he joined Motorola Japan and had been engaged in research and development on CDMA2000 system. In 2000 he joined Lucent Japan and had been engaged in research and development on W-CDMA system, IMS and Parlay. In 2005 he joined Tokyo University of Information Sciences and has been engaged in research on Ubiquitous Network.
He is currently a Professor in the Department of Information Systems.