Periodic Route Optimization for Handed-off Connections in Wireless ATM Networks

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

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

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

Abstract

In Wireless ATM networks, user connections need to be rerouted during handoff as mobile users move among base stations. The rerouting of connections must be done quickly with minimal disruption to traffic. In addition, the resulting routes must be optimal. A reasonable approach is to implement handoff in two phases. In the first phase connections are rapidly rerouted and in the second phase a periodic route optimization procedure is executed. The route optimization should impose minimal signaling and processing load on the ATM switches. In this paper, we propose and study a periodic execution of route optimization for a two-phase handoff scheme. We study two types of execution: non-adaptive and adaptive. For the adaptive optimization, we consider two adaptation schemes that are dependent on network conditions. A simulation model is developed to study system performance. The adaptive route optimization is shown to minimize signaling and processing load while maximizing utilization of reserved resources.

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.

We present an alternative two-phase handoff scheme in which Handoff Permanent Virtual Paths (HO PVPs) are provisioned between every two adjacent (Mobility Enhanced Switches) 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 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.

In this paper, we focus on the route optimization of the second phase. In order to minimize signaling and processing load in WATM networks, the route optimization is periodically executed. In particular, we propose and study two types of route optimization execution: non-adaptive and adaptive. The adaptive optimization execution is dependent on network conditions. Two adaptation schemes are considered: range-based and continuous. A simulation model is developed to study system performance.

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. Section 4 presents a study of the non-adaptive optimization execution using simulation. In section 5, we propose and study the adaptive optimization execution, considering two adaptation schemes. Finally, section 6 contains the conclusion.

2. A Two-Phase Handoff

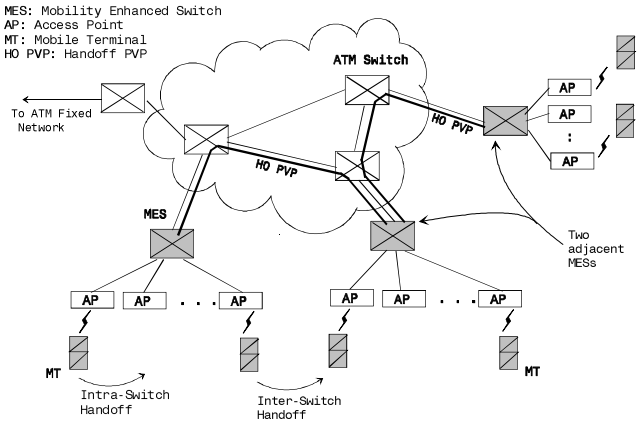


Figure 1. WATM network architecture

In this section, we briefly describe the two-phase handoff scheme proposed. 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 the 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.

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 briefly 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 in order to find the optimal path. The COS discovery is similar to Prior Path Knowledge COS discovery scheme proposed in [8], however no centralized connection server is used. The new MES computes the shortest path from itself and all candidate crossover nodes in the list. The new MES then builds a new connection segment between itself and the candidate crossover node constituting the shortest path. 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. Therefore, it is highly desirable to execute the route optimization periodically and at a suitable rate that would minimize message signaling and processing load.

4. Non-Adaptive Optimization

In this section, we study the performance of the non-adaptive optimization in which optimization periodic rate is constant and doesn't adapt to network conditions. In particular, we study the relation between the optimization rate and the reserved HO PVP bandwidth utilization. The bandwidth utilization is the percentage of the total bandwidth currently being used by connections/calls, i.e. the current number of connections within the HO PVP. 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.

According to the two-phase scheme, the bandwidth for a single HO PVP is affected by two factors: 1) allocation of connections due to inter-switch handoff arrival of λ_S, and 2) release of connections due to one of the followings:

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

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:

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

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. The rate follows a Poisson process.
- $1/\mu_M$: The mean of holding time of a call. Holding time is exponentially distributed.
- $1/\mu_R$: The mean residual or sojourn time of a call in a cell. Residual time is exponentially distributed.

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 $\lambda_S = \lambda_h$.

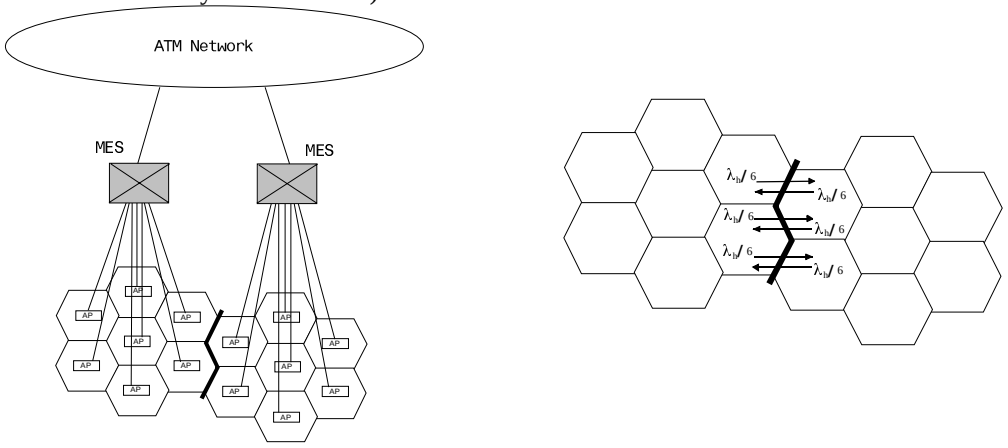


Figure 2. Inter-switch cell boundaries and handoff rates

Next we model and approximate the optimization procedure. 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, μz is distributed between these two adjacent MESs. We assume that μz is divided evenly between the adjacent MESs, with each MES having a mean optimization service rate of $\mu z/2$. As for λs, 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 $\rho = \lambda s/\mu z$.

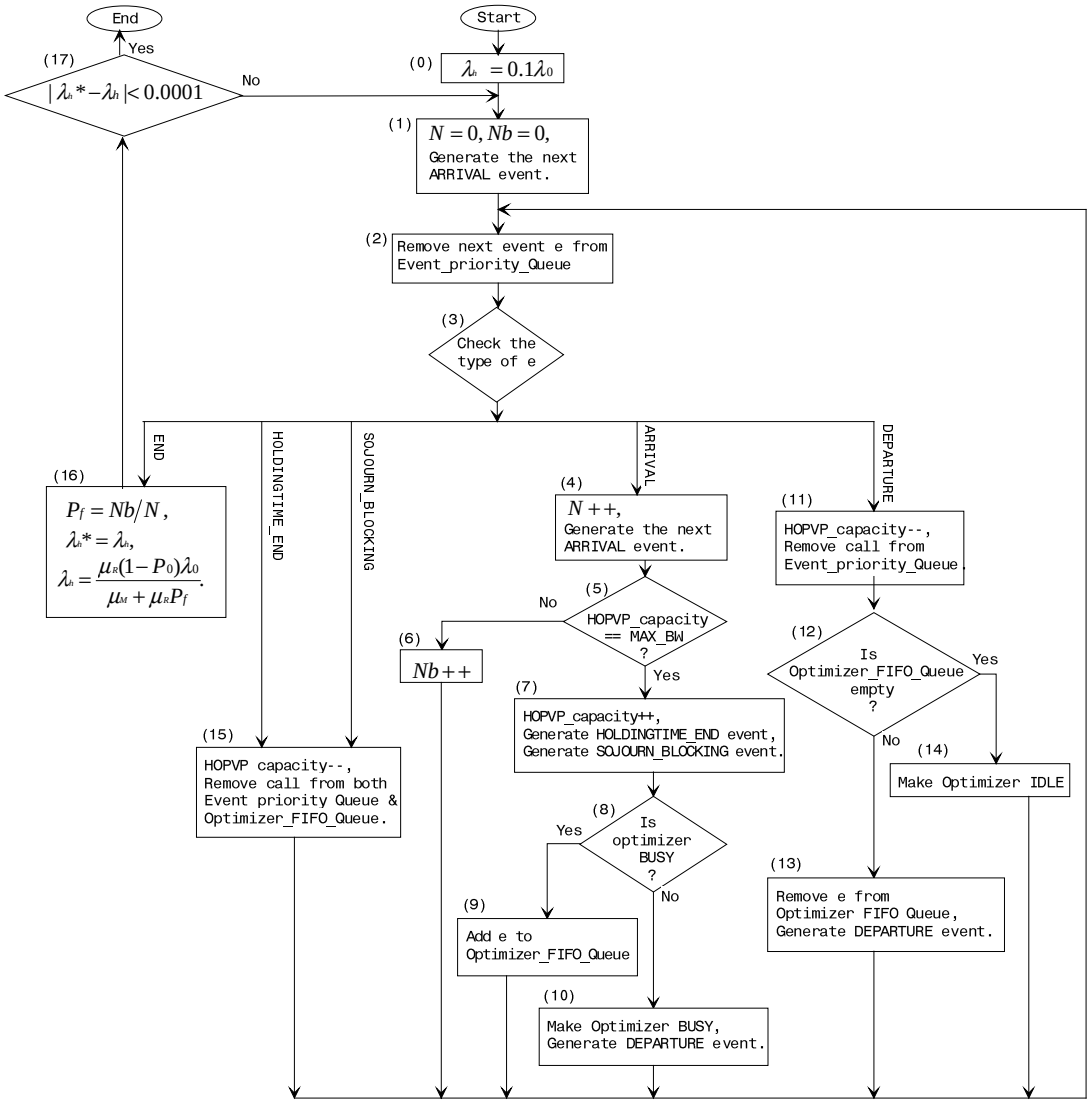


Figure 3. The simulation flow chart

A simulation flowchart is presented in Figure 3 to study the performance of the non-adaptive optimization rate. 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 an initial value for λ_h (e.g. $0.1\lambda_0$), 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.

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.

Next, we study the performance of the simulated non-adaptive optimization. In particular, we examine bandwidth utilization of a single HO PVP for different non-adaptive route optimization rates. 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. The HO PVP bandwidth is assumed to carry a maximum of 200 connections. Mean route optimization times of $1/\mu z$ are chosen to be 2.25, 2.00, and 1.75 seconds. We assume these times are sufficient to carry out the signaling and processing load involved in the optimization procedure. Each simulation iteration was run for 200 hours, i.e. the END event time was 200 hours.

Figure 4 shows the HO PVP bandwidth utilization as a function of the originating call rate λ_0. The utilization is plotted for different values of the mean route optimization time and when the route optimization process is turned off. As expected the utilization with optimization will always have to be less than the utilization with no optimization. The figure also illustrates the tradeoff that exists between HO PVP bandwidth utilization and optimization rate. In

the heavy load region $(\lambda_0 > 1)$, the utilization increases noticeably as the optimization rate decreases. While in the light load region $(\lambda_0 < 1)$, increasing the optimization rate results only in marginal reduction in utilization. It is also noted that the utilization is poor in the light load region.

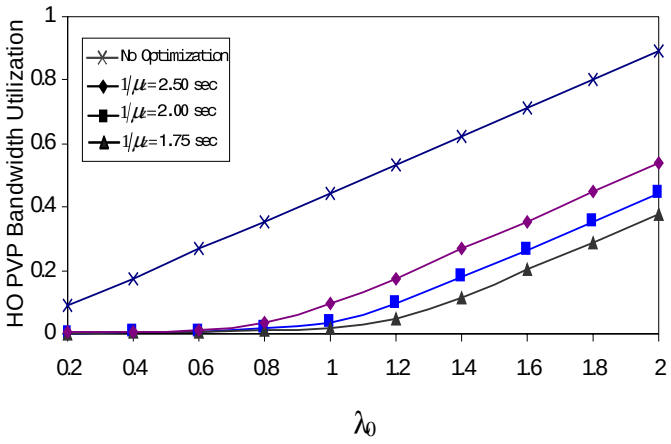


Figure 4. Non-adaptive optimization rates

5. Adaptive Optimization

In order to maximize utilization of the reserved bandwidth and minimize the signaling and processing load at the ATM switches due to the non-adaptive optimization rate, it would be more appropriate to have the optimization rate adapt to changes in network conditions. In real life the optimization rate is dependent on several network parameters: optimality of the current path, reserved bandwidth utilization of HO PVP, handoff blocking probability, connection QoS, connection lifetime (being old or new), number of hops, loop detection, etc. However, it would be difficult to model such a system based on all of these parameters. Also from implementation point of view, additional processing load and computation complexity would be imposed on the MES as it needs to monitor and compute all of these parameters. A simple parameter would be the reserved bandwidth utilization.

To have the optimization rate adapt to changes in the bandwidth utilization, we must allow the optimization rate to decrease when the utilization decreases. Also, when the utilization increases, the optimization rate should increase. Hence, the route optimization service time for any handoff can be expressed as $1/\mu_n = \tau + 1/\mu$. τ represents the route optimization urgency (i.e. how aggressive the optimization rate should be) and is dependent on the current bandwidth utilization.

Next we propose and study two adaptation schemes. One simple scheme is to employ a range-based service rate scheme in which a utilization range corresponds to a particular optimization rate. Another scheme is to have the relation between the optimization rate and bandwidth utilization expressed as a continuous function.

5.1 Range-based Adaptation

We shall refer to this as Scheme A. To implement this scheme in simulation, we employ 5 classes for service rate with each class having a different urgency value. The urgency value covers 20% of the current utilization of the HO PVP. The higher the utilization is, the smaller the urgency would be. The classes, utilization ranges, and corresponding urgencies are illustrated in Table 1.

Table 1. Utilization range and corresponding urgency

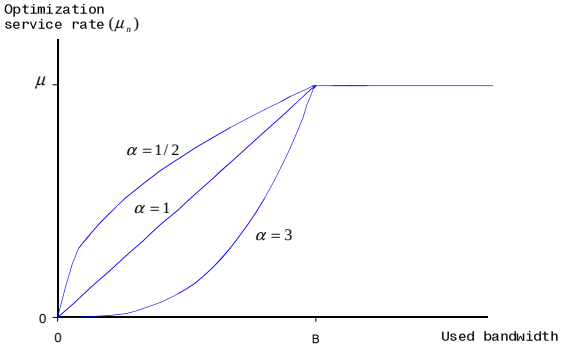
	Utilization Range	Urgency (sec)
Class A	0.8 – 1.0	0
Class B	0.6 – 0.8	1*T*
Class C	0.4 – 0.6	2*T*
Class D	0.2 – 0.4	3*T*
Class E	0.0 – 0.2	4*T*

T=5 sec.

5.2 Continuous Adaptation

We shall refer to this as Scheme B. For continuous adaptation, we have the relation between optimization rate μ_n and current utilization expressed according to following continuous function:

$$\mu_n = \begin{cases} \left(\frac{n}{B}\right)^{\alpha}\mu & (0 \le n < B) \\ \mu & (n \ge B) \end{cases}$$



where μ is the maximum optimization rate, n is the current used bandwidth and B is a threshold of the reserved bandwidth and is equal to a fraction (e.g. 80%) of the maximum reserved bandwidth. Note that $\left(\frac{n}{B}\right)$ represents the utilization of the HO PVP. The parameter α is a tunable parameter and controls how fast the optimization rate should change in relation to the bandwidth. For example if α =1, the rate is proportional to the bandwidth. Note that when α <1, the service rate becomes more aggressive, i.e. more responsive to changes in the utilization. When α >1, the service rate is less aggressive.

To implement the adaptive route optimization, we will have $\tau = [(\frac{B}{n})^{\alpha} - 1] \cdot T$. If the fraction $[(\frac{B}{n})^{\alpha} - 1]$ is negative, then τ will be set to zero, and hence the optimization is executed at a mean rate of $1/\mu$. The simulation model for both adaptive schemes stays the same as in Figure 3. However, the calculation of the DEPARTURE event would be different. The DEPARTURE event time would be the simulation clock value plus the urgency τ plus the random value generated from the exponential distribution with mean $1/\mu$.

Next, we study the performance of the simulated adaptive optimization and compare the results to that of non-adaptive. We choose 5 seconds for *T*, 2 for α and 80% of maximum bandwidth for *B*. We also assume 2 seconds for $1/\mu$.

Figure 5 plots the bandwidth utilization for four different types of optimization: no optimization, non-adaptive scheme, adaptive scheme A, and adaptive scheme B. The figure shows that we achieved better utilization of the HO PVP bandwidth when utilizing the adaptive optimization schemes. Scheme B yields better utilization when considering the utilization for the lightly loaded region when λ_0 is between .2 to 1.4. Scheme B is comparable to scheme A in the heavy loaded region when λ_0 is greater than 1.4. We next look at the corresponding average adaptive optimization rate as a function of λ_0. Remember that the optimization rate is an indicator of the signaling and processing load imposed on the WATM network node elements as a result of initiating route optimization.

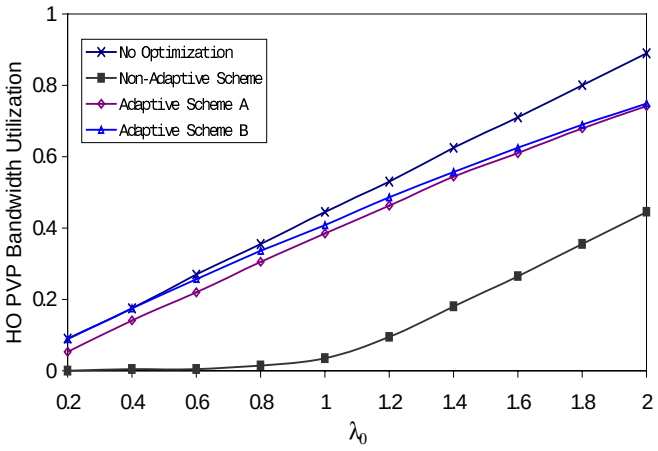


Figure 5. Non-adaptive vs. adaptive optimization rates

Scheme A and scheme B average adaptive optimization rates against the originating call arrival rate are plotted Figure 6. The figure illustrates how the optimization rates change as a function of the system load parameter λ_0. From the figure, it is apparent that a significant reduction in processing load and signaling overhead, especially in the light load region, is made. Rather than executing optimization always at a mean rate of .5, the optimization is now executed at a much slower rate that adapts to system load.

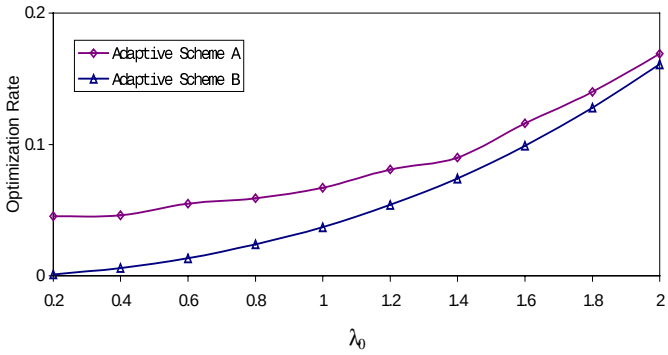


Figure 6. Range-based and continuous adaptive optimizations

It is noted that scheme B is always adapting to changes in utilization, and hence resulting in better utilization and less signaling and processing load than that of scheme A. One can even engineer a more desirable optimization rate and bandwidth utilization for scheme B by tuning the parameters T and α. As for scheme A, it is noted that the optimization rate is not smooth. This is so because the urgency changes only for 5 different utilization ranges. Within the 20% utilization range, urgency stays the same.

This is not true for scheme B where urgency always depends on the current value of the utilization. In order to achieve better adaptation for scheme A, additional classes for utilization and suitable delays (i.e. *T* values) are to be utilized.

6. Conclusion

We have proposed and studied a periodic execution of route optimization in WATM networks for a two-phase handoff scheme. Two types of route optimization were considered: non-adaptive and adaptive. The adaptive optimization was dependent on the current resource utilization. It was shown that the adaptive route optimization minimizes overhead and signaling messages while maximizing reserved resource utilization. For the adaptive optimization, we proposed and studied two adaptation schemes: range-based and continuous. Design and implementation issues were discussed. Continuous adaptation was shown to be more appropriate in terms of maximizing bandwidth utilization and minimizing processing overhead.

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.