This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the WCNC 2008 proceedings.

ECHO: A Quality of Service based Endpoint Centric Handover scheme for VoIP

John Fitzpatrick, Seán Murphy, Mohammed Atiquzzaman*, John Murphy

Performance Engineering Lab,
School of Computer Science and Informatics,
University College Dublin,
Ireland
E-mail: john.fitzpatrick@ucd.ie

**School of Computer Science,*
University of Oklahoma,
Norman, OK.
E-mail: atiq@ou.edu

***Abstract* – Existing terminal oriented handover mechanisms capable of meeting the strict delay bounds of real time applications such as VoIP do not consider the QoS of candidate handover networks. In this paper ECHO – a QoS based handover solution for VoIP– is proposed. ECHO is endpoint centric and does not require any network support; it leverages the SCTP transport protocol. ECHO incorporates network metrics that directly affect VoIP quality into the handover decision process. A dynamic variant of the ITU-T E-Model is used to calculate how the network metrics map to a user perceived voice quality metric known as the MOS. The MOS value is then used to make handover decisions between each of the available access networks. The results show that the addition of the QoS capabilities significantly improves the handover decisions that are made.**

I. Introduction

An increasing variety of options for wireless network access, coupled with increased capabilities of mobile end-terminals provide users with the possibility of seamlessly accessing the network via different radio access technologies. Before this can be realised, however, it is necessary to develop solutions for handover between these disparate networks.

While substantial work has been performed on this, and there exist some approaches to handover in this context, few can meet the more stringent demands imposed by delay sensitive applications. Voice traffic, in particular, is one traffic class which is very prevalent and also imposes stringent demands on the handover mechanism. With increasing penetration of VoIP, it is interesting to investigate handover mechanisms which can support VoIP without significant interruption in service.

Much work has been done on Mobile IP (MIP) for mobility support in IP networks [1, 2]. However, MIP is seeing little success in terms of network deployment, despite the widespread availability of equipment for some time. While the reasons for this are complex, one key challenge for this technology is that it requires large scale rollout before becoming very useful. This creates an opportunity for alternative solutions which do not require such sweeping network infrastructure modifications. They are the focus of this work.

Endpoint centric handover solutions require no network infrastructure modifications and therefore have fewer barriers to deployment. SIGMA is one such solution, which is based on the SCTP transport protocol. Although SIGMA is capable of providing seamless handovers for VoIP with no degradation in call quality [3], the handover decision is based on a simple comparison of the Received Signal Strength (RSS) from each available access network. The Quality of Service (QoS) in terms of Delay, Loss and Jitter that each network can offer is not considered. This can result in handovers to networks that cannot support the QoS required by the application if, for example, a network was congested.

In this paper, Endpoint Centric Handover (ECHO), is proposed, which addresses some of deficiencies of SIGMA. The scheme incorporates measured network performance parameters to effect a QoS-aware handover decision. In this way, the decision results in improved VoIP performance.

This rest of this paper is structured as follows. Section II describes related work. Section III gives an overview of SIGMA and the E-Model. Section IV describes the proposed QoS handover mechanism and Section V describes the experimental testbed to compare the handover performance of SIGMA and ECHO. Results are presented in Section VI, followed by the conclusion in section VII.

II. Related Work

Although some work has been done in the area of QoS handovers, existing mechanisms either fail to meet the delay requirements of real time applications or require the support of specific network nodes. In [4] a quality based handover trigger for transferring calls from WLAN to cellular is proposed. A handover is performed with the support of a PBX within the network. The handover trigger is based on metrics such as delay and loss experienced at the PBX.

Media Independent Handover (MIH) [5] is being standardised within the 802.21 working group to enable handovers between different access technologies. One important feature of MIH is the ability to maintain QoS before and after handover. This is achieved using a combination of reservation techniques, direct communication of signalling traffic and data forwarding between access points. MIH requires that each network node implements a MIH Function (MIHF). MIHF allows MIH devices and access points to communicate information about the availability of networks and the QoS that can be supported. Although, MIH can provide QoS handovers it requires that all network nodes be MIH capable and thus may suffer from the same rollout problems as MIP.

In [6] an endpoint centric handover solution for vertical handoff between WWAN and WLAN is proposed. The system

1525-3511/08/$25.00 ©2008 IEEE

introduces a connection manager (CM) which employs a combination of MAC layer sensing and FFT-Based Decay Detection of the RSS. The CM uses the 802.11 Network Allocation Vector (NAV) occupation time to estimate the traffic load and the QoS in the form of the network access delay to make handover decisions. Results are presented with respect to TCP throughput but no discussion of the handover delay is included or the solutions applicability to real time traffic, such as VoIP.

The VoIP handover mechanism proposed in this work differs from these existing solutions by focusing on a terminal oriented solution that does not require network support or modifications. Also, the proposed mechanism is QoS based and can meet the strict delay requirements of VoIP.

III. TECHNICAL BACKGROUND

A. SCTP

Stream Control Transmission Protocol (SCTP) is the third transport layer protocol to be ratified by the IETF [7]. It is a message oriented reliable transport layer protocol which inherited many of the core features of TCP such as congestion control and retransmission. In terms of this work, the main advantage of SCTP over other transport layer protocols is multihoming.

The multihoming feature of SCTP allows a single association1 to span multiple IP addresses. Each IP address can be bound to a separate IP interface connected to different physical networks. The IP diversity achieved through multihoming allows for simultaneous connections through multiple IP interfaces and enables soft handovers to take place. This element of SCTP is critical to the proposed QoS handover mechanism.

Table 1 - VoIP Call Ratings

R value (lower limit)	MOS (lower limit)	User Perception
90	4.34	Very Satisfied
80	4.03	Satisfied
70	3.6	Some users dissatisfied
60	3.1	Many users dissatisfied
50	2.58	Nearly all users dissatisfied

B. PR-SCTP

SCTP requires strict message ordered delivery and suffers from head of line blocking. Although SCTP can support unordered message delivery the head of line blocking problem still exists. To combat this, we used the Partial Reliability extension to SCTP [8] to provide varying levels of reliability to upper layer protocols.

Partial Reliability uses a 'Timed Reliability' parameter which allows the sender to specify a time to live parameter on a per message basis. The time to live parameter defines the duration for which the sender should attempt to transmit and retransmit the message. Essentially, PR-SCTP allows the sender to define how persistent the transport layer will be at attempting to deliver each message. The partial reliability extension to SCTP is not used by default in an association; an SCTP endpoint can use partial reliability only if it is supported by its peer. An endpoint is notified that partial reliability is supported by its peer during association establishment.

C. Applying the E-Model in Real Time

The E-model is a computational model for estimating the subjective quality of a VoIP call. It is standardized by the ITU-T (International Telecommunications Union Technical standards) as recommendation G.107 [9]. The E-Model combines loss and delay impairments based on the concept that perceived quality impairments are additive. The primary use of the E-model is in the design of codecs and transmission networks. The output of the E-model algorithm is a scalar rating of call quality called the *R* value.

The E-model output *R* can be converted into the more commonly known metric Mean Opinion Score (MOS) that measures how a user rates call quality. Table 1 shows the non-linear mapping of the R rating to MOS.

While the E-model was primarily designed as a network planning tool, some work has been done on applying the E-Model to real time environments [12]. Default values can be chosen for parameters within the E-Model. This simplified variant can then be used in a real time context. Using the work in [12] the E-model algorithm becomes:

$$R = 93.34 - Id - Ie$$

Impairments due to network transmission are represented by *Id* and take into account network delay and jitter parameters. The equipment impairment factor *Ie* is loss and codec dependent.

D. SIGMA

Seamless IP-diversity based Generalized Mobility Architecture (SIGMA) [10] incorporates SCTP multihoming and the Dynamic Address Reconfiguration Extension (DAR) [11] to perform seamless handovers between wireless networks. The DAR extension to SCTP allows addresses to be dynamically added and deleted from an association. DAR defines a new message type called an Address Configuration Message (ASCONF). The ASCONF message can be transmitted by either endpoint to inform its peer of IP addresses through which it is reachable. This can be done dynamically during an active association and is the main feature that enables SCTP to support seamless handovers.

The SIGMA handover process can be defined in the following 4 steps.

1. Acquire new IP address – When the MH moves into the coverage area of a wireless access network it is assumed that it can detect the availability of this network. For example in the experimental testbed which will be described later, 802.11 wireless access points were used and network detection is done via the APs beacon frame router advertisement.
2. Add IP - Once the MH has acquired a new IP address it must add this to the association by informing the CN of the new IP. This is done using DAR.
3. Handover Decision - The handover decision is based on the RSS of each available AP. When the RSS of

[1] A central concept in SCTP is the definition of an *association*. An *association* in SCTP is analogous to a connection in TCP.

the newly available AP becomes greater than that of the existing AP, a handover is triggered. To perform the handover SIGMA must redirect the data flow to the CN via the new AP. The handover is done by sending a '*Set Primary*' ASCONF message to the CN containing the new Primary address.

4. Delete IP - The final step in the handover process is to remove the old IP address from the association, so that no data is transmitted to the MH via the old access network which may no longer be available. Once again, an ASCONF message is transmitted to the CN containing the '*Delete IP*' parameter.

IV. ECHO MOBILITY MECHANISM

This section gives an overview of the proposed QoS enhancement to the SCTP based SIGMA mobility scheme. ECHO builds on SIGMAs RSS based handover mechanism by considering QoS metrics thereby providing a better handover decision. As SIGMA is endpoint centric, any QoS based handover mechanism should also be endpoint centric. ECHO meets this requirement.

ECHO begins in the same manner as SIGMA until the handover decision at step 3 (see section III – D). ECHO increases the complexity of the step 3 handover decision as outlined below.

A. ECHO Handover Decision

Once in the overlapping coverage region of two access networks, a handover decision may be necessary. The MN begins to monitor the RSS of each available AP. When the RSS of the new AP becomes greater than that of the existing AP the handover decision process is triggered. At this point SIGMA would have immediately performed a handover to the new access network, without considering the QoS that will be obtained after handoff.

ECHO bases the handover decision on both the RSS and the available downlink QoS from each access network. In order to calculate the downlink QoS for each link independently, data must be transmitted over both links simultaneously. To do this, when the MN moves into the overlapping region of two access networks it informs the CN to begin transmitting data over both links. The MN uses the downlink traffic streams to obtain network metrics for each access network. These metrics are then used to calculate the MOS for each link using the E-Model algorithm on which the handover decision is based. Once a handover decision is made the MN informs the CN to stop transmitting data over both links and only use the network to which a handover decision has been made.

A flow chart of the handover process is shown in figure 1. As can be seen, ECHO assesses the QoS of the new network before making the handover decision. To do this the MN transmits a 'Begin Duplication' message to the CN. On reception of this the CN begins simultaneously transmitting all data to each of the MN IP addresses specified in the association. This allows the MN to individually measure the downlink delay, jitter and loss of each access network. The obtained metrics for each access network are then converted to a MOS score using the E-Model algorithm, the implementation of which is discussed in Section IV - B.

A MOS score threshold is used to decide if the call can be supported on the newly available access network. An appropriate MOS threshold was chosen from table 1.

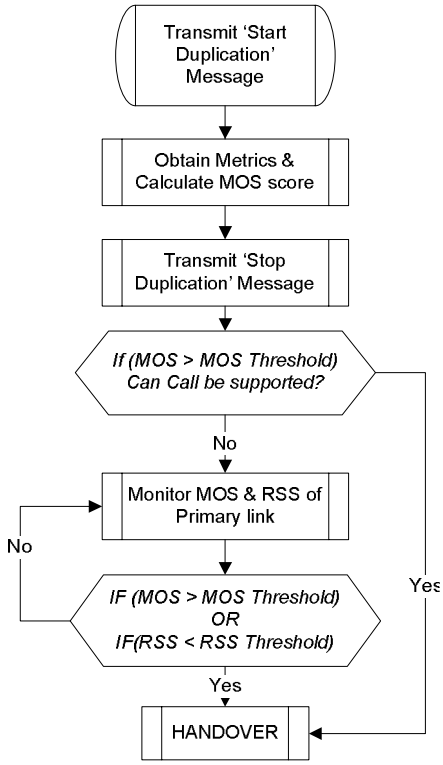

Figure 1 - ECHO Flowchart

The handover decision leads to two scenarios described below.

Scenario 1 – The call can be supported by the new access network. This scenario is straightforward. Since the call can be supported a handover is immediately performed to the new access network.

Scenario 2 – The call cannot be supported by the new access network. In this situation the MN assesses the MOS score obtained over the existing connection; if the call quality is above the MOS threshold then no handover takes place. The MN then continues to monitor both the RSS and MOS of the existing network connection. Two threshold values are then used to initiate a handover. A handover is triggered if the call quality falls below the MOS threshold. Likewise if the RSS falls below an RSS threshold, a handover is triggered. Details of the threshold value used will be discussed in Section VI.

B. Calculating MOS

The network metrics required to calculate the MOS are delay, jitter and loss. Inter-arrival jitter and packet loss are obtained using the individual downlink VoIP streams over each access network. Jitter is calculated using the E-Model recommended RTP jitter algorithm. Loss is calculated using a moving average of packet loss with a window size of 100 packets. This value was chosen as it is the minimum value at which a resolution of 1% can be achieved. The loss value required by the E-Model uses a series of predefined loss values in increments of 1% to calculate the impact on user perceived quality.

Since only downlink data is being used to obtain network metrics, one-way-delay values cannot be accurately obtained using the VoIP data. Instead, a transport layer delay metric is used.

SCTP maintains a Smoothed Round Trip Time (SRTT) measurement for each address in the association. The SRTT is calculated using the RTT but incorporates a smoothing factor which has the effect of low pass filtering, thereby removing any delay spikes. In this work, we use the SCTP SRTT as the delay metric for MOS calculation. This provides a sufficiently accurate measurement of delay for calculating the MOS. Experimental verification of the accuracy of this approach was carried out, but due to space constraints these results are not presented.

A SRTT must be obtained for each address in the association. While the MN may have only one destination address specified in the association during handover, the CN will have two or more addresses specified – one for each IP address of the MN. Current SCTP implementations only retain state information for each active destination specified in the association. As the MN may only have one address for the CN specified in the association it cannot obtain SRTT information for both paths. Since the CN will have both interfaces of the MN specified as destination addresses in the association the SRTT for each path must be acquired at the CN. The SRTT values obtained at the CN are independently encapsulated into the downlink traffic over each link and transmitted to the MN. The MN then uses these SRTT values as the delay metric in the E-Model to calculate the MOS score for each candidate handover network.

C. *Final Stages of Handover Process*

To perform a handover the data flow must be redirected over the new access network. The handover is accomplished by sending a 'Set Primary' ASCONF message to the CN containing the IP corresponding to the new network. On successful reception of the ASCONF message the CN modifies the primary address through which it will communicate with the MN.

The final step in the handover process is to remove the old IP address from the association, so that no data is transmitted to the MN via the old access network which may no longer be available. Once again, an ASCONF message is transmitted to the CN containing the 'Delete IP' parameter.

V. VoIP Testbed

This section gives an overview of the experimental testbed and the client/server applications used for the experiments.

A. *Testbed Architecture*

As is shown in figure 2, the testbed for testing the performance of ECHO consists of two 802.11b WLAN access points, two desktop PCs to act as gateways and an Ethernet LAN on the University network. Also, included is a Mobile Node (MN) and a Correspondent Node (CN); these are the two endpoints between which handover takes place. The MN and CN were built on laptop computers running Fedora Core 5. The MN has two WLAN cards[2] to allow simultaneous connection to both APs giving the IP diversity required by SIGMA. The CN is single homed and connects directly to the Ethernet LAN.

The Linux Kernel implementation of SCTP (LKSCTP) was installed on each endpoint with both the Dynamic Address Reconfiguration (DAR) and Partial Reliability (PR) extensions enabled.

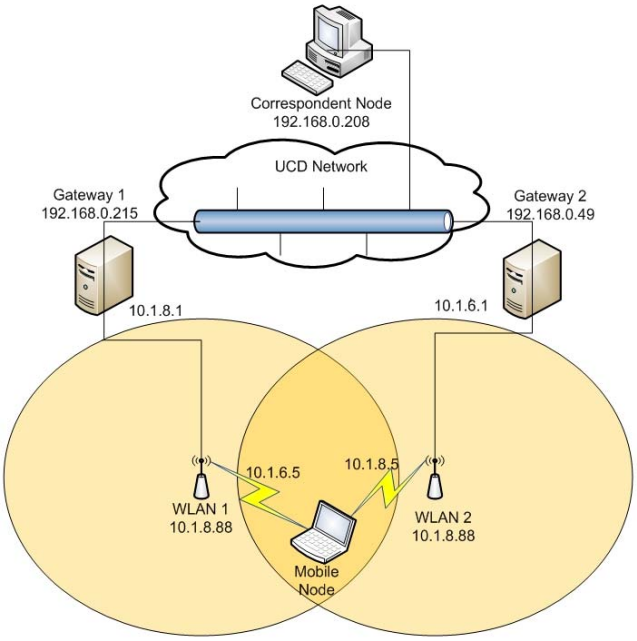

Figure 2 – Testbed Topology for ECHO Handover

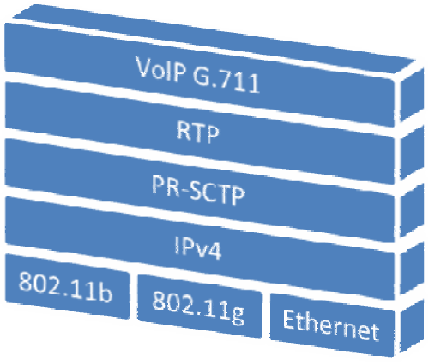

Figure 3 - Client/Server Stack

B. *Client/Server Application*

A Client/Server application that uses PR-SCTP and emulates Constant Bit Rate (CBR) full duplex VoIP data was developed. Each full duplex[3] stream is comprised of two simplex streams, one for both the uplink and downlink. The G.711 codec with a default frame size of 10ms is used as the

[2] Two different WLAN cards were used. An internal Broadcom 802.11b/g card and a 3Com PCMCIA 802.11a/b/g Card.

[3] The motivation behind using voice calls as full duplex was based on the most commonly used VoIP program Skype, which uses full duplex (i.e. no silence suppression is used). Skype uses full duplex for two reasons. Transmitting silent packets maintains UDP bindings at NAT (Network Address Translation). Also, if data is being transmitted over TCP, the silent period packets prevent a reduction in the congestion size during the silent period.

VoIP data source .The client encapsulates dummy VoIP data in RTP packets and transmits each packet to the server using PR-SCTP. Both the uplink and downlink VoIP streams are independent of one another; this realistically emulates a full duplex VoIP call.

The client and server applications run on the MN and CN, respectively. The system stack for both the client and server is shown in Figure 3.

VI. RESULTS

Each handover experiment involved the MN moving from AP1 toward AP2 at walking speed. The current study focused on WLANs for which walking speed is realistic, higher speed mobility would most probably require the use of longer range radio technologies. It is worth noting that as the MN moves away from an AP the RSS decreases and standard Link Adaptation (LA) occurs; however, this did not have any significant impact on results.

A. SIGMA Handovers

To illustrate non-optimal handover decisions in SIGMA, the following handover experiment was setup. The initial AP was uncongested and offers good QoS. The AP to which handover will take place is highly congested and is experiencing high packet loss.

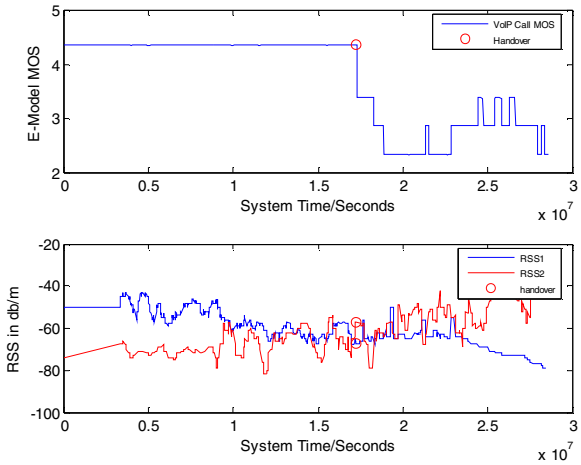

Figure 4 - SIGMA handing over to a network with high loss

The results of this experiment are shown in Figure 4, where the voice call quality measured in MOS is initially at the maximum attainable value of 4.4 with the G.711 codec. The MN moves at walking speed from the coverage area of AP1 towards AP2. As the MN moves further into the coverage area of the second AP, the RSS from AP2 becomes greater than that of AP1, at which point SIGMA performs a handover to AP2 at approximately 17 seconds. Since AP2 was highly congested, the call quality immediately dropped to an unacceptable level.

In this case, although the RSS of the alternative network was higher than that of the primary, it did not give an accurate reflection of the achievable call quality. Therefore, *there is a need for better handover schemes* that can consider the application and transport layer metrics to maximise the QoS obtained by the MN.

B. ECHO Handover

Figure 5 shows the result of an experiment for the same scenario as above, except that, the ECHO handover mechanism was used. When the RSS of the new network becomes greater than that of the existing network the QoS mechanism assesses the call quality that can be achieved over the new network. In this scenario, the secondary link provides a low quality MOS score of approximately 2.3. Since the call cannot be supported no handover takes place and the call is maintained over the existing AP with high quality even though the alternate network has a higher RSS.

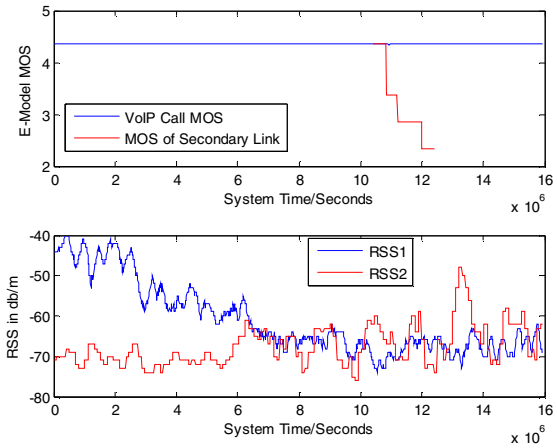

Figure 5 - MOS based handover to high loss network - No Handover

C. RSS Threshold

As was shown in Section VI-B above, by considering the achievable MOS a better handover decision can be attained and higher voice quality can be maintained. In the previous experiment, no handover took place as the call could not be supported by the secondary network. However, if the MN continues to move away from the primary AP the RSS will continue to decrease, eventually packet loss will begin and coverage will be lost. It is desirable to maintain connectivity even at reduced quality than to lose connectivity completely. To prevent loss of coverage, the MN must handover prior to moving completely out of the coverage area of the AP. Packet loss can give a good indication of being at the edge of a coverage area; however it is very difficult to differentiate between this type of packet loss and packet loss due to other network issues such as congestion. In this work an RSS threshold is used to estimate loss of coverage.

The RSS threshold is defined as the mean RSS at which packet loss begins to occur. To obtain this value, experiments were performed with the MN moving away from an 802.11b access point while measuring MOS, packet loss and RSS. One such experiment is shown in figure 6. As can be seen when an RSS of -76dbm is reached substantive application layer packet loss begins to occur and rapidly increases. Twenty such experiments were performed and the mean RSS at which packet loss begins to occur was found to be -78dbm.

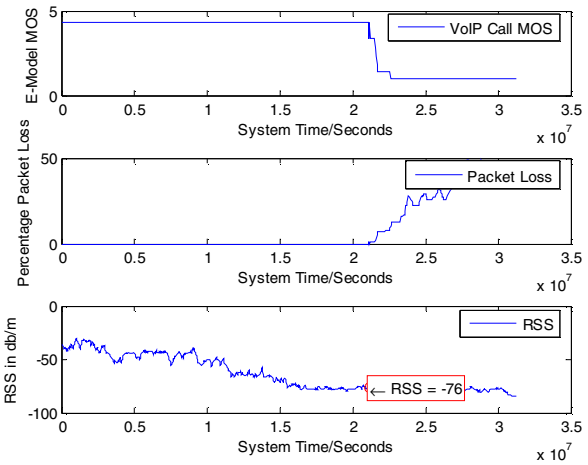

Figure 6 - Calculating the RSS Threshold

D. *QoS & RSS Handover*

The RSS threshold was then incorporated into the ECHO scheme as described in section IV of the paper. This addition allows connectivity to be maintained with reduced call quality. Figure 7 shows results utilising this feature. As in the previous experiments when RSS1 becomes less than RSS2 the quality of the secondary link is assessed and since it cannot support the call at a high quality no handover takes place. The MN then continues to move away from the primary access point. When the RSS threshold is reached on the primary link a handover is triggered. The handoff is performed before coverage is lost which would prevent the handover from taking place.

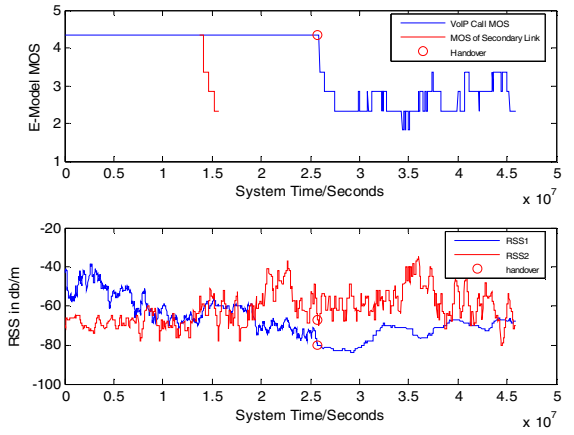

Figure 7 - RSS based Handover

In this scenario the QoS handover mechanism achieved a higher quality VoIP call for 11 seconds longer than would have been achieved if the traditional SIGMA handover process was used. After handover the call can continue albeit at lower quality. This is a better alternative than terminating the call due to lack of radio coverage.

VII. CONCLUSION

This paper presented an endpoint controlled QoS handover solution for VoIP based on the SIGMA mobility mechanism. It was shown that SIGMA's inability to consider QoS metrics resulted in poor handover decisions in cases in which there were congested networks. In the proposed ECHO scheme, multiple cross layer metrics that impact VoIP quality are obtained independently for each access network. These metrics are then mapped to user perceived quality using the E-Model algorithm to obtain a MOS value for each network. The resulting MOS value is used to make handover decisions.

Results from experimental evaluation show that the ECHO handover mechanism maintains high call quality by using the best available access network and minimising non essential handovers. This is a significant improvement over SIGMA.

Future work will investigate issues relating to simultaneous handovers as would be the case when multiple nodes need to affect handover at approximately the same time.

ACKNOWLEDGEMENTS

The support of the Irish Research Council for Science, Engineering and Technology (IRCSET) is gratefully acknowledged. The work of M. Atiquzzaman was supported by NASA grant NNX06AE44G.

REFERENCES

[1] C. E. Perkins, "Mobile IP," Communications Magazine, IEEE, vol. 35, no. 5, pp. 84-99, 1997.

[2] H. Fathi, S. Chakraborty, and R. Prasad, "Mobility management for VoIP: Evaluation of Mobile IP-based protocols," in Communications, 2005. *ICC 2005, IEEE International Conference on Communications 2005*, vol. 5, pp. 3230-3235 Vol. 5, 2005.

[3] J. Fitzpatrick, S. Murphy, M. Atiquzzaman, and J. Murphy, "Evaluation of VoIP in a mobile environment using an end-to-end handoff mechanism," *In Mobile and Wireless Communications Summit, 2007*. 16th IST, pp. 1-5, 2007.

[4] I. Marsh, B. Gronvall, and F. Hammer, "The design and implementation of a quality-based handover trigger," *5th IFIP Networking Conference*, pp. 580-591, 2006.

[5] IEEE Std p802.21/d02.00, "IEEE standard for local and metropolitan area networks: Media independent handoff services," tech. rep., September 2006.

[6] Q. Zhang, C. Guo, Z. Guo, and W. Zhu, "Efficient mobility management for vertical handoff between WWAN and WLAN," *Communications Magazine, IEEE*, vol. 41, no. 11, pp. 102-108, 2003.

[7] R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson, "Stream Control Transmission Protocol." RFC 2960 (Proposed Standard), Oct. 2000. Obsoleted by RFC 4960, updated by RFC 3309.

[8] R. Stewart, M. Ramalho, Q. Xie, M. Tuexen, and P. Conrad, "Stream Control Transmission Protocol (SCTP) Partial Reliability Extension." RFC 3758 (Proposed Standard), May 2004.

[9] ITU-T Recommendation G. 107, "The E-Model - A computational model in use in transmission planning," tech. rep., March 2003.

[10] S. Fu, M. Atiquzzaman, "SIGMA: A Transport Layer Handover Protocol for Mobile Terrestrial and Space Networks," *Invited book chapter in e-Business and Telecommunication Networks*, J. Ascenso, L. Vasiu, C. Belo, M. Saramago, (Eds.) Springer, 2006, pp. 41-52.

[11] R. Stewart, Q. Xie, M. Tuexen, S. Maruyama, and M. Kozuka, Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration." RFC 5061 (Proposed Standard), Sept. 2007.

[12] Psytechnics, "Estimating E-Model Id within a VoIP network," tech. Rep available from Psytechnics, 2002.