Performance Study of a Two-Phase 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

Tzilla Elrad
Dept. of CS, Illinois Institute of Technology.
elrad@charlie.cns.iit.edu

ABSTRACT. *This paper presents an analytical and simulation study of a two-phase handoff scheme for rerouting user connection in Wireless ATM networks. The two-phase handoff scheme provides a rapid rerouting of user connections in the first phase utilizing permanent virtual paths reserved between adjacent Mobility Enhanced Switches (MES). In the second phase, a non-realtime route optimization procedure is executed to optimally reroute handed-off connection. In this paper, we study the performance of such a scheme as a function of various system load parameters. These parameters include originating call arrival rate, call holding time, and radio cell residual time. We examine the relation between the required bandwidth resources and optimization rate. Also we calculate and study the handoff blocking probability due to lack of bandwidth for resources reserved to facilitate the rapid rerouting.*

1. INTRODUCTION

Wireless Asynchronous Transfer Mode (WATM) technology combines two of the hottest technologies in communication these days: wireless and ATM. WATM will provide multimedia traffic for mobile terminals with high quality of service. However, WATM faces many technical challenges. One of the most important is supporting mobility of the user while maintaining communication. This requires the implementation of handoff. In WATM handoff, connections need to be modified as users move from one radio cell to another. The rerouting of connections must be done quickly with minimal disruption to traffic. Also the resulting routes must be optimal [1]. Figure 1 shows the WATM network model and its network elements.

A number of schemes to reroute connections during WATM handoff has been proposed in literature. Two well-known schemes are path extension [2, 3, 4] and path rerouting [5, 6, 7]. 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. The AP must be capable of managing pre-provisioned connections, and 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 node 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 connection segments 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 and study a two-phase handoff scheme in which Handoff Permanent Virtual Paths (HO PVPs) are provisioned between every two adjacent MESs. The HO PVPs, shown in Figure 1, are used to rapidly reroute user connections during inter-switch handoffs eliminating the connection processing load and delays at intermediate switches. Therefore, the handoff latency is minimal. Also HO PVPs reduce system cost as they eliminate the need for additional physical connections between adjacent (Mobility Enhanced Switches) MESs. The rapid reroute of user connections is followed by a non-realtime second phase in which a 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 and manage 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 two-phase handoff scheme is briefly described for both intra- and inter- switch handoffs. Section 3 describes the route optimization of the second phase. Performance study using analysis is presented in section 4. Section 5 presents a simulation study. In section 6, results obtained from analysis and simulation are discussed. Finally, section 7 contains the conclusion.

2. TWO-PHASE HANDOFF

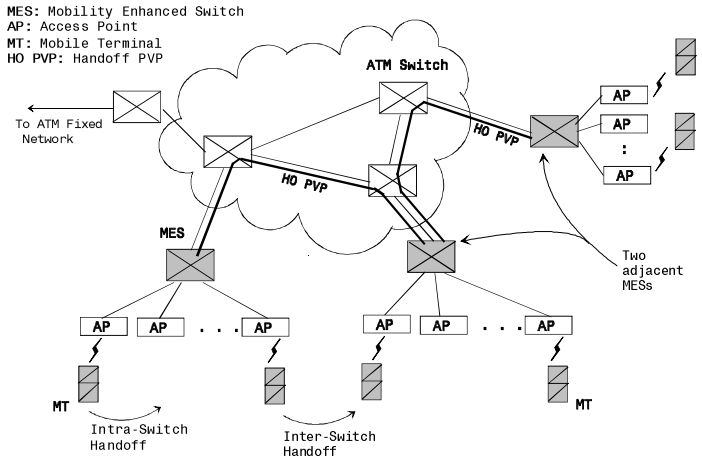

Figure 1 WATM network model

In this section, we briefly describe the two-phase handoff scheme. We describe how the 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. Intra-switch handoff requires only one new connection 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. Hence, for intra-switch handoff, there will be no need to execute a second phase. However, inter-switch handoff becomes more involved as more new connections need to be set up. The number of new connections is dependent on the network topology and may span number of ATM switches . With the use of HO PVP between adjacent MES, the management and establishment of new connections are simplified. Only two new connections need to be established and managed: one is within the HO PVP and the other is between the new MES and the new AP. After a successful rapid inter-switch handoff, a request for route optimization is initiated.

3. ROUTE OPTIMIZATION

In order to optimize the connection route resulting from rapid rerouting using HO PVP, a non-realtime route optimization is executed by the new MES. The route optimization procedure can be described as follows: 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 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 or COS 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 original connection establishment. 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 [8], however no centralized connection server is used. In order to find the optimal path, the shortest path from the new MES to all candidate crossover nodes in the list is computed. The new MES then builds a new connection segment between itself and the selected COS. Buffering and switching functions are then 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. Lastly, the old path segment is released. This will include the release of the connection within the HO PVP.

Based on the description of the route optimization procedure above, signaling and processing load would be imposed on the WATM network. In particular, processing load would be imposed on the MES and crossover nodes, and signaling messages would be exchanged between new and old MES as well as between new MES and crossover nodes. We will study this optimization overhead in relation to the required HO PVP bandwidth. The optimization overhead will be represented by the optimization rate μz .

4. ANALYSIS

The performance of the two-phase handoff scheme is studied in this section using analysis. The following assumptions are made:

1) Each call uses one connection. Every call/connection has an identical bandwidth requirement.
2) Each connection is bi-directional. This means a connection has two virtual circuits or VCs.
3) Resource 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 require route optimization.

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

$$P_f = \frac{\frac{[\lambda_S E(T_S)]^{N_S}}{N_S!}}{\sum_{n=0}^{N_S} \frac{[\lambda_S E(T_S)]^n}{n!}}, \qquad (1)$$

where N_S is the number of connections in the HO PVP, λ_S is the total inter-switch handoff request rate, and $E(T_S)$ is the expected holding time of a connection in the HO PVP.

First we find λ_S, the total inter-switch handoff request rate. In [9], the handoff call arrival rate in a radio cell is given as follows:

$$\lambda_h = \frac{(1-P_0)[1-R^*(\mu_M)]\lambda_0}{\mu_M \; E(R)\,[1-(1-P_f)\,R^*(\mu_M)]},$$

where:

- P_0: The originating call blocking probability
- P_f: The handoff blocking probability, (i.e. the probability that call is dropped due to lack of bandwidth.)
- λ_0: The originating call arrival rate in a cell. It follows a Poisson process.
- $1/\mu_M$: The mean of holding time of a call T_M. T_M has exponential distribution.
- $E(R)$: The mean residual time R of a call in a cell. The cell residual time is the time the MT resides in a cell before it moves out to another cell. R has a general distribution. The cell residual times, $R^{(1)}, R^{(2)}, R^{(3)} \cdots$, resulting from the movement of the MT, are all random variables which are independent and identically distributed.
- $R^*(s)$: The Laplace-Stieltjes transform (LST) of the random variable R.

We assume a generic environment consists of hexagonal-shaped cells with uniform movement in all six directions. The handoff rate across any cell boundary, contributed by one cell, is $\lambda_h/6$. As shown in Figure 2, there are three cell boundaries contributing to the total inter-switch handoff. Therefore $\lambda_S = 3 \cdot 2 \cdot \lambda_h/6$, and hence $\lambda_S = \lambda_h$.

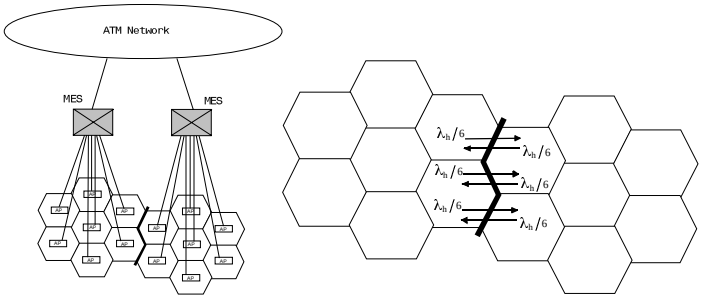

Figure 2 Inter-switch cell boundaries and handoff rates

Now we find $E(T_S)$. Suppose the MT moves across one of the inter-switch cell boundaries and has a successful first-phase handoff, i.e. a new connection got established in the HO PVP. This connection will remain established until it is released due to one of the followings:

1. Route optimization (executed at a mean rate of μ_Z).
2. Call holding time expiration.
3. Handoff blocking as a result of MT journey.

Hence, the connection holding time T_S within the HO PVP can be written as:

$$T_S = \min(T_M, T_Z, T_R), \qquad (2)$$

where:

- T_M is the holding time of a call/connection. Since T_M has exponential distribution, $F_{T_M}(t) = 1 - e^{-\mu_M t}$.
- T_Z is the route optimization time of one connection for a single HO PVP. According to our proposed route optimization procedure, the initiation of optimization for handed-off connections within a single HO PVP is performed by the two adjacent MESs. Hence, μ_Z is distributed between these two adjacent MESs. We assume that μ_Z is divided evenly between the adjacent MESs, with each MES having a mean optimization service rate of $\mu_Z/2$. As for λ_S, it is also divided evenly among these two MESs. This is so because every MES performs route optimization for the "incoming" handed-off connections. The term "incoming" refers to handed-off connections towards the MES. Handed-off connections towards the other MES will be considered "departing" connections and will be handled by the other adjacent MES. At the inter-switch cell boundaries the incoming and departing handoff rates are equal, since movement within a cell was assumed to be uniform. So for each MES, the mean optimization request rate is $\lambda s/2$. Therefore, one can approximate the optimization process by two independent or parallel *M/M/1* queues with each having a mean service rate of $\mu_Z/2$ and a mean arrival rate of $\lambda s/2$. Hence, the two independent *M/M/1* queues are equivalent to one *M/M/1* queue with $\rho = \lambda_S/\mu_Z$. The distribution function of T_Z is given by $F_{T_Z}(t) = 1 - e^{-(\mu_Z - \lambda_S)t}$. For simplicity, it is assumed that the route optimization will always result in releasing the connection.
- T_R is the total sojourn time of N cells where MT generating the call resides before handoff blocking.

The distribution of T_S can be expressed as

$$F_{T_S}(t) = 1 - [1 - F_{T_R}(t)]\, e^{-(\mu_M + \mu_Z - \lambda_Z)t}.$$

By the definition of LST properties,

$$E(T_S) = -T_S^{*'}(0), \text{ and}$$

$$T_S^{*}(x) = \int_0^{\infty} e^{-xt}\, dF_{T_S}(t).$$

Let $v(x) = \mu_M + \mu_Z - \lambda_Z + x$, then

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

Next we find $T_R^{*}(v(x))$. Remember that T_R is the total residual time of N cells before handoff blocking. This means $T_R = R^{(1)} + R^{(2)} + R^{(3)} + \cdots + R^{(N)}$. R is the cell residual time in a cell. Note that N is the number of cells the MT resides in before the handoff blocking. Therefore N is a random variable and has a geometric distribution. And thus

$$P(N = n) = P_f (1 - P_f)^{n-1}, \quad n = 1,2,3...$$

The LST of T_R is given by

$$T_R^{*}(s) = N[R^{*}(s)]$$

where $N[R^{*}(s)]$ is the generating function of the random variable N, and described as

$$N[R^{*}(s)] = \frac{p_f\, R^{*}(s)}{1 - (1 - p_f) R^{*}(s)}.$$

Therefore

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

Taking the derivative of $T_S^{*}(x)$ and evaluating x at 0, we get

$$E(T_S) = \frac{1 - R^{*}(\mu_M + \mu_Z - \lambda_Z)}{(\mu_M + \mu_Z - \lambda_Z)[1 - (1 - P_f) R^{*}(\mu_M + \mu_Z - \lambda_Z)]}.$$

R has a general distribution. If R has an exponential distribution, then

$$R^{*}(s) = \frac{\mu_R}{s + \mu_R},$$

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

$$E(T_S) = \frac{1}{(\mu_M + \mu_Z - \lambda_Z) + \mu_R P_f}. \qquad (3)$$

Special Case:

Let us consider a special case when the route optimization process is turned off. This means that the connection within the HO PVP is released due to two of the following conditions: 1) call completion or 2) handoff blocking. Hence, the connection holding time T_S can be written as:

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

Carrying out the previous derivations, we get

$$E(T_S) = \frac{1}{\mu_M + \mu_R P_f}. \qquad (4)$$

Applying numerical operations to Eq. (1), (2), (3), and (4), one can find N_S and P_f.

5. SIMULATION

A simulation model, shown in Figure 3, is developed to study and validate analysis performance results. In our simulation, a user connection will remain established in the HO PVP until it is released due to one of the following events: 1) HOLDINGTIME_END, 2) SOJOURN_BLOCKING, or 3) DEPARTURE. HOLDINGTIME_END event is the expiration of the holding time of a call. The SOJOURN_BLOCKING is the event for handoff blocking due to MT journey. It is the total sojourn times of N cells that the MT visits before handoff blocking. N is a random variable and has a geometric distribution. Route optimization procedure is simulated as an *M/M/1* queue with DEPARTURE and ARRIVAL events.

The simulation first chooses a λ_h value (e.g. 0.1 λ_0), then simulates the behavior of the handoff procedure to obtain P_f. A new λ_h value is computed, and a new simulation iteration is conducted using the new λ_h value. The procedure repeats until λ_h converges.

The details of the simulation are given in Figure 3. Step 1 initializes the simulation. Then the first handoff ARRIVAL event is generated. The next event is removed from the Event_priority_Queue in step 2 and is processed based on its type in step 3. The simulation clock is advanced to the time of the event. The Event_priority_Queue is the queue of events and its priority is based on time.

For an ARRIVAL event, N is incremented and the next ARRIVAL event is generated (step 4). The capacity of HO PVP is checked in step 5. If bandwidth is not available, meaning the HO PVP capacity is equal to maximum available bandwidth, then the call is blocked and *Nb* is incremented (step 6). Otherwise, we use one connection of bandwidth and hence we increment the HO PVP capacity and generate HOLDINGTIME_END and SOJOURN_BLOCKING events. Both events are inserted in the Event_priority_Queue (step 8). Immediately we begin serving the handoff arrival for route optimization. If the route optimization server is busy (step 8), we insert the event into the Optimizer_FIFO_Queue, otherwise the

server is idle and the call can be served instantly. Hence, we make the server busy and generate a DEPARTURE event (step 10). The DEPARTURE event is inserted into the Event_priority_Queue.

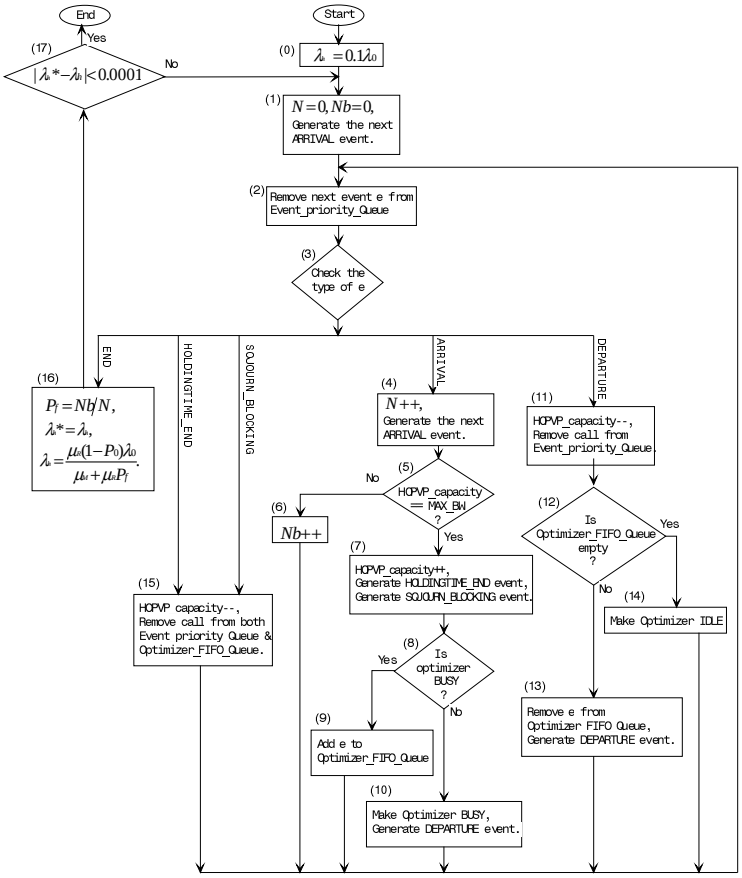

Figure 3 The simulation flow chart

A DEPARTURE event indicates that the route optimization server has completed. In such an event, we release the used bandwidth by decrementing the HO PVP capacity and remove any events for that call from Event_priority_Queue (step 11). Events that might exist for that call include HOLDINGTIME_END and SOJOURN_BLOCKING events. Events for a particular call are identified by a handoff_id field, which is part of the event data structure. Step 12 then checks to see if any other event is waiting to be served in the Optimizer_FIFO_Queue. If not, we make the route optimization server idle (step 14); otherwise, we remove an event from the Optimizer_FIFO_Queue and compute its route optimization completion time and generate for it a DEPARTURE event (step 13).

A call may also terminate when either a HOLDINGTIME_END or SOJOURN_BLOCKING event occurs. In such case (step 15), we first release the used bandwidth by decrementing the HO PVP capacity. We then remove any events for that call from both Event_priority_Queue and Optimizer_FIFO_Queue. Events that might exist for that call in the Event_priority_Queue include HOLDINGTIME_END or SOJOURN_BLOCKING events. If a DEPARTURE event existed for that call, it should not be deleted, because our optimization server is assumed to be a non-preemptive server.

For an END event, the simulation iteration terminates and P_f and the new λ_h are computed (step 16). The new λ_h value is compared with the old λ_h* value (step 17). If the absolute difference is within 0.1%, then the simulation terminates. For our numerical examples, we ran each simulation iteration for 200 hours, i.e. END event time was 200 hours.

6. NUMERICAL EXAMPLES

In this section we study the performance of the two-phase handoff scheme as a function of system offered load. These parameters include mean originating call arrival rate, call holding time, and residual time. We examine the relation between the required HO PVP bandwidth and optimization signaling and processing load. Also, we study the handoff blocking probability due to lack of HO PVP bandwidth.

For our numerical examples, we assume a mean cell residual time of 6 minutes and a mean call holding time of 3 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 varied from 1.3 to 0.6 Sec. We assume these times are sufficient to carry out the processing and signaling load involved in the optimization procedure. The dashed line in the figures represent some of the results obtained by simulation.

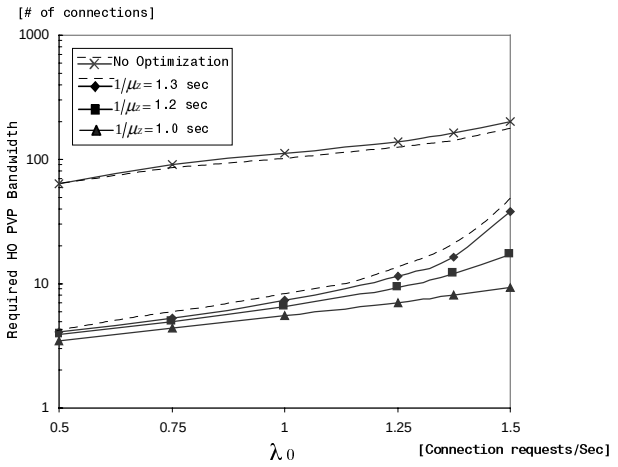

Figure 4 Required HO PVP bandwidth vs. originating call arrival rate

We first study the required HO PVP bandwidth as a function of the originating call arrival rate. Figure 4 shows the required HO PVP bandwidth for different mean route optimization times with no optimization. The figure illustrates the tradeoff that exists between the HO PVP bandwidth and optimization rate. In heavy load region $(\lambda_0 > 1.25)$, the HO PVP bandwidth increases considerably as the optimization rate decreases. While in light load region $(\lambda_0 < 1.25)$, increasing the optimization rate results only in marginal reduction in the required bandwidth. We can also observe a significant bandwidth saving as a result of executing the route optimization procedure. We conclude that the route optimization procedure is a desirable procedure not only for optimizing routes, but also for significantly saving HO PVP bandwidth.

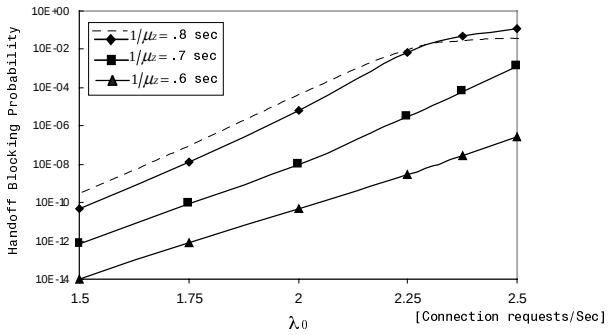

Figure 5 Handoff blocking probability vs. originating call arrival rate

We next study the handoff blocking probability for different mean route optimization times and different range of the originating call arrival rate, as depicted in Figure 5. In this case we assume the maximum number of connections that the HO PVP can hold is 115. The figure illustrates the relation between the handoff blocking probability and the optimization service rate. Since the optimization releases connections within the HO PVP, it results in decreasing handoff blocking probability. The faster the optimization rate is, the smaller the blocking probability becomes.

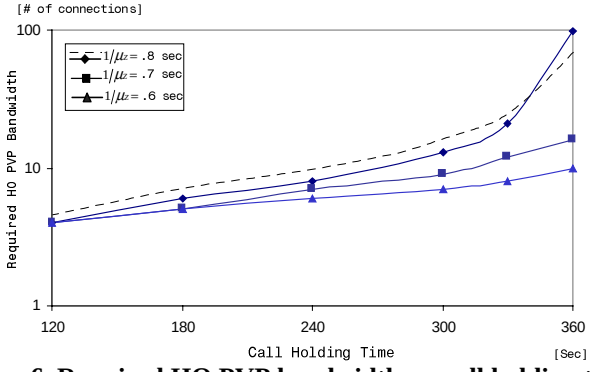

Figure 6 Required HO PVP bandwidth vs. call holding time

The relation between the required HO PVP bandwidth and call holding time when $\lambda_0 = 1.25$ is shown in Figure 6. We vary the mean call holding time from 120 to 360 seconds. The mean cell residual time is chosen to be 360 seconds. The figure shows the required bandwidth for different values of $1/\mu_z$. When the holding time is greater than 300 seconds, varying optimization rate has small impact on the required bandwidth. It is to be noted that the residual time is directly proportional to the handoff arrival rate. The longer the holding time of a call is, the chance of the call to be handed off among radio cells will be bigger. In addition, it is noted from the figure that between 300-360 seconds the handoff rate will be greater as the call holding time becomes close in value to the assumed cell residual time of 360 seconds. This explains why at larger call holding times more HO PVP bandwidth is required.

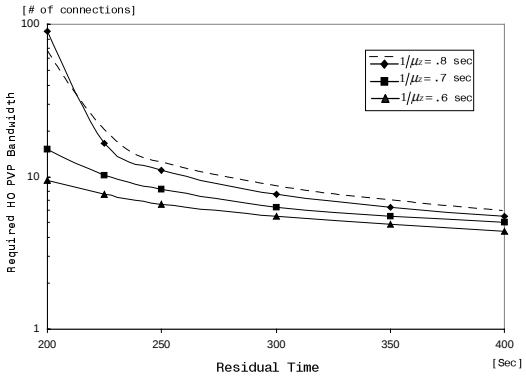

Figure 7 Required HO PVP bandwidth vs. residual time

The relation between the required HO PVP bandwidth and cell residual time when $\lambda_0 = 1.25$ is shown in Figure 7. We vary the residual time from 200 to 400 seconds. The mean call holding time in this case is 200 seconds. It is to be noted that the residual time is inversely proportional to the handoff arrival rate. The smaller the residual time is, the higher chance of the MT to move out of the cell will occur, i.e. the higher the handoff arrival rate. This explains why at smaller residual times more HO PVP bandwidth is required .

From the figures above, it is apparent that the analysis and simulation results are in good agreement. The curves tend to take the same shape with a small marginal error.

7. CONCLUSION

We have presented and studied the performance of a two-phase handoff scheme. The performance was studied using both analysis and simulation. We considered a number of system offer load parameters. These parameters included mean originating call arrival rate, call holding time, and residual time. We examined the relation between required HO PVP bandwidth and

optimization rate. Also we calculated and studied the handoff blocking probability due to lack of HO PVP bandwidth. Results obtained by analysis and simulations were in good agreement. Results indicate a tradeoff exists between required bandwidth and optimization rate. It was shown that the route optimization procedure is a desirable procedure not only for optimizing routes, but also for significantly saving network bandwidth resources utilized to facilitate the rapid rerouting in the first phase.

REFERENCES

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

[2] P. Agrawal, et. al., “SWAN: A Mobile Multimedia Wireless Network,” IEEE Personal Communications Magazine, Vol. 3, No. 2, Apr. 1996, pp. 18-23.

[3] M. Veeraraghavan, et. al., “Handoff Scheme for Mobile ATM Networks,” ATM Forum/96-1499/WATM, 1996.

[4] S. Lee and D. Sung, "A New Fast Handoff Management Scheme in ATM-based Wireless Mobile Networks", In Proceedings of IEEE GLOBECOM, 1996, pp. 1136-1140.

[5] A. Massarella, “Wireless Mobile Terminal/Network Anchor Switch Handover Model,” ATM Forum/97-0265/WATM, 1997.

[6] P. Shieh, et. al., “Handover Schemes to Support Mobility in Wireless ATM,” ATM Forum/96-1622/WATM, 1996.

[7] A. Acharaya, et. al., “Signaling for Connection Rerouting for Handoff Control in Wireless ATM,” ATM Forum/97-0338/WATM, 1997.

[8] C. Toh, "Performance Evaluation of Crossover Switch Discovery Algorithms for Wireless ATM LANs", In Proceedings of the IEEE IC3N, Mar. 1996, pp. 1380-1387.

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