A Two-Phase Inter-Switch Handoff Scheme for Wireless ATM Networks

Khaled Salah
Dept. of CS, Illinois Institute of Technology
Tellabs Operations, Inc.
ksalah@tellabs.com

Elias Drakopoulos
Dept. of CS, Illinois Institute of Technology
Lucent Technologies
edrakopoulos@lucent.com

Abstract

Supporting mobility in Wireless ATM networks poses a number of technical issues. An important issue is the ability to reroute on-going virtual connections during handoff as mobile users move among base stations. We propose a two-phase inter-switch handoff scheme using permanent virtual paths reserved between adjacent Mobility Enhanced Switches (MES). The virtual paths are used in the first phase to rapidly reroute user connections. In the second phase, a distributed optimization process is initiated to optimally reroute handoff connections. The proposed handoff scheme yields small handoff latency and optimal routes while decreasing system cost and complexity. The paper also describes wireless control and ATM signaling capabilities required for supporting this scheme. Specific ATM UNI/NNI protocol extensions are presented. Using analysis we calculate and study the bandwidth requirement for the reserved virtual paths. We also study the second-phase optimization process overhead as well as the effect of other system design parameters.

1. Introduction

In future mobile communication networks, Wireless ATM (WATM) technology promises support for multimedia traffic such as voice, video, and data with QoS guarantees . A key feature of any wireless network is the ability to support the mobility of a user while maintaining communication. This requires the implementation of handoff. In WATM handoff, VC routes need to be modified as users move during the life time of a connection. The rerouting must be done fast enough with minimal disruption to traffic.

For the purpose of this paper, the network model shown in Figure 1 is adopted. This model has been used in project Magic WAND (Wireless ATM Network Demonstrator) [1] and is being used as a reference network configuration in the ATM Forum [2].

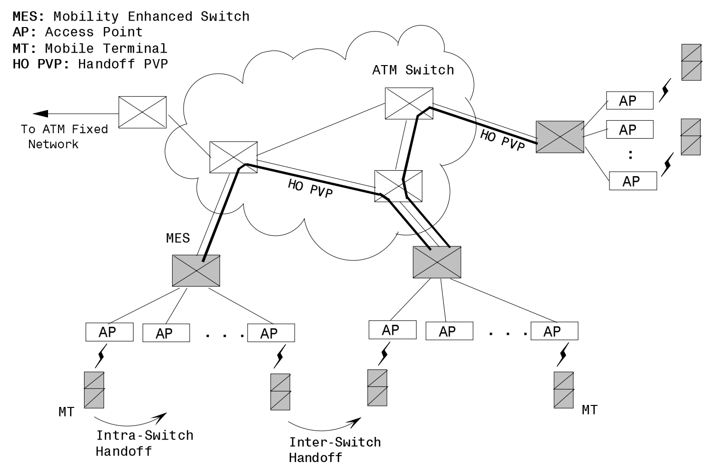

Figure 1 WATM network architecture

In order to solve the problem of rerouting user connections in WATM handoff, number of handoff schemes have been proposed. Two of the most well-known schemes are path extension [3], [4] and path rerouting [5], [6]. In path extension, the connection is extended from the old AP (Access Point) to the new AP. Pre-provisioned connections are typically established between APs in order to reduce connection setup time. While this scheme promises low rerouting latency, the resulting route is often not optimal. Also, it increases the complexity of the AP as it must have buffering and switching capabilities to all adjacent AP links. Increasing complexity of the AP will lead to increase in the total system cost as the AP will be one of the most widely deployed nodes. In path rerouting, a portion of the connection is rerouted at a Crossover Switch (COS). The COS is a rerouting point where the new partial path meets the old path. The idea is to re-use as much of the existing connection as possible, creating only a new partial path between the COS and the new AP. The scheme provides only partial route optimization and requires an implementation of a COS selection algorithm during handoff. The handoff latency of this scheme depends largely on the time involved in selecting the COS and the delay involved in setting up new VCs for the establishment of the new partial path. This delay will be highly variable and will depend on the number of intermediate switches and the processing load at each switch. The delay is more noticeable in the inter-switch handoff as the number of intermediate switches increases.

In this paper, we present an alternative solution in which we overcome these drawbacks. In the new scheme, Handoff Permanent Virtual Paths (HO PVPs) are provisioned between adjacent MESs to rapidly reroute user VCs during inter-switch handoffs eliminating the connection processing overhead and delays at intermediate switches. Therefore, the handoff latency is minimal. The rapid reroute of user VCs is followed by a non-realtime second phase in which a distributed route optimization procedure is initiated to find optimal paths. This scheme keeps AP complexity and cost low. The AP is simple and doesn't require having switching or buffering capabilities. It requires only mapping capabilities of user cells received on the wireless link to the wired link connected to the MES. Also, provisioning HO PVPs between adjacent MESs is more efficient in terms of bandwidth and management resources. It is more expensive to provision permanent connections between adjacent APs or between border APs and their adjacent MESs.

The rest of the paper is organized as follows. In Section 2, the protocol stacks for WATM nodes and signaling protocols are given. In Section 3, the first phase of the proposed scheme is described along with signaling protocols for both Intra- and Inter-Switch handoffs. Section 4 describes the route optimization of the second phase. Section 5 presents the analytical model used to evaluate the proposed scheme. Section 6 studies performance results. Finally, Section 7 concludes the paper.

2. Protocol Stacks

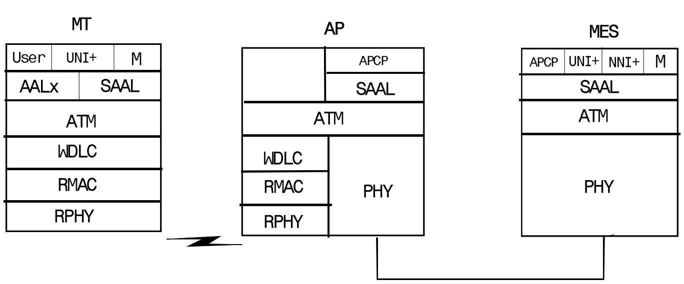

Figure 2 Protocol stacks for WATM nodes

Figure 2 depicts the protocol stack of the WATM network elements: MT, AP, and MES. The radio-physical (RPHY) and radio multiple access (RMAC) layers in the MT and AP are responsible for carrying ATM cells over the radio link. The wireless data link control (WDLC) layer provides error control mechanisms to enhance the performance of the radio channel. Similar to fixed ATM networks, end-to-end connections between mobile/fixed terminals are provided by the ATM adaptation layer (AAL) present in the end-user protocol stack.

It can be observed that ATM cells are exchanged all the way between end-user terminals, thus implying that ATM cells are carried over the radio interface. This approach yields a simple and homogeneous network architecture with end-to-end ATM cell delivery through standard ATM service access points.

In our network architecture, the MT uses the standard ATM signaling protocol [7], [8] (Q.2931, ATM Forum UNI) for connection establishment and handoff. Mobility-enhanced signaling protocol (named UNI+) is used by the MT to communicate with the MES. Access Point Control Protocol (APCP) is used for MES-AP communication. Communication between MESs is done using NNI+ protocol. Mobility control function, which may include registration, location update, and authentication uses "M" signaling protocol.

3. Description of the proposed scheme

In this section, we describe how the proposed two-phase handoff scheme can be applied to Intra-Switch handoff as well as Inter-Switch. Intra-Switch handoff occurs when an MT (Mobile Terminal) moves from an AP connected to an MES to another AP connected to the same MES. Inter-Switch handoff occurs when an MT moves from an AP connected to an MES to another AP connected to a different MES (see Figure 1). Assuming MT traffic is carried over a two-way VC (Virtual Connection), Intra-Switch handoff requires only one new VC to be established between the MES and the new AP, and the resulting route is optimal (assuming the original path to the MES was optimal). Since the new AP is directly connected to the MES, the HO PVP is not involved. Therefore for the Intra-Switch handoff, there will be no need to execute the handoff in two phases. The handoff is simple enough to be executed in one phase without the use of HO PVP. However, Inter-Switch handoff becomes more involved as more new VCs need to be set up. The number of new VCs is dependent on the network topology and may span number of ATM switches . With the use of HO PVP between adjacent MES, only two new VCs need to be established: one is within the HO PVP and the other is between the new MES and the new AP.

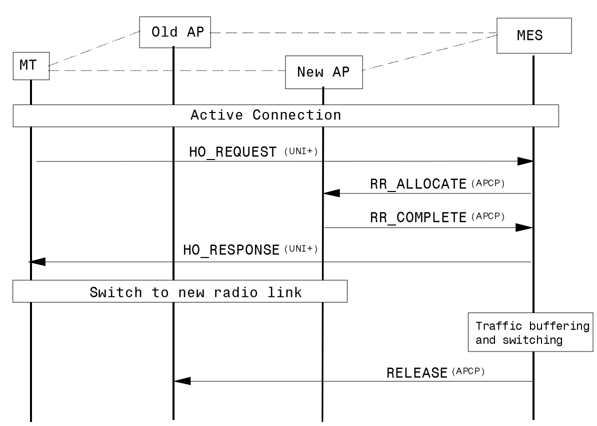

Figure 3 Intra-Switch handoff signaling protocol

A signaling protocol for Intra-switch handoff is shown in Figure 3. The figure also shows what signaling protocol is used to carry each message. The protocol for Intra-switch handoff can be described briefly as follows. During a call setup the user communication path to the MT is established using switched VCs. When the MT moves to a new cell, it determines a handoff is needed by using signal strength measurements. So, it sends to its MES (via its AP) a HO_REQUEST message requesting a handoff to a new AP. The MES upon reception of the HO_REQUEST allocates a new VC for the new AP. The MES requests the new AP (using RR_ALLOCATE message) to allocate radio resources according to expected QoS and bandwidth requirement. The VC allocation is completed when RR_COMPLETE message is received by the MES. The MES then returns to the MT a handoff response message via the old AP. The handoff response message includes the new VC id and possible QoS modifications. The MT then establishes a new radio link with the new AP. Buffering functions need to be performed at the MT and MES to coordinate switching of traffic to ensure in-order delivery of cells and no cell loss. An example of switching and buffering between two nodes will be illustrated in Section 3. Finally, old VC and radio resources are released.

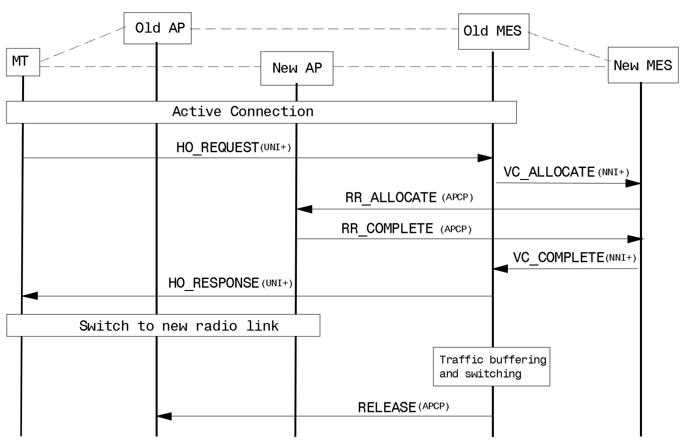

Figure 4 Inter-Switch handoff signaling protocol

In case of Inter-Switch handoff, the signaling protocol is similar except the new MES is involved (see Figure 4). When the old MES determines that the new AP is connected to an adjacent MES. The old MES sends VC_ALLOCATE message to the new MES. The new MES allocates two VCs: one between itself and the new AP and the other within the HO PVP. After a successful Inter-switch handoff, a request for route optimization is initiated. The route optimization is described next.

4. Route Optimization

In order to optimize the connection route resulted from the rapid rerouting using HO PVP, a non-realtime route optimization is executed by the new MES. We propose a distributed route optimization procedure in order to distribute processing overhead and minimize signaling to a centralized node. The urgency for route optimization is dedicated by several factors: optimality of the current path, utilization of HO PVP, QoS degradation, call duration (being old or new), number of hops, loop detection, etc. Once the need for optimizing the new handed-off path is detected, the route optimization procedure is executed.

The protocol for the route optimization procedure is described in the following steps:

1. The new MES requests path information of the handed-off connection from the old MES. Path information is requested using an ID that uniquely identifies the handed-off connection. The ID is provided to the new MES during handoff in the HO_REQUEST message. The requested information includes connection QoS parameters, source and destination ATM addresses, and a list of addresses for all candidate crossover nodes along the path. A crossover node in this case is basically a regular ATM switch which has the added functionality of coordinating traffic switching and buffering with the new MES. The list of candidate crossover nodes is built during connection original establishment. Current ATM Forum and ITU-T standards for UNI and NNI signaling can support building such a list. Call SETUP and CONNECT messages can carry such information as the original connection segments are built hop by hop. If the MT, the local host, is the called/destination node, the MES will extract the list from the SETUP message. However if the MT is the caller/source node, the MES will extract the list from the CONNECT message. A crossover node along the path processes the SETUP or CONNECT message and adds its address to the message using additional IEs (Information Elements). A non-crossover node merely processes the message and passes it to the next node. A crossover node adds its address to the message if the message has one or more crossover IEs. The initial crossover IE is added by the crossover node nearest to the remote host. If the remote host is the caller, the IE is added in the SETUP message, otherwise it is added in the CONNECT message. The crossover node nearest to the remote host uses location management information and addressing to determine if the end host is mobile or fixed. The list of candidate crossover nodes is kept in hierarchical order, i.e. the first node on the list means that the node is nearest to the remote host.

2. Based on path information received from the old MES, the new MES performs COS discovery. This scheme is similar to Prior Path Knowledge COS discovery scheme proposed in [10], however no centralized connection server is used in our proposed procedure. In order to find the optimal path, the shortest path from the new MES to all candidate crossover nodes in the list is calculated. Since the PNNI routing scheme is a link-state routing scheme (and not a “distance-vector” scheme), this operation can be computed using the existing PNNI protocol [11]. The candidate crossover node with the shortest path will be selected as the crossover node. If multiple candidate crossover nodes have the

same shortest path (e.g. minimum-hop count), then the node nearest the remote host will be selected.

3. The new MES then probes the selected crossover node for optimization rerouting. A crossover node receiving the rerouting message will accept or deny the request based on its own knowledge of network topology and state. If the selected crossover node denies the request, another crossover node (one next to the best) is probed.

4. The new MES then builds the best route to the selected crossover node in the from of a hierarchically complete source route known as a Designated Transit List, or DTL, as specified in [11]. The establishment of the new connection segment between the selected crossover node and the new MES can be initiated by either the crossover node or by the new MES. In order to minimize signaling of path information to the crossover node and allow for faster selection of another crossover node in case of segment setup failure, the establishment of the new segment is initiated by the new MES.

5. After the new segment has been set up, buffering and switching functions need to be performed at the new MES and crossover node to ensure lossless rerouting. The new MES and crossover node will use in-band signaling prior to connection switch-over. For example in the ingress direction (towards the crossover switch), when the new MES receives successful segment establishment from the crossover node, it sends a special "Tail" signal cell after the last user cell on the old connection segment. "Tail" signals are special cells sent on the same VC as the user cells (in-band signals). They could be RM (Resource Management) cells. If user cells arrive at the crossover node on the new segment prior to the reception of the "Tail" signal, they will be buffered. These buffered cells will be sent after the reception of the "Tail" signal. Similarly, switching and buffering can be done in the egress direction (towards the new MES.) In-band signaling was used in [9] to implement lossless handoff.

6. Lastly, the old path segment is released. This may include the release of the VC within the HO PVP, if it is not part of the new segment. Since the HO PVP is a critical resource, releasing its VCs need to be done first. Also database information about the connection is deleted from the old MES and stored in the new MES with updates to connection parameters and the list of candidate crossover nodes

It is worth noting that the optimization phase can be transparent to both the AP and MT, unless there is a need for QoS re-negotiation or a need for bandwidth adjustment which requires the MT involvement.

5. HO PVP Bandwidth and Optimization Rate

Two important design parameters in the proposed scheme are the required bandwidth for HO PVPs and the processing overhead for route optimization at the MES. In this section, we analytically study these parameters. In [12], the handoff call arrival rate in a cell is given as follows:

$$\lambda_h = \frac{(1-P_0)\mu_R[1-M^*(\mu_R)]\lambda_0}{\mu_M[1-(1-P_f)\ M^*(\mu_R)]}$$

where:

- P_0 : The originating call blocking probability
- P_f : The handoff blocking probability
- λ_0 : The originating arrival call rate in a cell which follows a Poisson process.
- $1/\mu_M$: The mean of holding time of a call T_M. T_M has exponential distribution.
- $1/\mu_R$: The mean of residual time of a call T_R in a cell. T_R has a general distribution.
- $M^*(\mu_R)$: The Laplace-Stieltjes transform (LST) of T_R.

In addition, the following assumptions are made: 1) MT calls carry only CBR traffic, i.e. all have the same QoS and traffic characteristics. Therefore the required bandwidth for each call is the same and has a fixed size. 2) Each call uses one VC and each VC uses one radio channel. The bandwidth of the HO PVP is measured by the number of VCs/radio channels. 3) VC allocation never causes call blocking for originating calls or during route optimization. 4) Radio resources are sufficient not to cause blocking during handoff. 5) All inter-switch handed-off connections need route optimization.

Under the above assumptions, the handoff blocking probability P_f due to lack of VCs in HO PVP can be expressed using Erlang-B formula:

$$P_f = \frac{\frac{[\lambda_S E(T_S)]^{N_S}}{N_S!}}{\sum_{n=0}^{N_S} \frac{[\lambda_S E(T_S)]^n}{n!}}$$

Where N_S is the number of VCs or channels, λ_S is the total inter-switch handoff request rate, and $E(T_S)$ is the expected holding time of one VC.

To find λ_S, we assume a plane of square-shaped cells with MT uniform movement in the four direction of each cell. Handoffs may happen across a cell boundary in two directions. Assuming J cell boundaries contributing to the total inter-switch handoff, then $\lambda_S = J \cdot 2 \cdot \lambda_h/4$. For the purpose of this analysis, we will assume each MES covers 4 cells in an office environment (see Figure 5). Therefore, $J = 2$ and $\lambda_S = \lambda_h$. Note that in a highway environment where coverage area is made of square-shaped cells laid out in a rectangular fashion, $\lambda_S = \lambda_h$, since movement in a cell occurs only in two directions and $J = 1$.

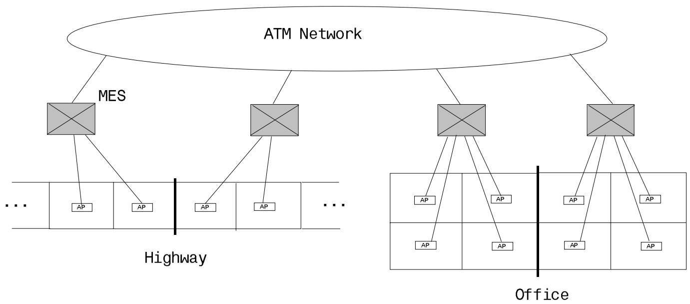

Figure 5 Inter-switch cell boundaries in office and highway environments

Now we find $E(T_S)$. Suppose that MT moves across one of the inter-switch cell boundaries and has a successful first-phase handoff, i.e. a new VC got established in the HO PVP. This VC will remain established until it is released due to one of the following: 1) call completion, 2) route optimization, or 3) next handoff blocking. Hence, The VC holding time T_S can be written as:

$$T_S = \min(T_M, T_Z, T_R)$$

Where:

- T_M is the holding time of a call/connection.
- T_Z is the route optimization time of one VC within the HO PVP. The route optimization time can be approximated by an $M/M/1$ queue with λ_S handoff rate and mean optimization rate of μ_z. Hence, $F_{Tz}(t) = 1 - e^{-(\mu_z - \lambda_S)t}$ is the distribution function of T_Z. For simplicity, it is assumed that the route optimization will always result in releasing the VC connection in the HO PVP.
- T_R is the residual time of a call in a cell. T_R as proposed by [4] has a geometric probability distribution and is the total sojourn time of N cells where MT generating the call resides before the next handoff blocking.

The distribution of T_S can be found as:

$$F_{Ts}(t) = 1 - [1 - F_{TR}(t)]\; e^{-\mu_M t} \cdot e^{-(\mu_z - \lambda_s)t}$$

The LST of $F_{Ts}(t)$ can be written as:

$$S^*(x) = \frac{v(0)}{v(x)} + \left[1 - \frac{v(0)}{v(x)}\right] R^*(v(x))$$

where: $v(x) = \mu_M + \mu_z - \lambda_S + x$,

$$R^*(v(x)) = N[M^*(v(x))] = \frac{M^*(v(x)) \cdot P_f}{1 - (1 - P_f)\; M^*(v(x))}$$

Therefore, $E(T_S)$ can be expressed as:

$$E(T_S) = \frac{1 - M^*(\mu_M + \mu_z - \lambda_s)}{(\mu_M + \mu_z - \lambda_s)[1 - (1 - P_f)\; M^*(\mu_M + \mu_z - \lambda_s)]}$$

If T_R is exponentially distributed, then:

$$M^*(\mu_M + \mu_z + \lambda_s) = \frac{\mu_R}{(\mu_M + \mu_z + \lambda_s) + \mu_R} \quad ,$$

$$\lambda_s = \frac{\mu_R\,(1 - P_0)\;\lambda_0}{\mu_M + \mu_R P_f} \quad ,$$

and $E(T_S)$ can be simplified to:

$$E(T_S) = \frac{1}{(\mu_M + \mu_z - \lambda_s) + \mu_R P_f}$$

6. Numerical Examples

In this section we study the performance of the proposed scheme as a function of system offered load. In particular, we examine the required bandwidth for HO PVP and the processing overhead at the MES due to route optimization. We assume the mean residual time of 4 minutes and a call holding time of 2 minutes. Originating calls are assumed to be blocked with probability of 0.01, while handoff blocking probability is assumed to be 0.001. Mean route optimization times are chosen to be 2.0, 1.9, and 1.5 Sec. These time estimates for the execution of the route optimization procedure are based on the VC setup measurements for the existing ATM switches. Switched VC setup latency through one node ranges from 10 ms to 125 ms [13]. Hence, if we assume an extreme case of a fairly large ATM network, the latency involved in establishing the new segment (step 4 of the route optimization procedure) among 200 nodes with average of 10ms VC setup latency per node would be 2 Sec. The latency due to signaling and overhead processing involved in other steps of the optimization procedure would be dependent on the number of crossover nodes. We assume this time will not take more than 100 ms.

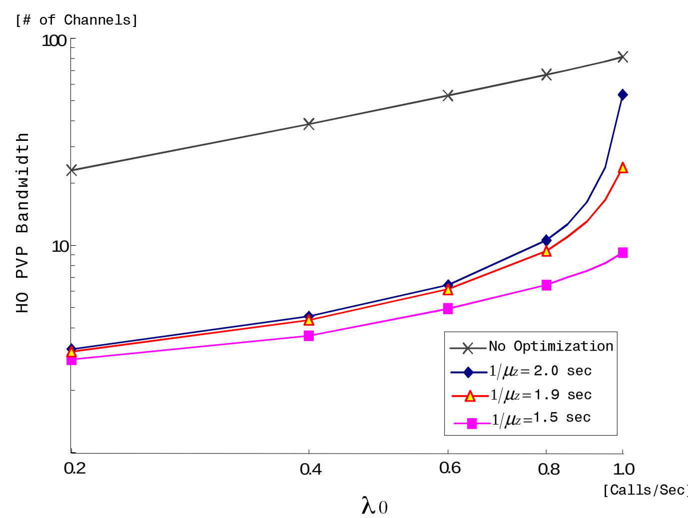

Figure 6 Required HO PVP bandwidth as a function of originating call rate.

We first study the required HO PVP bandwidth as a function of the originating call rate. Figure 6 shows the required HO PVP bandwidth for different values of the mean route optimization time and when the route optimization process is turned off. The figure illustrates the trade-off that exists between HO PVP bandwidth and optimization rate. In heavy load region $(\lambda_0 > 0.7)$, the HO PVP bandwidth increases considerably as the optimization rate decreases. While in light load region $(\lambda_0 < 0.7)$, increasing the optimization rate results only in marginal reduction in the reserved bandwidth. Thus, based on the above analysis a suitable operating point for the bandwidth and optimization rate can be engineered.

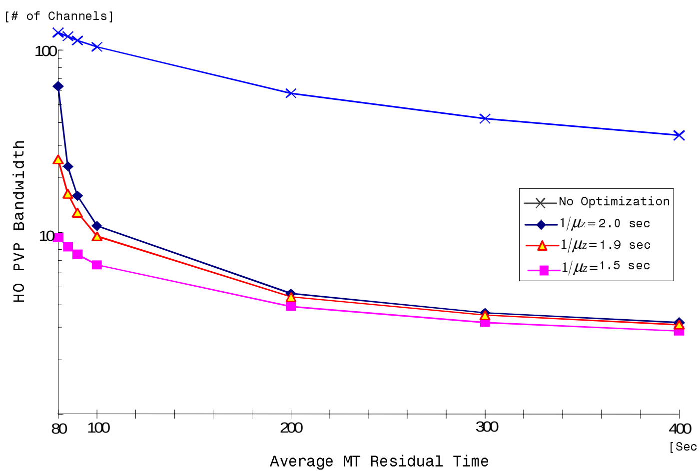

Figure 7 Required HO PVP bandwidth as a function of average residual time

The impact of MT residual time when $\lambda_0 = 0.2$ is shown in Figure 7. We vary the residual time from 80 to 410 seconds. The mean call holding time in this case is 200 seconds. The figure shows the required bandwidth for different values of $1/\mu_z$. When mean residual time is greater than 160 seconds, varying optimization rate at MES has small impact on the reserved bandwidth. It is to be noted, in both Figure 6 and Figure 7, that the reserved bandwidth in light load is small. This is due to the fact that the VC holding time within the HO PVP becomes smaller due to optimization rerouting. Under our assumptions, the optimization procedure gets always executed for every inter-switch handoff. In order to keep the HO PVP bandwidth better utilized in light load, optimization rate must be decreased. Optimization rate might be best studied as a function of the reserved bandwidth utilization. This issue is under on-going study.

7. Conclusion

We have proposed a two-phase inter-switch handoff scheme for Wireless ATM networks. Signaling and control protocols to support the two phases were described. The proposed handoff scheme does not require a complex AP or impose stringent latency requirement on COS selection algorithm, but utilizes reserved virtual paths between adjacent MES to rapidly reroute user connections for inter-switch handoffs. Optimal paths are accomplished in the second phase using a distributed rerouting optimization process carried out by the new MES. The required bandwidth for the HO PVP and the overhead at MES associated with optimization process were studied analytically and shown to be of small significance in light load. Our results indicate that a simple fast handoff phase followed by a route optimization phase is sufficient for supporting handoff in WATM networks.

References

[1] J. Mikkonen and J. Kruys, "The Magic WAND: a wireless ATM ATM access system", Proceedings ACTS Mobile Telecommunication Summit, Granda, Spain, 1996.

[2] Baseline Text for Wireless ATM Specifications, BTD-WATM-01.07, ATM Forum, WATM WG, April 1998.

[3] P. Agrawal, et. al, "SWAN: A Mobile Multimedia Wireless Network," IEEE Personal Communications Magazine, April 1996

[4] S. Lee and D. Sung, "A New Fast Handoff Management Scheme in ATM-based Wireless Mobile Networks", GLOBECOM '96.

[5] A. Massarella, "Wireless Mobile Terminal/Network Anchor Switch Handover Model," ATM Forum/97-0265/WATM.

[6] P. Shieh, T. Tedijanto, and R. Rennison, "Handover Schemes to Support Mobility in Wireless ATM," ATM Forum/96-1622/WATM.

[7] ITU Study Group 11, Q.2931, Caluse 1, 2, and 3. S. Shiraishi and H. Yao, June 1994.

[8] ATM User-Network Interface Specification, Version 3.0 ATM Forum, Prentice Hall, 1993.

[9] H. Mitts, et al., "Lossless Handover for Wireless ATM," ACM/Blatzer Mobile Networks and Appls., Dec. 1996.

[10] C-K Toh, "Crossover Switch Discovery Schemes for Fast Handovers in Wireless ATM LANs", ACM Journal on Mobile Networks & Applications - Special Issue on "Routing in Mobile Communication Networks", Vol 1, No 2, Nov. 1996.

[11] Private Network-Network Interface Specification Version 1.0 (af-pnni-0055.0000), ATM Forum, March 1996.

[12] Y. Lin, S. Mohan, and A. Noerperl, "Queueing priority channel assignment strategies for PCS hand-off and initial access," IEEE Trans. Veh. Tech., vol. 43, pp 704-712, Aug. 1994.

[13] The ATM Report, "Inside: The ATM Report Buyer's Guide to ATM Backbone Switches," February, March, April 1997.