Providing Seamless Communications in Mobile Wireless Networks *

Bikram S Bakshi P. Krishna N. H. Vaidya D. K. Pradhan
Department of Computer Science
Texas A&M University
College Station, TX 77843-3112
E-mail: {bbakshi,pkrishna,vaidya,pradhan}@cs.tamu.edu
Phone: (409) 862-2599
April, 1995

Technical Report # 95-046

Abstract

This paper presents a technique to provide seamless communications in mobile wireless networks. The goal of seamless communication is to provide disruption free service to a mobile user. A disruption in service could occur due to active handoffs (handoffs during an active connection). Existing protocols either provide total guarantee for disruption free service incurring heavy network bandwidth usage (multicast based approach), or do not provide any guarantee for disruption free service (forwarding approach). There are many user applications that do not require a "total" guarantee for disruption free service but would also not tolerate very frequent disruptions. This paper proposes a novel staggered multicast approach which provides a probabilistic guarantee for disruption free service. The main advantage of the staggered multicast approach is that it exploits the performance guarantees provided by the multicast approach and also provides the much required savings in the static network bandwidth.

The problem of guaranteeing disruption free service to mobile users becomes more acute when the static backbone network does not use any packet numbering or does not provide retransmissions. Asynchronous Transfer Mode networks, the future of B-ISDN, display these properties. To make our study complete, we present a possible implementation of our scheme for wireless ATM networks.

*Initial Version of this paper was submitted to Mobile Computing and Networking - MCN'95, in April, 1995. Research reported is supported in part by AFOSR under grant F49620-94-1-0276, and Texas Advanced Technology Program under grant 999903-029.

1 Introduction

Mobility has opened up new vistas of research in networking. With the availability of wireless interface cards, mobile users are no longer required to remain confined within a static network premises to get network access. Users of portable computers would like to carry their laptops with them whenever they move from one place to another and yet maintain transparent network access through the wireless link. Integrated voice, data and image applications are going to be used by millions of people often moving in very heavy urban traffic conditions.

On the downside, mobility brings along with it a myriad of network management problems. The problems could be broadly classified as mobility management related and connection management related. In this paper we will primarily deal with a key problem in mobile wireless networks related to connection management. The problem deals with providing disruption free service to mobile users.

Future personal communication networks (PCN) will allow users to engage in bi-directional exchange of information including but not limited to voice, data, and image, irrespective of location and time, while permitting users to be mobile. Even though, near term personal communication services (PCS) are going to be voice-oriented, PCN are expected to support multimedia PCS in the long term [13]. This will spur requirements for high capacity wireless networks.

A typical PCN with mobile users [8, 9, 10] comprises of a static network and communication links between them. Some of the fixed hosts, called *base stations* (*BS*)[1] are augmented with a wireless interface and they provide a gateway for communication between the wireless and static network. Due to the limited range of wireless transreceivers, a mobile user can communicate with a *BS* only within a limited geographical region around it. This region is referred to as a base station's *cell*. A mobile user communicates with one *BS* at any given time. Each *BS* is responsible for forwarding data between the mobile user and the static network.

When a mobile host is engaged in a call or data transfer, it will frequently move out of the coverage area of the mobile support station it is communicating with, and unless the call is passed on to another cell, it will be lost. Thus, the task of forwarding data between the static network and the mobile user must be transferred to the new cell's *mobile support station*. This process, known as *handoff*, is transparent to the mobile user. Handoff helps to maintain an end-to-end connectivity in the dynamically reconfigured network topology.

As the demand for services increase, the number of cells may become insufficient to provide the required quality of service. *Cell splitting* can then be used to increase the traffic handled in an area without increasing the bandwidth of the system. In future, the cells are expected to be very small (less than 50 meters in diameter) covering the interior of a building. The reduction in the cell

[1] Base stations are sometimes called *mobile support stations*.

size causes an increase in the number of handoffs, thereby increasing the signalling traffic (network load) due to the handoff protocol messages. In addition, handoff also causes a disruption in service if it is not done in a fast and efficient manner. In this paper we will primarily deal with design and implementation issues of handoff protocols to ensure disruption free service.

Providing connection-oriented services[14, 15, 16, 17, 18] to the mobile users requires that the user always be connected to the rest of the network in such a manner that its movements are transparent to the users. Providing disruption free service is a stronger requirement than mere connection-oriented services. In addition to maintaining the connection, the network will need to ensure that the delay experienced by the data packets over the network is less than a fixed time called the deadline. The deadline is in turn determined by the *quality of service* (QOS) required by the users. The goal of seamless communication is to provide disruption free service to a mobile user. A disruption in service could occur due to active handoffs (handoffs during an active connection). This is because traditional protocols require the old BS to forward data packets to the new BS. Thus, every time a mobile user moves into a new cell during the connection (active handoff), the user will see a break in service while the data gets forwarded to it from the old BS via the new BS.

We first present the proposed approach for providing disruption free service to mobile users. Our work differs from existing protocols in that the network load incurred by the proposed approach is significantly lower as compared to others. The number of disruptions seen by the user will depend on the number of handoffs incurred during the lifetime of the connection. The number of handoffs in turn depends on the mobility pattern of the user. In this paper we use two mobility models to analyze the proposed approach. In the first model, the user spends very little time in a cell (handoffs occur frequently), while in the other model the user spends a long time in a cell (handoffs occur infrequently). Analysis shows that for both these models, the proposed approach significantly reduces the network bandwidth usage without violating the quality of service (QOS) requirements specified by the user application.

The problem of guaranteeing disruption free service to mobile users becomes more acute when the static backbone network does not use any packet numbering or does not provide retransmissions. Asynchronous Transfer Mode networks, the future of B-ISDN, display these properties. The second half of the paper deals with implementation issues of the proposed approach. The backbone network has been assumed to be an asynchronous transfer mode (ATM) network. The vast transmission capacity offered by an ATM broadband network can provide communication services to a wide range of applications including video and audio. It is thus a natural choice for multimedia services. ATM is basically a connection-oriented switching technology. Users need to establish a fixed route called a *virtual channel (VC)* before any information can be exchanged. To make maximum use of available bandwidth, multiple *VC*s can be statistically multiplexed over the same link. Issues related to ATM have been comprehensively treated in [19, 22, 23].

While ATM promises to do away with the present problems faced by the telephony community, it raises a number of issues for the mobile computing industry. As mentioned before, existing ATM protocols do not offer any packet numbering and prohibit packet reordering. In this scenario maintaining a continuous (disruption free) communication link to the mobile host becomes complicated. We thus need to ensure that once a handoff takes place no packet is lost and deadlines are met, i.e., the packets that have been transmitted to the previous BS and which have not reached the mobile host due to handoff, are somehow delivered to it within the given time constraint. Keeping these problems in mind, we propose an easily implementable technique to provide disruption free service to mobile hosts in wireless ATM networks.

The rest of this paper is organized as follows. In section 2 we briefly review related work. The basic idea behind our scheme is presented in section 3. Section 4 presents the issues related to implementation of the proposed approach using ATM as the backbone network. Concluding remarks are presented in section 5.

2 Related Literature

Keeton et.al. in [2] proposed a set of algorithms to provide connection oriented network services to mobile hosts for real time applications like multimedia. Their solutions lay excellent groundwork for work in this area but did not guarantee disruption free service. In fact their scheme was shown to suffer from extended intervals of time when service to the mobile host was disrupted. A study done in [1] shows that if the handoff protocol required forwarding data between the BSs connected by physical links, then a high bandwidth (between 48Mbps and 96Mbps) is required just to forward these data packets. Moreover, loops can be formed in the connection path if forwarding is employed. This will lead to inefficient network utilization.

A multicast based solution was proposed in [1]. In this approach, the data packets for a mobile host are multicast to the BSs of the neighboring cells so that when the host moves to a new cell, there are data packets already waiting for it and thus, there is no break in service. It is evident, however, that this scheme is not cost effective. As the number of users in the network increases, the amount of network bandwidth used up by the multicast connections is going be prohibitively high. In [1], the cost of such a multicast scheme was determined to be the buffer overhead at the BSs. Our view of the problem is that the major component of cost incurred in a multicast based approach will be the *amount of extra bandwidth used*, and not the buffer overhead at each BS. This argument is supported by the availability of cheap memory but expensive network bandwidth[2].

As pointed out in [3], the *network call processor*[3] in a static network becomes the bottleneck

[2]The cost of a 30 minute call from USA to Japan is approximately equal to the cost of 1 Mbyte of RAM.

[3]The role of the *network call processor* is to establish a path or route at connection setup time. While doing so it takes into account the network load so as to balance the load on each network node.

in an environment where handoffs are frequent and require excessive interaction with a base station – an inherent problem associated with the work in [2]. To alleviate this problem, the authors in [3] proposed a new network architecture which made use of *virtual circuit trees* to minimize handoff processing. However, it does not discuss about providing disruption free service when a handoff takes place – the mainstay of applications like multimedia [23].

We find that while existing literature is a rich source of protocols and models for tackling the problem in hand, there does not exist a cost-effective solution for providing disruption free service.

3 Proposed Approach

Traditional multicast-based schemes require the packets to be multicast throughout the length of the connection. This leads to wastage of network bandwidth. The communication links from the switch to the BSs other than the BS of the cell where the mobile host is currently located get unnecessarily loaded. As the number of mobile hosts increase in a cell, the total network usage due to multicast connection for each host will become enormous. Due to this extra network usage, new connections might be blocked because the network capacity is exceeded.

The thrust of our approach is to avoid unnecessary multicast. A multicast throughout the length of the connection may prove to be unnecessary if the network had some information – e.g., how long is the mobile host going to remain in the same cell (this period is called *cell latency*). If the network has such information, then the multicast need not be done during that period of time.

The main idea of the proposed approach is to "stagger" the multicast initiation by the amount of time one is sure that the host remains within a cell, i.e., for a time interval equal to the *cell latency*. The *cell latency* will solely depend on the mobility model of the host. In this paper we will analyze the proposed approach based on two mobility models. One model is *pessimistic* in nature, and the other *optimistic*. By *pessimistic* we mean that the *cell latency* for a mobile host is very small. On the other hand in the *optimistic* model, the mobile host remains in a cell for a longer time.

We will now present the staggered multicast approach.

3.1 Staggered Multicast

If the mobility pattern of a user could be modeled in such a way that it can be ascertained with a certain probability that the user is going to remain in the same cell for t_s amount of time, multicast could be avoided for this amount of time. The value of t_s then[4] gives us a measure of the stagger

[4]Note that the actual stagger time is less than t_s, as the time to set up the multicast connections should be taken into account. This has been dealt in greater detail in Section 4.3 for a wireless ATM network.

time than can be safely introduced before initiating a multicast. This way, we will save on the network usage, and still guarantee disruption-free service with a certain probability.

Let P_i be the probability of disruption during the *i-th* handoff, and t_i be the *cell latency* before the *i-th* handoff. Let t_{mi} be the time spent in multicast mode before the *i-th* handoff. A disruption occurs when a mobile host initiates a handoff before multicast has been initiated. Then the probability of disruption during the *i-th* handoff can be given as,

$$P_i = Pr[t_s > t_i]$$

Let the number of handoffs occurring over the length of the connection time T_c be N_h. Let $P_{disrupt}$ be the average probability of disruption during a handoff. $P_{disrupt}$ is determined as,

$$P_{disrupt} = \frac{1}{N_h} \sum_{i=1}^{N_h} P_i$$

The value of $P_{disrupt}$ can now be used as a measure of the Quality of Service (QOS). There are a number of applications that cannot tolerate disruptions during the time of connection, i.e., $P_{disrupt} = 0$. Two existing examples of such applications are *telemedicine*, and *video conferencing.* With increased availability of mobile computing applications, a large number of hitherto unexplored applications will emerge. The applications mentioned here are but only a small sample.

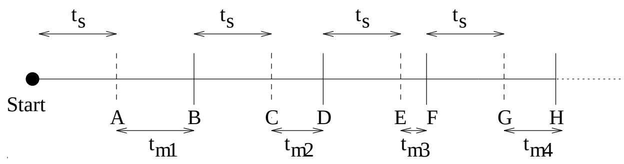


Figure 1: Total Guarantee

Figure 1 presents an example showing the times of handoffs and multicast initiations. The times B, D, F, and H represent the time at which handoff takes place. The times A, C, E, and G represent the time at which multicast is initiated. The cell latencies for Figure 1 are $t_1 = t_s + t_{m1}$, $t_2 = t_s + t_{m2}$, and so on. For total guarantee, the following should hold.

$$\forall i, 1 \leq i \leq N_h, t_s < t_i$$

i.e., for all handoffs a multicast is initiated within the associated *cell latency* interval.

However, there are a lot of applications that do not have a strict requirement of disruption free service during every handoff. A probabilistic guarantee is sufficient for such applications, i.e., $P_{disrupt} \geq 0$. Examples of such applications include *ftp*, *audio channels* and *movies.* If the QOS requirement can be expressed as a probabilistic guarantee for disruption free service, then the

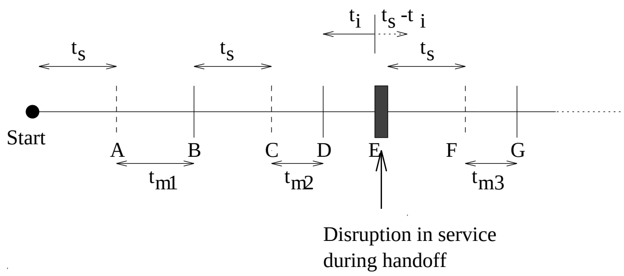


Figure 2: Probabilistic Guarantee

multicast initiation could be further staggered resulting in an even greater reduction of network usage.

To illustrate this probabilistic scheme we present an example as shown in Figure 2. This figure shows the times of handoffs and multicast initiations in the multicast scheme that provides a probabilistic guarantee. The times B, D, E, and G represent the time at which handoff takes place. The times A, C, and F represent the time at which the multicast is initiated. As noticed in the figure, there is a disruption in service during handoff at time E, because, there was no multicast initiated before the handoff. Thus, a disruption occurs during the *i-th* handoff when the stagger time t_s is greater than the cell latency time t_i.

In the absence of any empirical data for user mobility, we propose to evaluate the effectiveness of our scheme using two mobility models, which we believe cover a wide range of user mobility. At this point we would like to mention that the main aim of this paper is not to show that the two models cover the whole spectrum of user mobility, but, to show that with the aid of user mobility information, we can drastically reduce the network load and still provide disruption-free service. We will be able to correctly estimate the benefits obtained from the proposed approach only if we can accurately model the user mobility.

3.2 Mobility Models

3.2.1 *Optimistic* Model

The *optimistic* model is based on the two dimensional random walk model. In such a model, the user tosses two coins every T seconds. Based on the resulting head-tail combination, the user will decide to take a step of size s meters in a specific direction (e.g., head-head results in a step in the north-east direction). Let the distance of the user with respect to the center of the circular cell at time t be $r(t)$. As derived in Appendix 1, the probability that a mobile user will remain in the same cell at time t is given as,

$$Prob(r(t) < R) = 1 - e^{-\frac{R^2}{2\alpha t}} \tag{1}$$

where, R is radius of the cell, and $\alpha = s^2/T$.

Let us consider a picocellular environment, which is more suited for pedestrian traffic. Let s = 0.4m, T = 0.25 sec. Therefore, $\alpha = 0.64$. We vary the radius of the circular cell R from 10m to 50m. The variation of the probability with time is illustrated in Figure 3. As seen in the

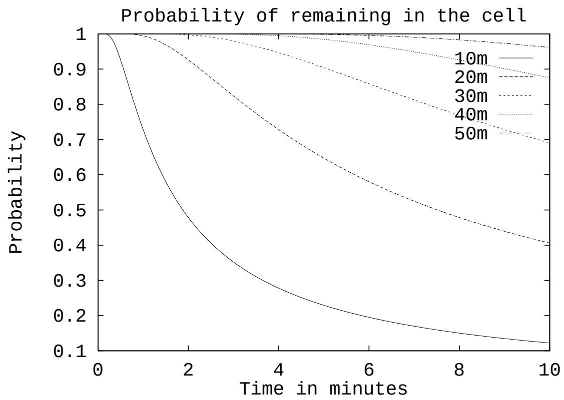


Figure 3: Probability of being in a cell

figure, the probability of the mobile host remaining in a cell decreases with time. An interesting observation however is that even after 5.1 minutes (306 seconds), the probability that the user is still in the same cell ($R = 30m$) is as high as 90%. This mobility model represents the class of users who spend a lot of time in a cell.

3.2.2 *Pessimistic* Model

The *pessimistic* model is based on the mobility model proposed in [5]. In this model the mobile user is assumed to be moving at an average velocity of V. The direction of movement is uniformly distributed over $[0, 2\pi]$. The mobile users are assumed to be uniformly distributed over the cell area with a density of ρ. If the length of the cell boundary is L, and the cell area S, the number of mobile users crossing the cell boundary per unit time is given by $\frac{V\rho L}{\pi}$. If ρ can be assumed to remain constant over the entire cell area, the average cell crossing rate of a mobile user is given by $\frac{VL}{\pi S}$. For circular cells, L will correspond to the perimeter of a cell, and thus $\frac{L}{S} = \frac{2}{R}$. It follows that the average *cell latency* of a mobile user is given $\frac{\pi R}{2V}$. As in [5], we will assume that the *cell latency* of a mobile user is exponentially distributed with a mean $\frac{\pi R}{2V}$.

3.3 Performance Analysis of the Staggered Multicast Approach

The overhead of the staggered multicast scheme can be characterized by the total time T_m spent in the multicast mode as compared to the length of connection T_c. T_m is determined as

$$\sum_{i=1}^{N_h} t_{mi}$$

where, t_{mi} is the time spent in multicast mode before the *i-th* handoff, and N_h is the number of handoffs occurring over the length of connection. The total time spent in the unicast mode, T_u, is then given by the difference, $T_c - T_m$. We determine the overhead of the multicast scheme as the fraction of the total connection time spent in the multicast mode,

$$Overhead = \frac{T_m}{T_c}$$

The QOS measure of the staggered scheme (characterized by $P_{disrupt}$) is now given as

$$QOS = 1 - P_{disrupt}$$

3.3.1 Performance of *Optimistic* Model

In this section we present the results of the staggered multicast scheme obtained using the *optimistic* model. We performed simulations to analyze the staggered multicast scheme. The radius of the circular cell R was varied from 10m to 50m. The time of connection T_c, was fixed to be 100 minutes. The step size s was chosen to be 0.4m, and the time interval between two tosses T was chosen to be 0.25 s.

As stated earlier, we characterize the overhead as T_m/T_c. Figure 4 illustrates the variation of overhead with the stagger time t_s. It is noticed in Figure 4 that the overhead reduces as the stagger time increases. This is because as stagger time increases, the amount of time spent in multicast mode reduces. Thus, the overhead, determined as T_m/T_c, reduces. It can also be noticed that for a given stagger time, the overhead increases with an increase in cell radius. This is because as the radius increases, the time interval between handoffs increases. If stagger time is kept constant, we are not making use of the potentially extra time available due to increased cell radius. As a result the fraction of time spent in multicast mode increases.

We also evaluated the probability of a disruption during a handoff, $P_{disrupt}$. As stated earlier, a disruption occurs only if multicast is not initiated before a handoff occurs. Figure 5 illustrates the variation of probability of disruption with stagger time t_s. Higher the stagger time, higher is the probability of disruptions. It can also be noticed that for a given stagger time, the probability of disruption increases as the radius of the cell decreases. This is because as the radius decreases, the probability of remaining in a cell reduces for a given stagger time. Therefore, the probability of disruption increases.

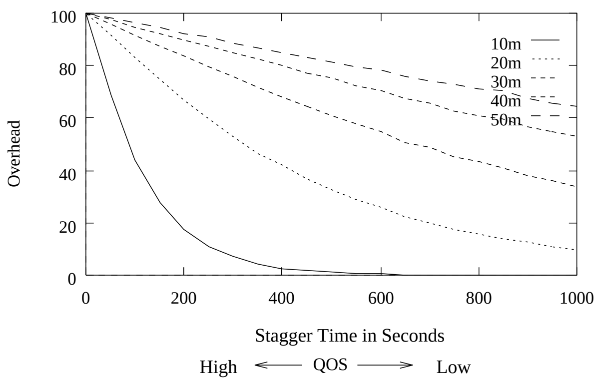


Figure 4: *Optimistic* Model : Overhead Vs Stagger Time

Using these results, the network can determine the appropriate stagger time for a user. Let us illustrate it with an example. Let the radius of the cells in the network be 30 m. Suppose that the users in a network are maintaining non-critical connections. This means that a probabilistic guarantee will suffice. Let the QOS demanded by the users be 75 %. This means that $P_{disrupt}$ = 25%, i.e., on an average three out of four handoffs will be guaranteed to be disruption free. Then, using Figure 5, we can determine the appropriate stagger time, which is 650 seconds (approx. 11 minutes). Therefore, the multicast initiation can stagger by 11 minutes and we will still provide the desired QOS to the users. The overhead of such a staggered scheme can be determined using Figure 4 to be 50%. Therefore, the network spends only 50% of the total connection time in multicast mode for the user. In a traditional multicast based solution for disruption free service, the network spends 100% of the connection time in multicast mode [1]. Comparing it to the traditional multicast based solutions, there is a 50% savings in network bandwidth.

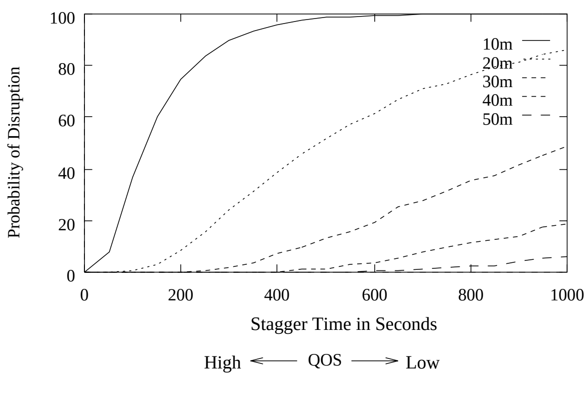


Figure 5: *Optimistic* Model : Probability of Disruption Vs Stagger Time

Suppose on the other hand, a user is maintaining a critical connection which demands total guarantee of disruption free service, i.e., the QOS demanded by the user is 100 %. Even though a non-zero value of stagger time could be obtained for $P_{disrupt} = 0$ in the *optimistic* model (e.g., $t_s = 3$ minutes for $R = 30m$ in Figure 5), this may not be true in general for other models. In fact the next model shows that for total guarantee of disruption free service, stagger time has to be zero. In other words, for total guarantee of disruption free service, multicast should be done throughout the length of the connection.

3.3.2 Performance of *Pessimistic* Model

Simulations were performed to analyze the multicast scheme using the *pessimistic* mobility model. The mobility model for this part was same as the mobility model proposed in [5]. The radius of the circular cell R was varied from 10m to 50m. The time of connection T_c, was fixed to be 100 minutes. The average velocity V was chosen to be 1.6 m/s (approx 5.7 km/hr, for a pedestrian user).

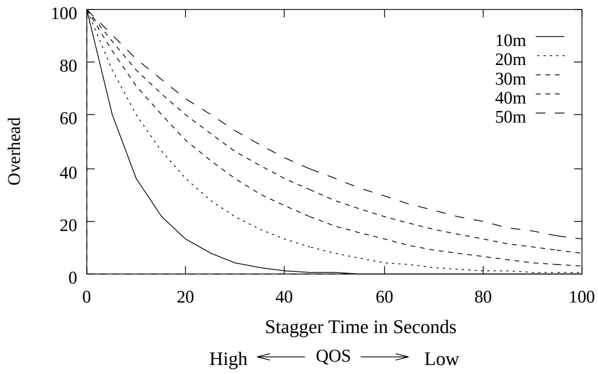


Figure 6: *Pessimistic* Model : Overhead Vs Stagger Time

The trends in the variation of overhead (Figure 6) and probability disruption (Figure 7) with respect to stagger time for the pessimistic mobility model are similar to the optimistic model. But as was expected, the allowable stagger time in the pessimistic model for a particular QOS is very low compared to the allowable stagger time in the optimistic model. For example, when R = 30m, QOS = 75 %, the allowable stagger time in the optimistic model is 650 seconds. On the other hand for a pessimistic model, the allowable stagger time is only 9 seconds.

Another noticeable difference with the optimistic model is that there is no stagger time allowable for total guarantee service (i.e., when $P_{disrupt} = 0$). Thus, for a user whose mobility pattern can be modeled with the pessimistic model, multicast has to be done throughout the connection time if the user desires total guarantee of disruption free service. On the other hand, if the user requires only a probabilistic guarantee, then a non-zero stagger can be introduced. For

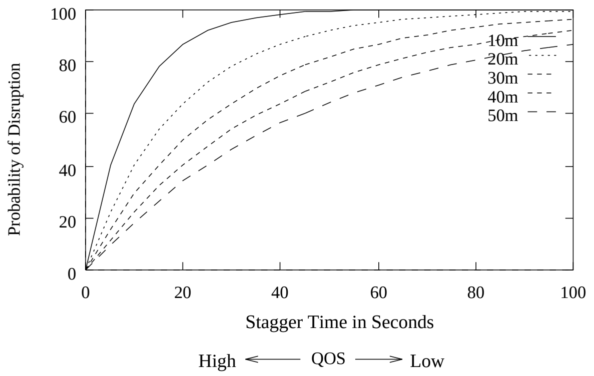


Figure 7: *Pessimistic* Model : Probability of Disruption Vs Stagger Time

example, if the QOS demanded by the user is 75 %. Then, using Figure 7, we can determine the allowable stagger time to be 9 seconds. The overhead of such a staggered scheme can be determined using Figure 6 to be 73 %. Therefore, when compared to the traditional multicast schemes, there is a 27% savings in network bandwidth.

3.3.3 Discussion

In this section we have presented a staggered multicast approach. The main features of this approach are that it saves network bandwidth by providing a probabilistic guarantee for disruption free service. We analyzed the proposed approach for two mobility models. These models represented two different classes of mobile users – those with high *cell latency*, and those with low *cell latency*. The results indicate that regardless of the mobility model, the proposed approach provides tremendous savings in network bandwidth for applications that require a probabilistic guarantee. We expect the performance gains of the proposed approach for a typical user mobility model to lie somewhere in between the performance gains obtained for the two models considered.

In the next section we will present an implementation of the proposed approach on a wireless ATM network.

4 Implementation on ATM Network

4.1 System Model

We view the future personal communication network as a two tier network - a backbone static ATM network and a peripheral wireless network. This model is similar to the one proposed in [4]. Figure 8 shows ATM switches connected to base stations which in turn provide service to the mobile hosts. ATM cells are received by the base stations from the static network and forwarded to the mobile hosts.

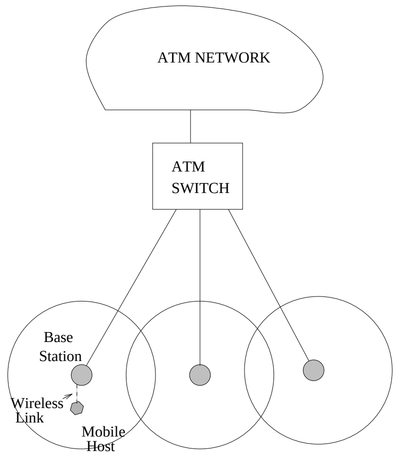


Figure 8: PCN Model

4.2 Protocol

We define a multicast group g_i as the set of base stations that are included in the multicast operation for the mobile host i. The base stations maintain a table which maps each mobile host in its cell to its multicast group members. The group members for a mobile host can be determined based on some hints (direction, velocity). If no hints are available, the default multicast group members will be the neighboring base stations [1].

The connection management problem can be divided into two phases, namely, connection establishment phase and connection maintenance phase. The source mobile host initiates the connection establishment phase by sending a *connection request* message to its base station. The base station forwards this message to its switch. The switch assigns a VCN (virtual circuit number) for the source mobile host. The switch then initiates a *locate* procedure for the destination mobile host [10, 11, 12]. Upon getting the location information of the destination mobile host, a connection is set up between the source and the destination mobile host via the switches at the source and the destination.

Our work differs in the connection maintenance phase. Please refer Figure 9 for the discussion. The thick lines in Figure 9 represent the data packets being transferred over the static network, and the thick dashed lines represent the data packets being transferred over the wireless medium between the base station and the mobile host. The thin lines represent the control messages being transferred over the static network, and the thin dashed lines represent the control messages being transferred over the wireless medium.

Once a connection is established, the switch SW is in the unicast mode, i.e., it forwards the data packets to only the "current" base station $BS1$, which in turn forwards it to the mobile host

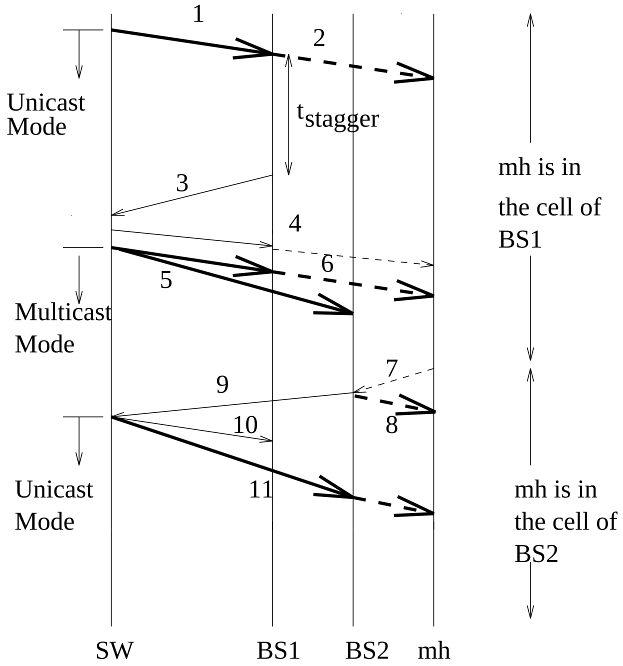


Figure 9: Connection Maintenance

mh (steps 1-2). After $t_{stagger}$ units of time[5], $BS1$ sends a *multicast initiate* message to SW (step 3). The multicast group members g_{mh} are tagged along with the message. The switch SW then determines the crossover point for the multicast group members. The VCNs to the base stations in g_{mh} are assigned, and the switch SW sends back the list of VCNs to $BS1$ which forwards it to mh (step 4). Upon receiving an acknowledgment from mh, SW enters the multicast mode. Let us suppose that the multicast group members are $BS1$ and $BS2$. SW multicasts the data packets to the base stations $BS1$ and $BS2$ (step 5). However, only the current base station which is $BS1$ forwards the data packets over the wireless medium to mh (step 6). This continues till the mobile host mh detects that it has to handoff to $BS2$. The mobile host mh then sends a *handoff initiate* to the new base station $BS2$ (step 7). The base station $BS2$ starts transmitting data to the mobile host (step 8). It also forwards *handoff initiate* message to the switch SW (step 9). The switch SW then terminates the connections to the multicast group members except for $BS2$ (step 10). SW then reenters the unicast mode and sends the data packets to only the "current" base station $BS2$, which in turn forwards it to mh (step 11).

4.3 Implementation Issues

Given the lossy nature of the wireless medium, there may be a need to frame groups of ATM cells at the BS and assign them some kind of sequence numbers. Additional bits to enable error correction

[5]It will be shown later that $t_{stagger} < t_s$, where t_s is the stagger time derived in Section 3.

and to allow recovery schemes may also be required for each frame. Likewise communication from the mobile host to a base station will consist of frames of ATM cells with additional bits as described above. Going by the philosophy behind ATM, it is likely that each frame will be small and of equal size. In line with this, we assume that all frames will be of fixed length containing F ATM cells.

Before we can apply our scheme to an ATM environment, we must take into account the various properties of ATM network protocols that make them differ from existing network protocols. As was mentioned before, ATM is a connection oriented switching technology where connections must be established for the entire duration of the call. Connection establishment consists of assigning a VCN (virtual circuit number) and/or a VPN (virtual path number), and allocation of resources both within the network and at the source and destination to support this connection. In a mobile environment we will thus need to ensure that before a mobile host hands off, connection has already been established between the new base station and the destination. For this purpose, we make use of a *dynamic virtual connection tree (dvct)* based network architecture, an extension to the idea proposed in [3]. For sake of completeness, we will describe the virtual connection tree in some more detail.

A virtual connection tree [3] is a set of cellular ATM switches and base stations in the static network that are chosen at call setup time to route ATM cells. The network is divided into *neighboring access regions* and the mobile host is assigned a set of *VCNs*, one for each base station in this region. As soon as the mobile host detects that it is entering another wireless cell, it starts transmitting its messages with the VCN assigned for that base station. This change in position of the mobile host is updated at the root of the virtual connection tree (an ATM switch that maintains the routing tables for this connection) as soon as the first ATM cell from the mobile host arrives bearing the new VCN. The study showed considerable reduction in load on the network call processor. The only time that the network call processor participates in a handoff is when the mobile host changes its neighboring access region. As noted by the authors, handoffs within this connection tree are handled entirely by the mobile itself in a totally distributed fashion.

A *dvct* differs from a virtual connection tree in that the choice of participating base stations and ATM switches depends on the current location of the end-points and may change *dynamically*, i.e., base stations and switches may be dynamically added and removed depending on the movement of the mobile host. All the base stations and ATM switches included in the multicast operation can now be viewed as a *dvct*. Figure 13 is an example of how bidirectional communication takes place between two end points – both of which may be mobile, in a *dvct* using the multicasting approach.

We consider an example to make the *dvct* approach more clear. Suppose that switch A is connected to base stations a and b. Let a be providing a connection between a mobile host $m1$ in its wireless cell and a mobile host $m2$ in the wireless cell of BS d which is connected to switch

D. The table shown in Figure 13 represents the routing information maintained by switch A. Such information is present at all switches in the ATM network. Data coming out of host $m1$ carries VC1 (for BS a) which was assigned at connection set up time. Switch A translates VC1 to VC2 after a look up of its routing table and sends out this data through port 2. Switch B further translates the header information so that it now carries the VC3. Finally switch D translates this to VC5 before passing it on to the BS d and from there to host $m2$. On the return path, $m2$ sends out data carrying the VC7. Suppose the multicast group members for mobile host $m1$ are base stations a, b and c. Then, switch D translates VC7 to VC8 followed by translation to VC13 (and VC9 for multicast) at switch B. Finally switch A translates VC13 to VC14 (and VC15 for multicast) for the multicast members a and b. The onward transmission to host $m1$ is done by a. Now if host $m1$ hands off to base station b, it will continue normal transmission but with VC16, and continue receiving with VC15. It is easy to see that allocating VCNs to all base stations that are included in a multicast, will greatly ease the handoff process.

The total delay experienced by an ATM cell over the network can be characterized by two main components [23, 22].

$$\tau_{delay} = D_{cons} + D_{var} \tag{2}$$

where D_{cons} represents the constant component and D_{var} represents the variable component of the delay. D_{cons} depends on the the physical delay of the medium and the distance an ATM cell has to travel between source and destination. D_{var} on the other hand is representative of the variation in queueing delays experienced by different ATM cells over the same connection in the network. Given the nature of delay variation experienced by different ATM cells, it is easy to see that different cells may experience different total delays over the same connection. This variation in cell delay is also referred to as *jitter*. [22] presents delay and delay variation objectives for two-way session audio and video services.

Crossover points within the network have significance when multiple connections are branching off from a common stream. Each connection in a multicast operation need not start from the source but may in fact find an intermediate switch that is handling the connection for some other base station (See Figure 11 for an example of crossover point location during handoff.). We model the delay experienced by an ATM cell over different routes starting from the crossover point to be bounded by the times τ_{min} and τ_{max}. The delay variation for each connection may now be viewed simply by a *delay pipe* as shown in Figure 10. The tail of the pipe represents the entry point of an ATM cell from the source.

It is evident from Figure 10 that at any given instant for a multicast operation, the tail of each pipe contains the same ATM cell. However, due to different delays experienced on different routes, the ATM cells coming out from the heads of different pipes to the respective base stations may not be the same.

If the mobile host is to get consistent information from a base station during and after

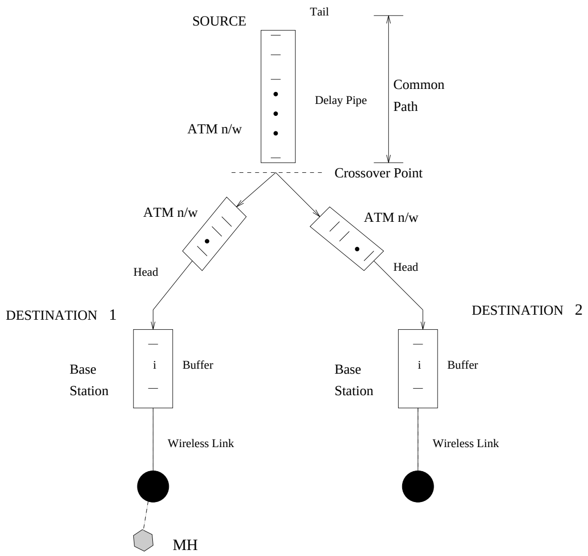


Figure 10: Delay Pipe Model

handoffs, then we have to make sure that the base stations involved in the handoff procedure have corresponding ATM cells in their buffers. Using this information and from the discussion above, it follows that the buffer requirement for ensuring that consistent information is present at the participating base stations is given by

$$Buffersize \geq BW_{conn} \times (\tau_{max} - \tau_{min}) \tag{3}$$

where BW_{conn} is the bandwidth of the connection.

The ATM cell stream originating at the source consists of cells arriving back to back with no way of differentiating between two data cells. Of course, special cells may be generated by setting the appropriate bits in their headers, but this is not the case with data cells in particular. Extensions to existing ATM protocols to suit the mobile environment are discussed in [4]. However, our solution does not require any changes in existing protocols but targets ATM switch fabrics to achieve its goals.

In Figure 11 , BS1 is the base station that is currently transmitting to the mobile host and BS2 is the base station that is required to join the multicast. After waiting for time $t_{stagger}$, BS1 sends out a request to the switch to include the base stations in the multicast group of MH (g_{MH}) in the multicast operation. The upper bound on time taken for this is represented by t_{setup}. Note that t_{setup} includes the time required to

- find the crossover point between BS1 and the base station farthest (in terms of number of intermediate switches) from it (BS2 in Figure 11),
- to update the multicast table entries in the crossover switch and

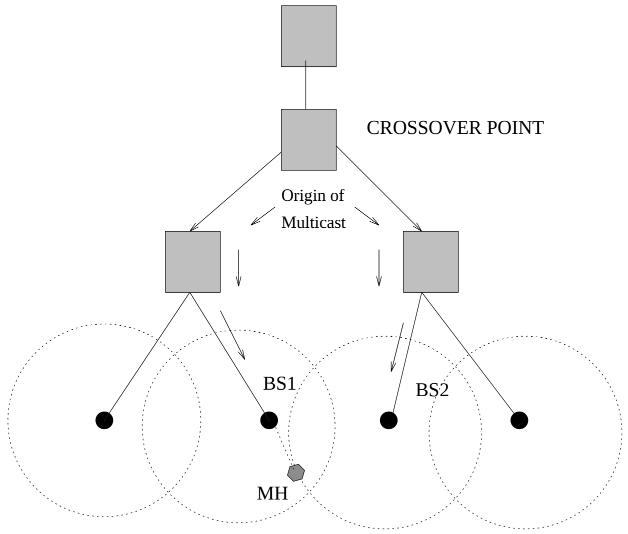


Figure 11: Handoff between Base Stations

- to send the newly allocated VCNs of each base station to the mobile host.

As mentioned earlier, each time the mobile host performs a handoff[6] it is necessary to ensure that the sequence of frames being received from the old and new base stations preserve their relative order. We propose to overcome this problem by generating a control cell at the crossover switch when a new multicast connection is admitted. This control cell will act as a reference point within the ATM cell stream to facilitate framing at each base station.

Implementation of our scheme will require minor modifications at the switch level. We would require the mobile host to maintain some kind of a record of the last frame number correctly received from a base station. A representative switch fabric that supports multicast (broadcast) [21] is shown in Figure 14. The modifications proposed to this switch architecture, however, are general enough to be applied to any other existing architecture. Our purpose is only to demonstrate how our scheme can be implemented. In the original switch architecture, CP is responsible for establishing both point-to-point and multicast connections. CN makes copies of the incoming ATM cells while the BGTs fill out the header information for each ATM cell generated by the CN (for multicast) as well as perform header translation for unicast cells. The DN distributes traffic over its outlets as uniformly as possible. For a comprehensive survey on switch architectures see [23, 22].

The modifications required are shown with dashed lines in Figure 14. On receiving a multicast join request, the CCGL will request the CN to generate an empty cell of 48 bytes while the CP is setting up the multicast connection. A BGT will then attach the header of this control cell and appropriate values of *Cell Loss Priority (CLP)* and *Payload type (PT)* bits will be filled in.

[6] Note that both BS1 and BS2 may not be connected to the same ATM switch. In fact the crossover point could require a number of hops to be made.

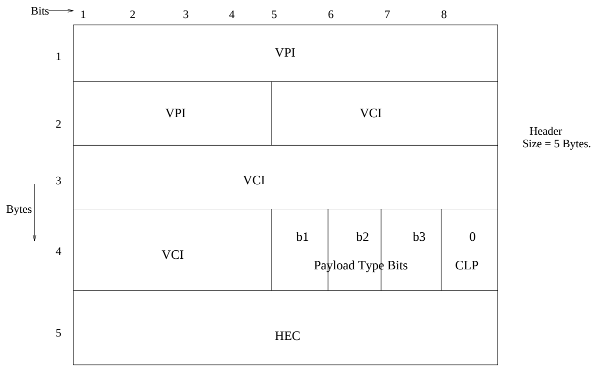


Figure 12: Possible header of a Control Cell

Note that the control cell will have its CLP bit set to 0 (high priority) and PT bits (b1, b2 and b3) may be set such that this control cell is distinguished from ordinary data cells. The VPI and VCI bits will be identical to their counterparts in the data cells. Figure 12 depicts a possible header configuration for a control cell.

When BS1 receives this control packet, say cp (which may be in the midst of regular cells all belonging to a single frame), it continues its framing process as before but sets a special flag in this frame before it goes out to the mobile host. The next frame to be transmitted will be numbered 1. The special flag that was set in the last frame to be transmitted will cause the mobile host to reset the frame counter it maintains to 0. Note that the mobile host does so only after it has received all previous frames from BS1 correctly. This will ensure that there is no confusion if requests for retransmission are generated by the mobile host on account of erroneous transmission from BS1. On receipt of cp, BS2 starts framing ATM cells (F cells per frame) and also starts numbering them from 1.

When handoff actually takes place, the mobile host will be able to specify the last frame completely received from BS1 (say n) so that BS2 can can send the next appropriate frame to the mobile host. If BS2 starts transmitting from frame $n+1$ onwards, it may result in one frame being completely duplicated at the mobile host in the worst case. The extent of duplication depends on the position of cp relative to the boundary of a frame being generated at BS1. Note that this duplication could be as small as a single ATM cell if the wireless protocol adopted transmits one ATM cell at a time instead of a larger frame. In any event loss of ATM cells will not occur.

Given below is the expression for the synchronization time (T_{synch}) required to ensure that duplication in receiving ATM cells at the mobile host is limited to at most one frame.

$$\begin{aligned} T_{synch} &= t_{setup} + 2\tau_{max} - \tau_{min} + \tau_{bs2} \\ &\quad + \frac{F}{BW_{WL}} + \tau_{WLL} \end{aligned} \tag{4}$$

where BW_{WL} is the bandwidth of the wireless link, and τ_{WLL} is the latency of the wireless link. τ_{max} represents the upper bound on the time taken for the first frame to be reach BS2. An additional $(\tau_{max} - \tau_{min})$ represents the upper bound on the time required to flush out the ATM cells which was already received by BS1 before cp arrived. The expressions τ_{bs2} and $\frac{F}{BW_{WL}}$ represent the processing time required for a frame at BS2 and the time required to transmit a frame over the wireless link respectively.

Note that this analysis assumes that the delay associated with ATM cells reaching $BS1$ is τ_{min}, and the delay for ATM cells reaching $BS2$ is τ_{max}. This analysis will produce the worst case value of T_{synch}.

The actual stagger time available for this connection is now given by

$$t_{stagger} = t_s - T_{synch} \tag{5}$$

where, t_s is the stagger time determined for a particular QOS requirement of the user (see Section 3).

Note that the synchronization time presented above and the chosen buffer size of $(\tau_{max} - \tau_{min}) \times BW_{conn}$ at each base station, will together ensure that the frame being currently transmitted to the mobile host is within the buffer for each base station in the multicast.

The scheme presented here may result in the duplication of a single frame of ATM cells as explained above. However, this duplication could be as small as a single ATM cell if the wireless protocol adopted transmits one ATM cell at a time instead of a larger frame. It is possible to avoid any duplication if control cells can be generated at the source itself. However, this may require a change in the existing ATM protocols. In this paper, we do not consider such a situation but it is evident that if such a change is brought about in the future, then we will be able to perfectly synchronize frame reception at the mobile host.

5 Conclusion

There are many user applications that do not require a "total" guarantee for disruption free service but would also not tolerate very frequent disruptions. An user will not not want to pay a high cost for such applications. Thus if a multicast based approach is used, the data packets will be multicast to the neighboring wireless cells throughout the connection. This will be prohibitively expensive. On the other hand, if forwarding is used during handoffs, the user will see a break in service during every handoff. With the decreasing cell sizes, the user might see a disruption every 5 seconds (in picocellular environments). Proposed in this paper is a novel staggered multicast approach which provides *probabilistic* guarantee for disruption free service. The main advantage of the staggered multicast approach is that it partially provides the benefits of the multicast approach and also provides the much required savings in the static network bandwidth.

In summary, the main features of the staggered multicast approach are the following:

- The network bandwidth usage is significantly reduced.
- A probabilistic guarantee for disruption free service is provided.

Using the ATM switch modifications as suggested in the implementation section of the paper, we can ensure lossless data delivery to the end user.

We are currently investigating staggered multicast schemes where the stagger time is determined dynamically during the handoff process. We believe that a dynamic stagger will more provide a much better performance than the static stagger scheme proposed in this paper. On the other hand, if there are sophisticated wireless adapters available that can provide an intermediate signal level which will notify the mobile host that a handoff will soon occur, then the multicast initiation could be staggered till this point.

References

[1] R. Ghai and S. Singh, "An Architecture and Communication Protocol for Picocellular Networks," *IEEE Personal Communications Magazine,* pp. 36-46, Vol.1(3), 1994.

[2] K. Keeton et.al., "Providing connection-oriented network services to mobile hosts," *Proc. of the USENIX Symposium on Mobile and Location-Independent Computing,* Cambridge, Massachussets, August 1993.

[3] Anthony S Acampora, Mahmoud Nagshineh, "An Architecture and Methodology for Mobile-Executed Handoff in cellular ATM Networks," *IEEE Journal on Selected Areas on Communications,* October, 1994.

[4] D Raychauduri and N Wilson, "ATM Based Transport Architecture for Multiservices Wireless Personal Communication Networks." *IEEE Journal on Selected Areas on Communications,* October, 1994.

[5] R. Thomas, H. Gilbert, and G. Mazziotto, "Influence of the Mobile Station on the Performance of a Radio Mobile Cellular Network," *Proc. 3rd Nordic Sem.,* paper 9.4, Copenhagen, Denmark, Sep., 1988.

[6] A. Papoulis, "Probability, Random Variables, and Stochastic Processes," Third Edition, McGraw-Hill, Inc.

[7] E. Kreyszig, "Advanced Engineering Mathematics," Fifth Edition, John Wiley & Sons, Inc.

[8] W. C. Y. Lee, *Mobile Cellular Communications Systems,* McGraw Hill, 1989.

[9] D. M. Balston and R. C. V. Macario, *Cellular Radio Systems,* Artech House, 1994.

[10] S. Mohan and R. Jain, "Two User Location Strategies for Personal Communication Services," *IEEE Personal Communications,* Vol. 1, No. 1, 1994.

[11] P. Krishna, N. H. Vaidya and D. K. Pradhan, "Location Management in Distributed Mobile Environments," *Proc. of the Third Intl. Conf. on Parallel and Distributed Information Systems,* pp. 81-89, Sep. 1994.

[12] P. Krishna, N. H. Vaidya and D. K. Pradhan, "Efficient Location Management in Mobile Wireless Networks," Technical Report, Dept. of Computer Science, Texas A&M University, Feb., 1995.

[13] C. Lo and R. Wolff, "Estimated Network Database Transaction Volume to Support Wireless Personal Data Communications Applications," *Proc. of Intl. Conf. Communications,* May, 1993.

[14] Pravin Bhagwat and Charles. E. Perkins, "A Mobile Networking System based on Internet Protocol (IP)," *Proc. of the USENIX Symposium on Mobile and Location-Independent Computing,* Cambridge, Massachussets, August 1993.

[15] J. Ioannidis et. al., "IP-based Protocols for Mobile Internetworking," *Proc. of ACM SIGCOMM,* 1991.

[16] J. Ioannidis and G. Q. Maguire Jr., "The Design and Implementation of a Mobile Internetworking," *Proc. of Winter USENIX,* Jan. 1993.

[17] Charles Perkins, "Providing Continuous Network Access to Mobile Hosts Using TCP/IP," *Joint European Networking Conference,* May 1993.

[18] F. Teraoka, Y. Yokote and M. Tokoro, "A Network Architecture Providing Host Migration Transparency," *Proc. ACM SIGCOMM Symposium on Communication, Architectures and Protocols,* 1991.

[19] C. Partridge, "Gigabit Networking," Addison Wesley, 1993.

[20] International Telecommunication Union Recommendation I.311 (03/93)

[21] J S Turner, "Design of a broadcast Packet switching network," *IEEE Transactions on Communications*, vol. 36, pp. 734-743, June 1988.

[22] Raif O Onvural, "Asynchronous Transfer Mode Networks : Performance Issues," Artech House, 1993.

[23] Martin de Prycker, "Asynchronous Transfer Mode, solution for broadband ISDN," Ellis Horwood 1991.

Appendix 1 : Two Dimensional Random Walk Model

We will first discuss the one dimensional random walk model as explained in [6], and then extend it to two dimensions.

Let the position of the user after t units of time be $x(t)$. Let the one dimension be the X axis. In the one dimension random walk model, every T units of time, the user tosses a coin, and based on the result the user either decides to go in the positive X direction or the negative X direction. For example, upon a head the user decides to take one step in the positive X direction, and upon a tail the user decides to take one step in the negative X direction. For the purpose of this discussion, we will assume that the step size if small enough so that a step in any other direction can be approximated by one of the four directions mentioned here.

The important parameters in the model are the time interval between two tosses T and the length of the step s. It is shown in [6] that for $t \gg T$, $x(t)$ is normally distributed with zero mean and variance αt as shown in the following equation.

$$f(x,t) = \frac{1}{\sqrt{2\pi\alpha t}}\, e^{-\frac{x^2}{2\alpha t}} \tag{6}$$

where $\alpha = s^2/T$. It is assumed that the user starts from the origin. It is also assumed that the successive steps are independent of each other.

We can extend this analysis to two dimensional random walk model. Let the two dimensions be X and Y. Let positive X axis represent the east direction, and the positive Y axis represent the north direction. In such a model, the user tosses two coins every T seconds. Based on the resulting head-tail combination the user will decide to take a step in a specific direction. For example, a head-head results in the user taking a step in the north-east direction, a head-tail results in the user taking a step in the north-west direction, a tail-head results in the user taking a step in the south-east direction, and a tail-tail results in the user taking a step in the south-west direction.

We assume that the movement in the X-dimension is independent of the movement in the Y-dimension, and that the distribution functions are identical for both the dimensions. Thus, the joint density of the two dimensional random walk will be given as follows:

$$f(x,y,t) = \frac{1}{2\pi\alpha t}\, e^{-\frac{x^2+y^2}{2\alpha t}}$$

Let us assume circular cells of radius R units. At time t, the user will be in the same cell if $x(t) < R$, and $y(t) < R$. Therefore,

$$Prob(x(t) < R, y(t) < R) = \int_0^R \int_0^R f(x, y, t)\, dx\, dy$$

Converting into polar coordinates (r, θ) we get,

$$Prob(r(t) < R) = \int_0^{2\pi} \int_0^R f(r, \theta, t)\, J\, d\theta\, dr$$

where J is **Jacobian** [7], which is given as follows:

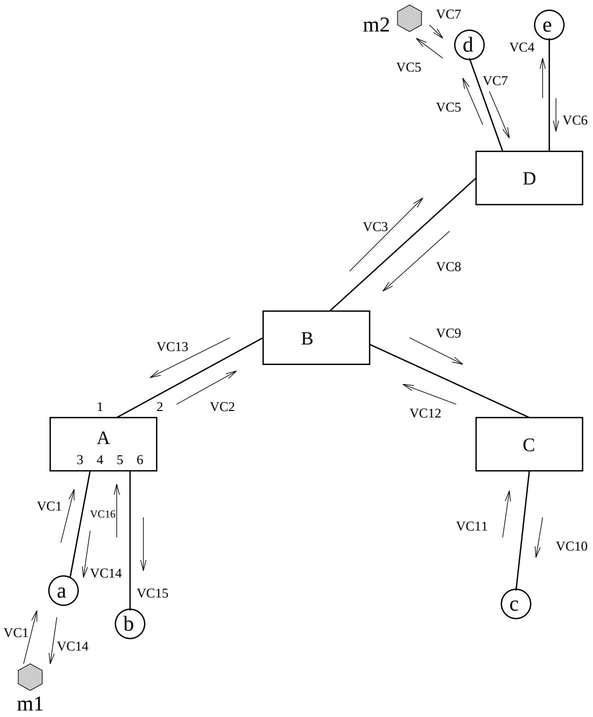
$$J = \frac{\partial(x, y)}{\partial(r, \theta)} = \begin{vmatrix} cos\theta & -rsin\theta \\ sin\theta & rcos\theta \end{vmatrix} = r$$

Replacing J, we get,

$$Prob(r(t) < R) = \int_0^R \frac{r}{\alpha t}\, e^{-\frac{r^2}{2\alpha t}}\, dr$$

Therefore,

$$Prob(r(t) < R) = 1 - e^{-\frac{R^2}{2\alpha t}} \tag{7}$$



VCN in	PORT in	VCN out	PORT out
1	3	2	1
16	5	2	1
13	2	14, 15	4, 6

Routing Information
For Switch A

ATM Switch

VC - Virtual Circuit #

Figure 13: Dynamic Virtual Connection Tree

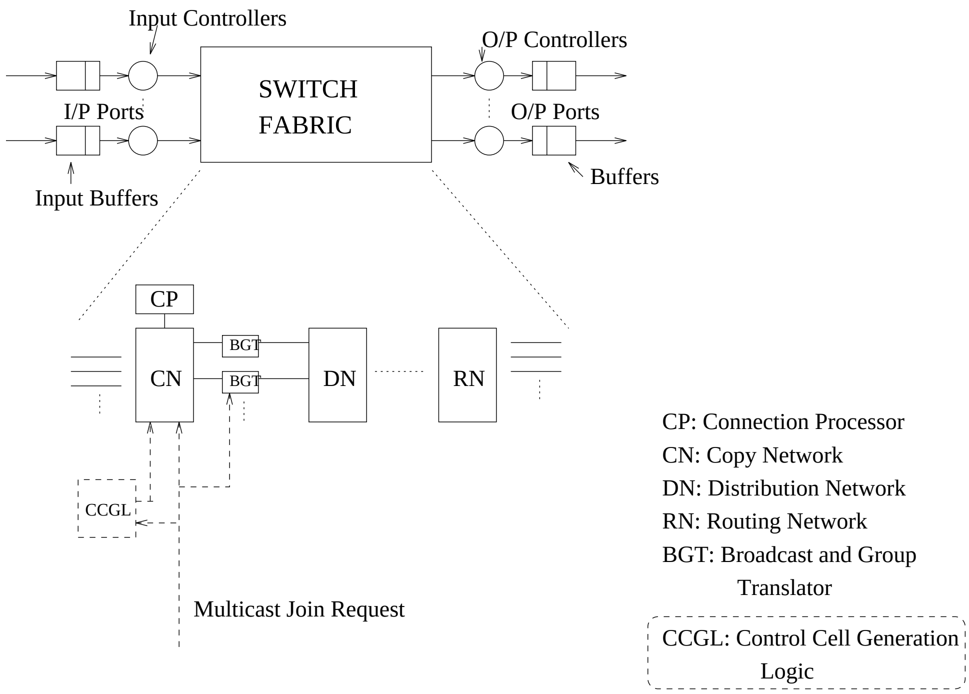


Figure 14: Switch Fabric modifications