Low Latency Handover in a Wireless ATM LAN

John Naylon, Damian Gilmurray, John Porter and Andy Hopper

Abstract— **The micro- and pico-cellular architectures proposed for wireless ATM LANs lead to wireless terminals frequently changing their point of attachment to the network. Because ATM connections have QoS guarantees which must be maintained, handover must be as seamless as possible. We present a novel architecture and protocol which primarily aims to keep the interruption period due to handover low, rather than seeking to keep the process entirely lossless. We compare the trade-offs made with those in other schemes from the literature and give quantitative results from an implementation of our scheme on a 10Mbit.s^{-1} prototype wireless ATM LAN.**

Keywords— **wireless LAN, asynchronous transfer mode, wireless ATM, mobility, handover, signalling protocols, quality-of-service.**

I. Introduction

Most of the proposed architectures for indoor wireless ATM LANs employ a micro- or pico-cellular architecture [1], [2], [3], [4], [5]. In such systens wireless terminals (WTs) will frequently change their point of attachment to the network due to movement of the WT, radio channel effects or network management intervention (for instance load-balancing). When this occurs, the WT's active connections must be transferred to the new access point (AP) so that connectivity is maintained. Ideally, applications running on the WT will see no effects, either transient or persistent, on data transport during this process. In order for the applications' desired QoS to be maintained in this way, the design and performance of the handover scheme responsible for the connections' transfer is of primary importance.

Schemes to support wireless ATM handover can broadly be divided into *virtual connection tree* [2], *path rerouting* [6], [4] and *path extension* schemes [7], [5]. A virtual connection tree describes a collection of paths from a root switch in the wired network to a group of APs. A WT connection, when admitted to a connection tree is allocated a set of VCs, one on each path to an AP. On handover, the WT switches to the appropriate path by changing the VCI of the connection. In a path rerouting scheme, a connection is rerouted on handover from the old AP to the new AP at a particular switch, termed the *crossover switch* (COS), in the wired network. The COS can be chosen either statically when a WT registers at a network or dynamically when a WT hands over to a new AP. It can be chosen for a group of connections or on a per-connection basis. In path extension schemes, a connection is extended on handover from the old AP to the new AP. This typically requires techniques to eliminate routing loops and inefficiencies.

The detailed requirements of a handover protocol are examined in Section II. Section III outlines the key architectural features of a path rerouting handover protocol satisfying these requirements. Sections IV and V describe and analyse the protocol in detailand it is compared with others in the literature in Section VI. The experimental 10Mbit.s^{-1} wireless ATM LAN that formed the basis for our experiments is described in Section VII. Measurements of the performance of our protocol and interpretation of these results are given in Section VIII. Finally a summary of this work is presented in Section IX.

II. Requirements

Outlined below are the most important criteria that a successful handover protocol for an indoor wireless ATM LAN must fulfil. For each criterion, a discussion of its importance is given.

A. Minimal service interruption

Clearly, it is desirable to minimise the length of time that a WT is out of touch with the wired network during handover. However, as we discuss further below, it is more important that the handover process continues to support the QoS properties of ATM.

B. No cell re-ordering or duplication and minimal cell loss

A fundamental principle in ATM networking is that the relay of cells within a single VC maintains cell sequence integrity. Thus it is essential that cells are not re-ordered or duplicated during handover. While ATM networks do not guarantee zero cell loss, unavoidable cell loss is undesirable and must be constrained so that connections' cell loss ratio (CLR) guarantees are honoured.

C. QoS maintenance

QoS guarantees associated with connections should be preserved during handover. Many of the important traffic parameters, such as cell loss ratio (CLR) and cell delay variation (CDV) will be limited by the length of transport interruption and degree of cell loss or re-ordering associated with handover. The relative importance of CDV versus CLR varies with the traffic type; for example, for delay-sensitive traffic such as voice we consider that it is more important to constrain the CDV than to ensure delivery of every cell. The opposite is true for loss-sensitive traffic such as data.

Manuscript received October 30, 1997; revised January 20, 1998.

The authors are with the Olivetti & Oracle Research Laboratory (ORL), 24a Trumpington Street, Cambridge, United Kingdom. E-mail: {jbpn,dpg,jdp,ah}@orl.co.uk.

John Naylon is also with the University of Cambridge Computer Laboratory, Pembroke Street, Cambridge, United Kingdom. Andy Hopper is also with the University of Cambridge Department of Engineering, Trumpington Street, Cambridge, United Kingdom.

D. *Scalable to many VCs and many WTs*

A mobile multimedia device is likely to have many VCs active at a given time; furthermore, our architecture allows for mobile networks, so it is important that the performance of a handover protocol should not degrade as the number of VCs to be transferred increases. Many WTs may initiate handover simultaneously either to or from a particular AP; again, handover performance should not degrade.

E. *Backward compatibility*

Any handover protocol should support existing applications, hosts and switches with no mobility enhancements. Applications should not need to be aware that they are operating over a wireless ATM network but this ought not to preclude mobile-enabled applications able to adapt to the effects of mobility. Although some parts of the network need to be modified to support mobility, unmodified hosts and switches should inter-operate transparently with those with mobility enhancements. Thus the modifications to existing signalling and routing protocols should be minimised. In particular, the handover protocol should not require specialised hardware such as additional buffering support in switches which would increase their cost and complexity.

F. *Resource utilisation*

Network resources should not be reserved for longer than is necessary to ensure QoS maintenance as above. In the local area, the wired network is likely to be over-provisioned, but wireless bandwidth will typically be more scarce. A compromise needs to reached between increasing the resources used in the wired network and increasing the disruption to cell transport during handover.

G. *Robustness and stability*

Interference or fading in the radio channel will cause a WT to initiate a handover. Thus, the handover protocol will need to operate when the radio channel is at its least stable. Any handover protocol which relies on communication over this channel must cope with both lost and duplicated messages. The handover procedure should exhibit stability such that the handover decisions of WTs remain valid over a period of significantly greater than the duration of a handover. This will reduce the likelihood of a WT oscillating back and forth between two APs.

III. Key Architectural Features

Our handover procedure allows for both hard backward and hard forward handovers, with backward handover being used preferentially. We do not consider soft handovers since these are not appropriate for the typical physical layer implementations being considered for wireless ATM. A fundamental principle in the design of the handover scheme is that it is more important to keep the interruption to data flow small than to aim for minimal cell loss. We describe here the major features of our scheme which make this possible.

A. *Path Rerouting Using a Static COS*

A novel aspect of this architecture is the use of path rerouting of all a WT's VCs at a single static COS.

When a WT registers at a network, a COS is allocated to it. For the LAN environemnt the position of the COS is of minor importance, since any routing inefficiencies introduced in a small network will be insignificant. In our scheme, a number of COSes are provisioned in the network and are allocated uniformly to WTs according to a simple algorithm. The position of the COS therefore does not necessarily provide optimal routing of connections to the initial AP but instead provides *acceptable* routing to all APs in the network. When a WT initiates a handover to another AP, therefore, there is no need to perform a search for a COS for each of its connections. In this way, we trade somewhat increased resource utilisation, in terms of potentially sub-optimal routing of VCs through the wired network, for speed, simplicity and scalability in the handover procedure.

For large networks with many APs, after the WT has performed a number of handovers, the inefficiency of this routing may become excessive. We make provision for the allocated COS to migrate in this case; the details may be found in [1].

B. *Two-Phase Handover*

There are two separate but overlapping procedures involved in the handover of a WT. One is the *network handover* where the virtual circuits of the WT are transferred from the old AP to the new AP. The other is the *radio handover* where the WT relinquishes its radio channel to its old AP and establishes a radio channel to the new AP. Niehaus *et al.* show in [8] that the time taken to establish a local-area connection using current implementations of ATM Forum UNI signalling [9] is in the range of tens of milliseconds. However, the radio handover can be reduced to the order of a few hundred microseconds.

Thus, the time spent duplicating the WT's connections to the new AP will dominate the total duration of handover. This is particularly true when the WT has a large number of connections. As we explain in detail below, this architecture decouples this lengthy process from the much shorter radio handover, thereby ensuring that a WT suffers an interruption to its connectivity only during the radio handover.

C. *User-Plane Handover Signalling*

There are two commonly proposed approaches to providing handover signalling support within a wireless ATM LAN. One approach is to modify existing ATM control-plane signalling protocols such as UNI and PNNI [10] to include specific support for handover signalling. A second, more flexible, proposal currently under consideration at the ATM Forum is to provide a generic signalling transport mechanism that could be utilised for handover signalling.

Rather than either of these methods, we have chosen in this architecture to employ *user-plane signalling channels*. These are simply switched user plane VCs used as channels to connect the software entities located on switches, APs and WTs which support a WT's mobility. These SVCs use AAL5 as their adaptation layer and provide reliable transport of messages by the combined use of acknowledgements and timers. We note that by requesting a QoS appropriate to a very low rate CBR connection (very low PCR, low CTD, CDV and CLR) for these user-plane signalling channels we are able to achieve deterministic performance of the handover protocol. In contrast, the hop-by-hop processing of UNI or PNNI messages can introduce large and variable delays into the system because such messages must traverse at least one protocol stack at each hop. Current ATM signalling systems are unable to offer any upper time bound on such delays, which will also be dependent on prevailing network conditions and individual switch implementations.

In addition to improving overall performance, employing user-plane rather than control-plane signalling avoids both the need to change existing signalling protocols and the upgrade effort associated with such changes. Furthermore, this approach enables future modifications to handover signalling to be made independently of changes to standard signalling protocols (and vice versa). Finally, integration with other systems requiring handover support but having incompatible signalling systems, for example PCS, is greatly eased by this approach.

IV. Handover Procedure

To describe the handover procedure fully, it is necessary first to outline the registration procedure and the aspects of our location management scheme relevant to connection setup to a WT. Some details are omitted and may be found in [1] and [11].

A. Registration

When a WT powers up or roams into a network, it is required to *register* with the network. The registration procedure allows the network's administration to authenticate the device and allocate resources to it. To register, a WT meta-signals its 48-bit IEEE MAC address[1] to its chosen AP; the AP responds to this by allocating a *mobile identifier* (MID) for the WT. A WT's MID is used to identify a WT uniquely to its current AP and is part of the extended ATM header used for each cell transmitted over the wireless link. The MID may be thought of as a virtual port number on the AP; in common with ports on normal switches, each such virtual port has a separate VPI/VCI space. Thus the full VPI/VCI space is available to each WT and transmissions to and from different WTs are distinguishable, even if on the same VPI and VCI, by the MID field of the extended cell header. Note that the extended header is used only over the wireless link; the AP provides a simple header translation function to effect this.

Once a MID has been assigned, a handover signalling channel using a well-known VCI over the radio link is automatically established between the AP and the WT. The authentication procedure is then invoked over this channel and, if successful, a static COS (and associated address—see below) is allocated to the WT. The COS establishes a handover signalling channel to the AP which can then send a confirmation to the WT thereby completing the registration.

B. Addressing and Routing for WTs

A WT is assumed to originate from a particular network, termed its *Home Network*. Services within this network have administrative responsibility for the WT and maintain information such as the current location of the WT and security details for authentication services. When a WT registers in a network other than its home network this network is termed its *Foreign Network*. Services within the foreign network are responsible for managing the small scale mobility of the WT within this network and for registering the current location of the WT.

A *Mobile Home Address* (MHA), specifying the home network, is permanently allocated to the WT which stores it in non-volatile memory. A possible mechanism for allocating this address is to derive it from the network prefix of the home network and the MAC address of the WT. Endpoints which wish to establish a connection with a WT use its MHA as the called party number in the UNI SETUP request. Similarly WTs which wish to establish a connection use their MHA for the calling party number in the SETUP request.

As described above, when a WT registers in a foreign network it is allocated a static COS and also a *Mobile Foreign Address* (MFA) by a service in that network. This address will normally be derived from the network prefix of the COS in the foreign network and the MAC address of the WT and will be topologically significant to the routing protocols of PNNI in the usual way. Registration of a WT results in the communication of its MFA to the WT's home network, where a location service maintains the authoritative mapping of MHA to MFA. This location service is likely to be structured along the lines of the Home Location Register/Visitor Location Register scheme employed in GSM [12]; the details are beyond the scope of this paper.

Figure 1 illustrates how an unmodified host in the wired ATM network establishes a connection to a WT which has roamed into a foreign network and completed its registration procedure. This fixed terminal has the address A.1.10 and wishes to make a connection to the WT whose MHA is C.2.11. It issues a SETUP message indicating A.1.10 and C.2.11 as the calling and called party numbers respectively. As the SETUP propagates through the network, as indicated by the light dashed arrows in the figure, it eventually reaches a switch running enhanced software capable of distinguishing C.2.11 as the address of a mobile device.[2]

[1] As suggested for use as the End System Identifier (ESI) in ATM Forum standards.

[2] We discuss in detail how this distinction may be made in [11].

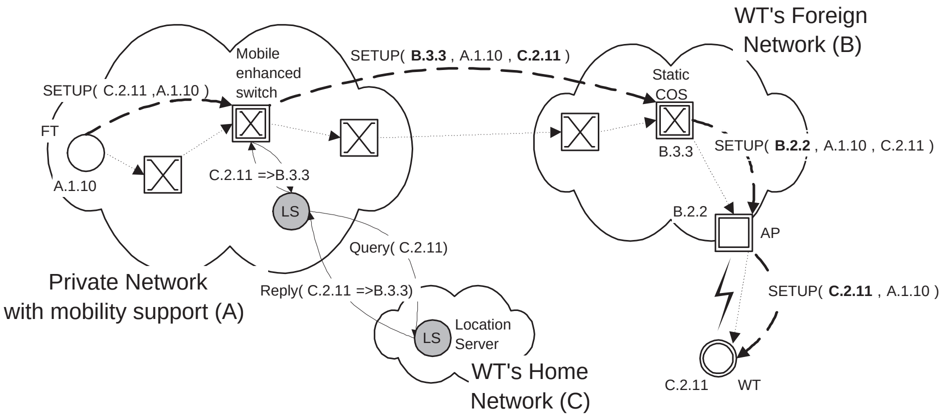

Fig. 1. Connection setup to a WT, showing our location management and addressing scheme.

This switch queries the location service to obtain the current mapping from MHA to MFA.

In the example shown, the response to this query is B.3.3. The mobile enhanced switch substitutes the called party number information element (IE) in the SETUP message with the MFA of the WT and encapsulates the MHA in another IE of the message.[3] The SETUP now proceeds through the network, eventually reaching the COS in the foreign network. As the COS is also running enhanced software, it again modifies the SETUP message, replacing the called party number IE with the address of the WT's current AP—here B.2.2. The message may now traverse any number of unmodified switches in the foreign network before reaching the AP. The original SETUP message is then reconstructed with the WT's MHA as the called party number. Both endpoints are therefore unaware of the WT's mobility and see only normal UNI signalling requests.

It will be noted that this approach apparently requires a mobility enhanced switch in every network that contains a host which might wish to communicate with a WT. If a network does not contain such a switch, the SETUP message will simply propagate to the WT's home network, where such a switch is guaranteed to exist. In this case, routing to the WT will be sub-optimal, but we have traded this off against ease of integration of mobility into current networks. In the medium term, modifying only a network's public gateway switch will give optimal routing across the public infrastructure with the minimum upgrade effort.

C. Handover Target AP Selection

In our architecture, an AP periodically broadcasts information about neighbouring APs to its attached WTs. This information includes each AP's identity, physical channel identifier and a resource loading metric for that AP. In order to determine suitable candidate APs for handover, a WT periodically switches its physical layer channel to an adjacent AP to determine the received signal strength indication (RSSI) of that AP's transmissions. This information is then combined with the QoS loading information to prioritise the candidate list.

When a WT attempts to handover to a particular AP, the COS will attempt to duplicate all the WT's connections to the new AP. If a recent change in the loading at that AP means that all the WT's connections cannot be supported, then the COS will send a HandoverFail message back to the WT to indicate that its choice of AP has insufficient resources to support its requirements. The course of action is now determined by the WT—if it has another candidate AP with a good enough RSSI, it will attempt to initate a handover to that AP. If no candidate is able to support the combined QoS requirements of the WT's connections, the WT must drop one or more of its connections and repeat its AP selection and handover procedure.

D. Backward Handover Procedure

The following is a description of the procedure for a backward handover of a WT from its current access point, AP$_1$, to another access point, AP$_2$. Figure 2 illustrates the exchange of messages and the path taken by data during the procedure.

Once the WT has decided to handover to AP$_2$, it transmits a HandoverRequest (1) message to AP$_1$ indicating AP$_2$. AP$_1$ acknowledges this immediately with a HandoverRequestAck (2). To begin the network handover, AP$_1$ forwards this message to the COS (3) which acknowledges it (4).

The COS initially creates a user-plane handover signalling channel to AP$_2$. The COS forwards the HandoverRequest (5) message on this channel, which AP$_2$ acknowledges (6). The COS creates, in parallel, duplicates of the COS–AP$_1$ segment of each of the WT's connections using standard UNI SETUP messages. Once the resultant CONNECT messages have been received by the COS, it switches the downstream traffic of these VCs to AP$_2$ while still receiving upstream traffic from AP$_1$; AP$_2$ will buffer this traffic until the WT finalises the handover. Once the process of duplicating the COS–AP VC segments is complete,

[3] The MHA can be encapsulated, for example, in the called party subaddress or the generic identifier transport IE in UNI 4.0.

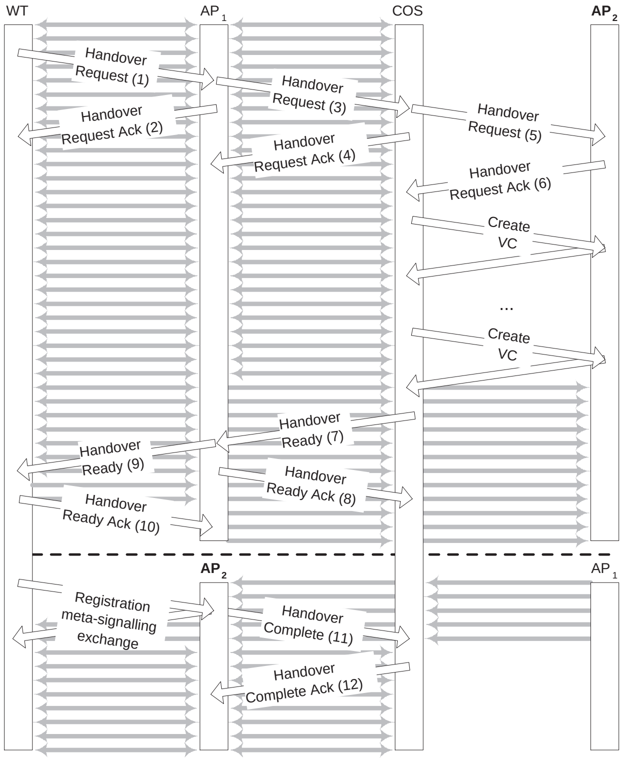

Fig. 2. Handover protocol diagram. The open arrows denote messages and the shaded arrows indicate the path taken by upstream (towards COS) and downstream (away from COS) data traffic during the handover. The heavy dashed line indicates the radio handover.

the COS sends a HandoverReady (7) message to AP$_1$ which acknowledges it (8). When AP$_1$ has flushed any remaining buffered cells to the WT, it forwards the HandoverReady message (9) which in turn causes the WT to reply with an acknowledgement (10). The WT now performs a radio handover by switching its radio channel to AP$_2$. The same procedure as for registration is used again here; the WT thereby obtains a new MID for use at AP$_2$. The WT updates the current MID for its connections with this new MID. Once this procedure is complete, AP$_2$ indicates this to the COS with a HandoverComplete (11) message. The COS switches upstream traffic from AP$_1$ to AP$_2$ and sends a HandoverCompleteAck (12) to AP$_2$. The handover is now complete, i.e. data is flowing bidirectionally to and from the WT via AP$_2$.

Later the COS uses RELEASE messages to close the handover signalling channel and the connection segments to AP$_1$.

E. Forward Handover Procedure

If a WT experiences a sudden degradation in the quality of its link with its current AP, it is unlikely to be able to initiate a handover via that AP. If this is the case, the WT selects a new AP based on its channel quality estimates and immediately performs a radio handover to that channel. It then follows the registration procedure described above, obtaining a MID and a handover signalling channel to the new AP. When the network invokes the authentication process, it is able to detect that the WT has already registered at the network and has ongoing calls. The network can then instruct the COS to create connection segments to the new AP and switch traffic to those segments immediately.

In this case, it is not possible to decouple the connection setup time from the handover period and the WT will thus experience a relatively large interruption to data transfer. This is a problem common to all forward handover schemes. It is for this reason that we employ backward handover whenever possible, only using forward handover as a fallback procedure when necessary, and in the following discussion we focus on the properties of the backward handover scheme.

V. Analysis of this Architecture

There are a number of salient advantages to the architecture as outlined above:

- The following description shows that per-VC cell sequence integrity is maintained during the handover process. Once the COS–AP VC segment duplication process is complete, downstream traffic from the COS is switched from the old AP to the new AP where cells are buffered until the WT performs a radio handover. Any remaining cells at the old AP are flushed to the WT before the radio handover occurs. Upstream traffic is received from the old AP until the COS receives the HandoverComplete message at which time it is switched to come from the new AP. Therefore, in either direction, only one path through the network is active at any one time.
- As discussed above, the use of user-plane handover signalling over AAL5 guarantees a performance gain over using integrated signalling and minimises the requirements imposed on the wired network in terms of per-node processing load. Furthermore, no modifications to standard signalling protocols are required.
- In this architecture, the endpoint VCIs used by the WT are maintained across handover. This means that applications and ATM protocol stacks on the WT need not be made aware of the handover procedure.
- The time taken for the network and radio handover is dominated by the time taken by the signalling system to duplicate the connection segments to the new AP. By employing a backward two-phase handover we minimise the interruption to the connectivity of the WT.
- By utilising a static COS, there is no requirement to locate a suitable COS—a potentially time-consuming procedure—for each of the WT's connections at handover time: this overhead occurs once only, at registration time.
- The use of a single, static COS enables the concept of VC grouping whereby the active connections to a WT can be grouped and handled collectively thereby reducing the overhead of the rerouting procedure. Furthermore, as we need only manage one COS per WT during handover, the overall complexity of the system is reduced and performance improved.
- A static COS isolates the potentially rapid small-scale mobility of a WT within a localised part of the wired ATM

network. The segments of connections between the COS and the remote endpoints remain unaffected during handover.

- The COS requires no additional hardware or buffering support; mobile-enabled software handles the handover procedure whilst buffering during handover is performed at the AP.

The primary disadvantage of a static COS is that within the wired network the routing of connections to a WT may not be optimal; we believe that this is a relatively minor concern in a LAN environment, especially when set against the reduction in the disruption of cell transport during handover.

Our architecture assumes that APs are able to buffer cells that arrive before the WT's radio handover. If the path lengths (in seconds) from the COS to the old AP and the new AP are t_1 and t_2, respectively, and a WT has N VCs, where VC i has a peak cell rate (PCR) of r_i, then an AP needs a maximum of:

$$(t_1 - t_2 + \delta) \sum_{i=1}^{N} r_i$$

cells' worth of buffering to support that WT's handover, where δ is the time between the COS issuing the HandoverReady message to AP$_1$ and the WT completing re-registration at AP$_2$. Now suppose that we wish to support M WTs handing over simultaneously to an AP. The total buffering requirement becomes:

$$(t_{1_j} - t_{2_j} + \delta) \sum_{j=1}^{M} \sum_{i=1}^{N} r_{ij}$$

where r_{ij}, t_{1_j} and t_{2_j} are the obvious extensions. In order to establish an upper bound on this, let us assume the AP has a radio interface with a data rate of 25Mbit.s^{-1}, that the maximum path length difference in the LAN is 3ms and that δ is 10ms. The sum of the PCRs of all VCs routed via an AP will be at most a small multiple of the data rate of the interface; the exact value of the multiplier depends on the burst tolerance. Let us assume a tenfold over-commitment of the bandwidth such that the value of the summation term above is 10×25Mbit.s^{-1}. Under these conditions, the total buffer requirement is 3.25×10^6 bits.

A major assumption made in the design of this protocol is that the WT will usually be able to communicate via its current AP in order to initiate the handover. We have optimised the system for these backward handovers rather than for forward handovers as in some other schemes. This is, however, a reasonable assumption since the coverage regions of neighbouring APs are likely to overlap significantly so as to provide uniform coverage of a building with no shadowed regions.

VI. Survey of Other Handover Architectures

We now compare our handover protocol to other schemes that have been proposed in the literature.

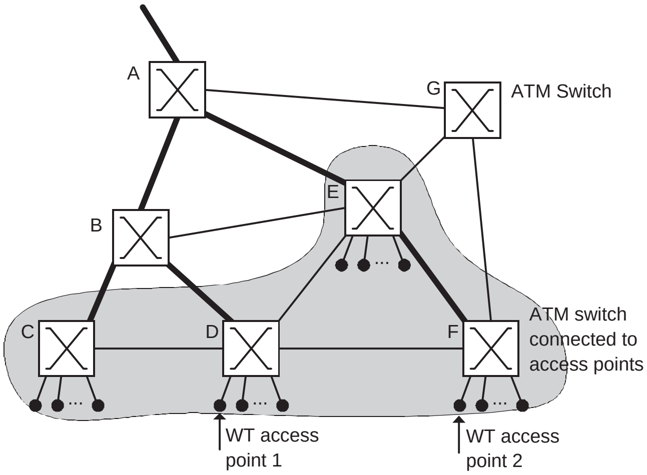

Fig. 3. Virtual connection tree concept as proposed by Acampora *et al.* [2].

A. Virtual Connection Tree

A virtual connection tree as depicted in Figure 3 is a collection of wired network switches, links and APs. The thickened lines represent links in the network which are part of the *virtual connection tree* (VCT) rooted at switch A. The shaded region represents the group of APs termed the *Neighboring Mobile Access Region* which comprise the leaves of this tree. When a connection is established to a WT, a set of VCs are established from the root of the VCT, one to each AP in the VCT. Only one VC, the VC to the WT's current AP, is *active* at any one time.

When a WT wishes to handover, it first performs a radio-level handover to the new AP. It then switches the VCI of each connection to the appropriate VCI at the new AP. When cells on these VCs reach the root of the VCT, a handover event for the WT is indicated, and the root updates the switch routing table entries for the fixed part of each connection to use the appropriate new VC.

In common with our scheme, this approach decouples the connection setup time from the duration of a handover. However, we over-provision a WT's connections only transiently, during one phase of the handover process, whereas here the connections to all the APs in the access region must be permanently pre-allocated. Since VCTs are designed to be large in scope these resource demands will be problematic even in local area networks. Increasing the size of the VCT also has the disadvantage of proportionately increasing the setup time for WT connections.

A second set of problems is due to the lack of explicit handover signalling combined with the exclusive use of forward handovers. In the downstream direction, after the WT performs a handover to the new AP, traffic buffered on the path to the old AP may be lost. If this traffic is not discarded, then it may introduce reordering in the cell stream if the WT quickly hands back to the old AP.In the upstream direction, depending on the relative path lengths in the VCT, cells from a previous AP may arrive at the root of the VCT, falsely indicating a handover event. The lack

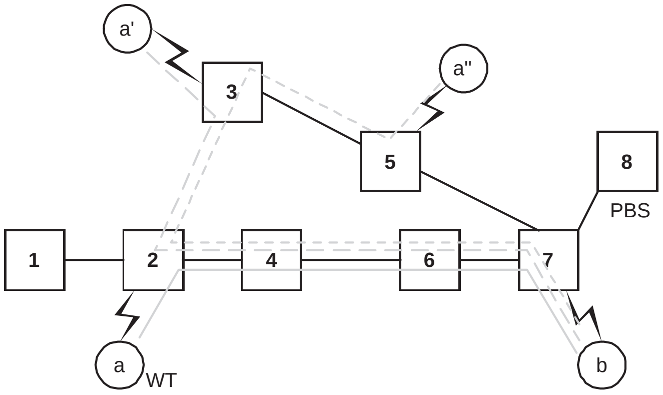

Fig. 4. Path rerouting for handover, as proposed by Eng *et al.* [4].

of explicit handover signalling also assumes the existence of timely upstream traffic to trigger the handover.

This architecture seems most well suited to networks with small numbers of WTs and APs. As we discuss further below, the pre-provisioning of VCs to all APs also makes it ideal for very high mobility environments, although in this case, the price in terms of wired network resource requirements will be steep.

B. Path Rerouting

The fundamental difference between the scheme outlined in Section IV and other schemes in this category described in the literature is the method for selecting an appropriate COS. In our architecture, a COS is selected once per WT per connected session. Most other schemes [4], [6] employ a more dynamic COS selection, where the COS is selected at handover, often on a per-VC basis.

B.1 BAHAMA

In the BAHAMA architecture, Eng *et al.* propose a Homing Algorithm to handle mobility of WTs. To illustrate the function of this algorithm, consider Figure 4. The numbered boxes represent interconnected *portable base stations* (PBSs) via which the WTs (a and b) can communicate. WT a is communicating with WT b and the data path is shown as a solid grey line. For this connection, PBS 2 is referred to as the *Source Home Station* and PBS 7 the *Destination Home Station*. Regardless of the location of a and b, cells from one to the other will always be routed first to PBS 2, then to PBS 7 and finally to b via its current point of attachment to the network. Thus, as a moves to locations a' and a'', the route of cells to b will be as shown by the two dashed grey lines. A Home Station can therefore be viewed as a COS. While Home Stations are associated with connections rather than WTs, typically all connections to or from a particular WT will share the same Home Station. The handover procedure in BAHAMA is a forward handover where the new PBS is used for control of the procedure. On handover, each connection from the WT is rerouted at the WT's Home Station from the old PBS to the new PBS.

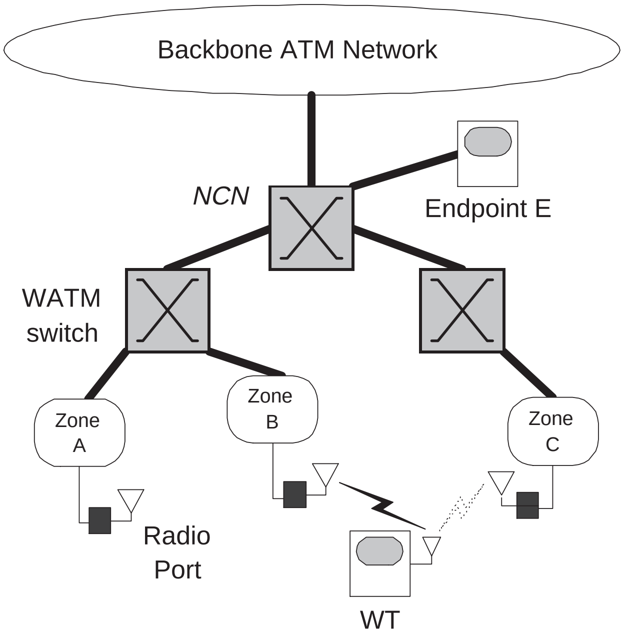

Fig. 5. NCN Rerouting, as proposed by Akyol and Cox [6].

In a similar manner to our scheme, the use of Home Stations through which all a WT's connections are routed enables collective management of those VCs. However, by typically choosing the local PBS as the Home Station, this has the disadvantage of requiring the frequent update of the Home Station, possibly after every handover. Migrating the Home Station requires the exchange of signalling messages with the destination Home Station for each connection and therefore incurs a significant overhead especially for large numbers of connections.

The exclusive use of forward handover means that special measures must be employed to reduce the latency of the handover and hence interruption to data transport. In BAHAMA a new non-standard connection setup procedure is used to reduce such latency. This involves provisioning a virtual path tree which connects all PBS's in the LAN together. Link and node resources in this tree are predivided between the PBS's. Thus, connections with QoS guarantees can be set up without incurring a hop-by-hop delay.

Modifications are also made to the ATM cell header to include sequence numbers. These sequence numbers are used to prevent reordering and duplication in the cell stream during handover. Furthermore, the VPI field of the ATM cell header is used in a non-standard manner as a destination identifier for PBS's. These non-standard signalling procedures, VPI concepts and sequence numbers preclude the BAHAMA architecture being applicable to a general purpose ATM network.

B.2 Cox and Akyol

Cox and Akyol [6] propose an algorithm called *Nearest Common Node Rerouting* (NCNR) to perform the rerouting of connections due to handover. In their network architecture, it is assumed that a group of radio ports is connected to the same wireless ATM network equipment. This group is termed a *zone* and is managed by a *zone manager*. It is assumed that these zones are interconnected by mobility-enhanced ATM switches.

The handover scheme proposed, as further detailed in [13], supports both forward and backward handover. When a backward handover is initiated to a radio port in a new zone, the NCNR scheme dynamically attempts to choose the appropriate COS for each connection to the WT. The COS or *Nearest Common Node* (NCN) for two zones is defined as the closest switch at which the routes to the old zone, the new zone and the remote endpoint are all on different ports.

To illustrate the NCNR handover procedure for a WT connection, consider Figure 5. The WT initiates a handover to zone C from zone B. The zone manager in zone B sends a *handoff start* message towards the remote endpoint E. When a switch receiving this message determines that it is the NCN, in this case the switch denoted *NCN*, it duplicates the connection to zone C. With the rerouting complete, the WT then performs the radio level handover to zone C.

To minimise the service interruption caused by handover, the NCNR procedure forwards data to both zones involved in the handover. However, as the authors acknowledge, this requires that the NCN is chosen such that the delay on the two paths from the NCN to each zone are equal relative to the radio handover delay. The NCN selection procedure therefore tends towards reusing as much of the existing path to the current zone as possible and as such can choose suboptimal paths in the wired network. However, with a suitable NCN, cell loss is minimised and cell sequence integrity in maintained. By employing separate procedures for time-sensitive and throughput-dependent traffic, QoS guarantees can be preserved.

The NCN is superficially similar to the COS of our scheme. However, the dynamic selection of an NCN per VC at handover time introduces a number of disadvantages. First, it will significantly increase the overall duration of a handover since NCN selection is a process requiring processing at each switch. Second, this introduces a requirement for all switches to be upgraded to be able handle the NCN selection message. Finally, the signalling load imposed by this scheme means that it becomes impractical for WTs with more than a few VCs.

C. Path Extension

The following two proposed schemes employ *path extension* rather than path re-routing where additional connection segments are created from a WT's current AP to its new AP and switching functionality at the AP allows data flow to be maintained.

C.1 NEC

In [7], a path extension scheme is described to handover a WT's connections as it moves to a new radio port. As depicted in Figure 6 the connections to the WT are extended though the addition of a subpath from the previous radio port to the new radio port.

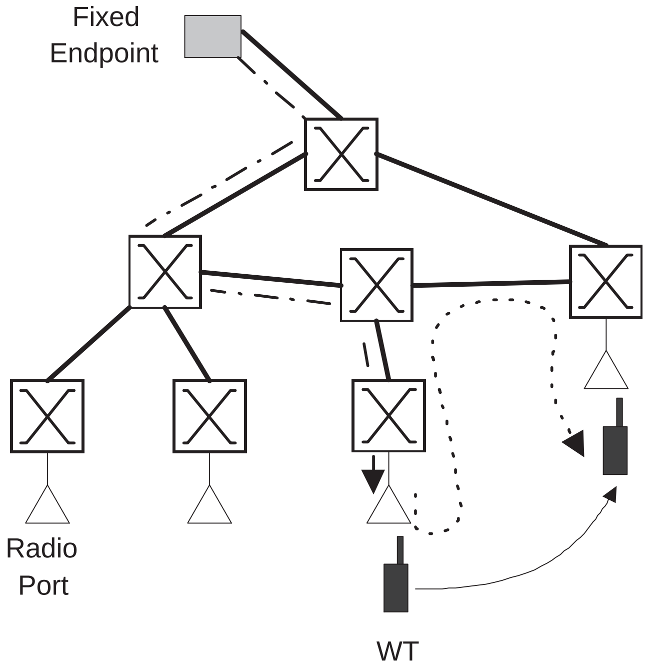

Fig. 6. Path Extension, as proposed by Acharya *et al.* [7].

Service interruption in this scheme is minimal as there is no COS selection phase and no cell delay introduced by buffering during a connection reroute. With no rerouting, cell sequence integrity is maintained and cell loss should be minimal. Path extension allows all the WT's connections to be handled collectively.

However, the major disadvantages of this scheme are that by extending the path, the end-to-end delay is increased and that routing loops can be introduced into the path as the WT repeatedly hands over between a set of radio ports.

To counter these problems, it is proposed that route optimisation be performed on the path in a lazy manner. Route optimisation involves rerouting the path of the WT's connections to produce a more optimal route.

To optimise a route, the loop must first be detected. This scheme proposes additional hardware in switches in order to detect such loops in the path. Buffering support for the data on the path is also required in these switches during the loop detection procedure. Once a loop has been detected, the path must then be rerouted to remove this loop and the buffered data forwarded on the new, optimised path.

C.2 SWAN

A similar scheme is proposed in [5]. In this architecture as shown in Figure 7, the network is divided into domains. As a WT hands off between APs in a particular local domain, its connections are extended to the new AP.

This scheme proposes that a routing loop is detected at an AP when that AP appears twice in a path. This loop

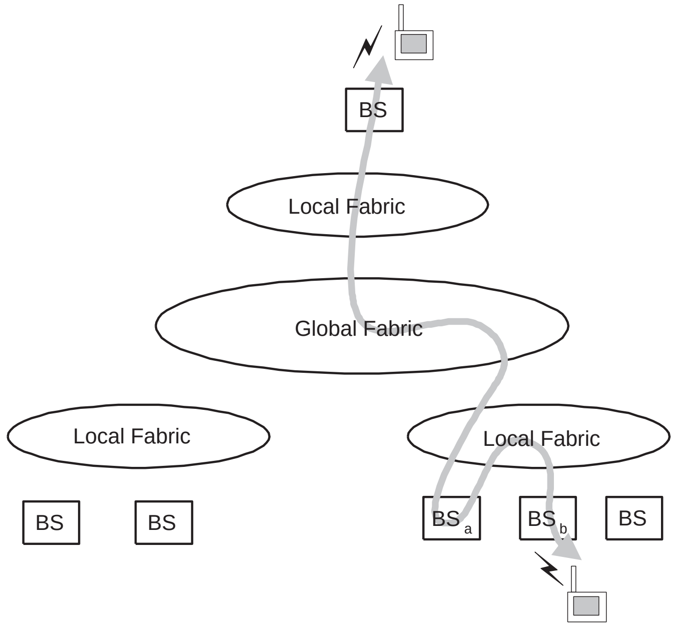

Fig. 7. Path Extension, as proposed by Agrawal *et al.* [5].

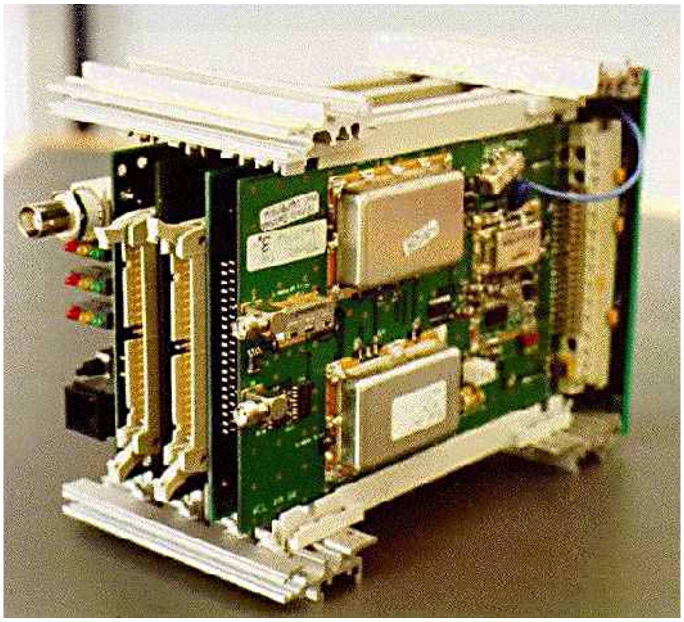

Fig. 8. Prototype implementation of an access point consisting of, from left to right, an ARM-based computer, MAC protocol board, IF board and 2.4GHz RF front end.

is then removed at the AP although no specific scheme for this procedure is described.

When a WT hands over to an AP in a different domain, path extension is not attempted. Instead the path is rerouted from a switch although again details of this procedure are not given.

In both schemes, the need for loop removal and path rerouting is recognised. Depending on the rate of handover, such path rerouting may be required regularly. While neither schemes details the procedure for path rerouting, it is clear that it will be at least as complex as the schemes outlined above.

VII. An Experimental Wireless ATM LAN

An outline only of the ORL experimental wireless ATM LAN is given here. Full details may be found in [1]. The goals of the system are to support mobile multimedia devices and to provide a test-bed for ongoing research.

A. Architecture

The ORL experimental wireless ATM LAN is an access-point based pico-cellular network employing frequency colouring between pico-cells to avoid interference and improve bandwidth reuse. A reservation-based TDMA MAC protocol, designed specifically to efficiently support ATM traffic, is used to share the radio channel among WTs. Each pico-cell is served by an AP which is directly attached to the backbone ATM infrastructure.

B. Implementation

The current realisation of the architecture outlined above is based around a single-board computer with an Advanced RISC Machines ARM processor, an ATM Forum standard 25Mbit.s^{-1} ATM interface, up to 32Mb of memory and a flexible expansion bus. A radio interface is attached to this bus, providing a highly programmable relay between the wired and wireless domains. Figure 8 shows one of these relay devices.

The current radio front-end operates in the 2.4–2.5GHz ISM band with a bandwidth of 10MHz per channel. Transmission power is less than 20dBm and in a normal office building this is sufficient for pico-cells with a radius of around 10 metres. QPSK modulation with a bit-rate of 10Mbit.s^{-1} is employed which reduces the symbol rate and allows us to keep within a 10MHz frequency bandwidth.

We have access to our own switch hardware which is again based around an ARM processor. A light-weight, real-time kernel runs on this hardware and handles signalling and other control functions. The COS functionality and the addressing modifications discussed in Section IV were implemented on this platform.

VIII. Results

In order to quantify the number of cells lost and the degree of cell delay variation (CDV) when handing over from one AP to another, the experimental apparatus shown in Figure 9 was used. The two end-points A and B are single-board computers of the same kind as described above, in this case equipped with GPS receivers and accurate 1MHz clocks. The GPS receivers provide a simultaneous one pulse-per-second (pps) output which is used to maintain synchronisation between the clocks running on the two computers.

Cells with a payload consisting of a sequence number and a time-stamp taken from the local clock are transmitted from end-point A to end-point B at a constant rate. When a cell arrives at B, it is stamped with its arrival time (ac-

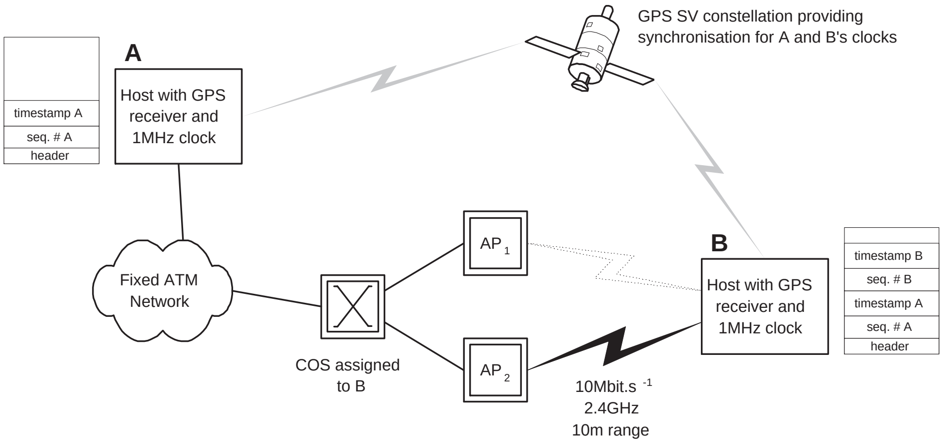
GPS SV constellation providing synchronisation for A and B's clocks
A
Host with GPS receiver and 1MHz clock
timestamp A
seq. # A
header
Fixed ATM Network
COS assigned to B
AP 1
AP 2
10Mbit.s -1
2.4GHz
10m range
B
Host with GPS receiver and 1MHz clock
timestamp B
seq. # B
timestamp A
seq. # A
header

Fig. 9. Experimental arrangement used to measure cells' end-to-end delay and loss rates during execution of our handover procedure.

cording to B's local clock) and then stored in memory for later examination. Other information, such as which AP transmitted the cell and how many times it was repeated are also recorded. The 1pps pulses of the GPS receivers used are synchronised to UTC with a nominal accuracy of 1μs, so the resolution with which we can measure the end-to-end delay incurred by a cell is limited by the precision of the 1 MHz oscillators in each computer. These have have a tolerance of 10 parts per million, so we would expect a maximum error of 20μs in the measured end-to-end delay values. Since at 10Mbit.s^{-1} this is a fraction of a cell transmission time, we consider this system's resolution to be adequate for analysing handover behaviour. Thus we can later reconstruct and analyse the pattern of cell loss and delay due to handover events.

In order to observe the behaviour of the protocol more easily, a slight modification is introduced into the scheme. Normally the COS would immediately issue a Handover-Ready message to the WT once the connection segments to the new AP have been set up (step (7) in Figure 2). For measurement purposes, the COS delays sending this message and switching the data path until it has seen a data cell with a particular sequence number. In the trace shown in Figure 10, the COS was allowed to send the message only after seeing a cell with a sequence number divisible by 3000. This allows us to be certain which effects are due to the handover protocol and which are brought about by interference in the channel. It does not, however, affect any of the measurements taken.

A. Data

Figure 10 shows the effect on cell transport of a number of handovers back and forth between two APs. In this case, handover is triggered every 3000 cells and it is interesting to note that since the end-to-end delay via the first AP is slightly different from that via the second, the handover events are easily observed. It may be seen that, handovers aside, there is some jitter in the cells' delay. This is predominantly due to the ARQ scheme which is employed to reduce the cell loss rate due to the error characteristics of the wireless link. Each repeat adds approximately 100μs to the delay of a cell and if a cell is repeated a great many times it may also delay its successor. This accounts for the small scale CDV in the delay plot and also for the occasional residual points such as those at about cell 900 which experience greater delays.

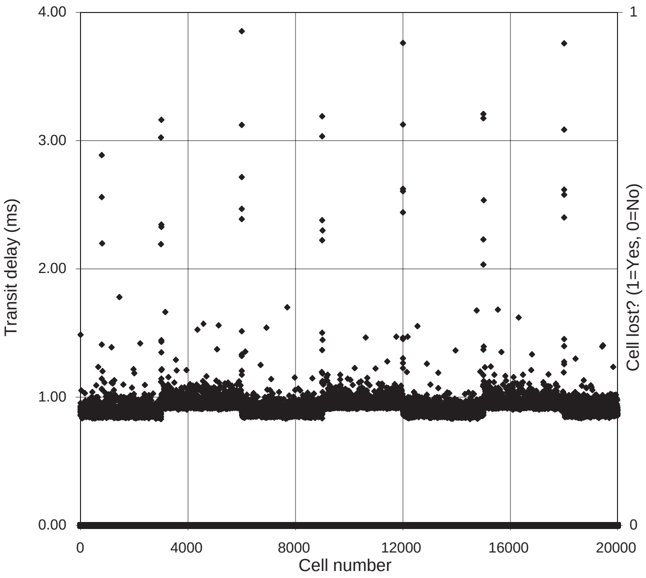

Fig. 10. Graph illustrating several handovers, occurring every 3000 cells. The upper line represents (on the left-hand axis) the measured end-to-end delay experienced by a cell; the lower line indicates (on the right-hand axis) whether each cell is lost or not. The data rate was 1.5Mbit.s^{-1}.

Referring to the data plotted against the right-hand axis in the figure, it is particularly important point to note that no cells are lost during handover. Figure 11 shows one handover in detail and it can be seen that only about 10 cells have any increase in delay due to the handover. The cell rate is 3000Hz so this corresponds to a period of disruption of approximately 3ms. The delay incurred by these 10 or

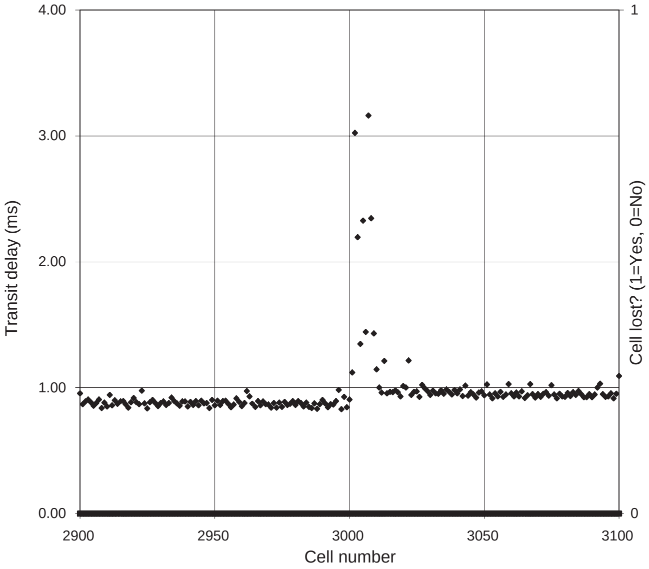

Fig. 11. Close-up of a section of figure 10.

so cells is due to being buffered in the new AP while the WT acknowledged receipt of the HandoverReady message, performed its radio handover and executed the necessary meta-signalling exchange with the new AP.

Figures 12 and 13 show a set of probability density functions of the end-to-end delay of cells. There are two features to note. Firstly, the main body of the pdf is less nearly vertical than would be expected for a wired network; secondly there is a shoulder where the delay sharply increases. Both of these features correspond to effects we discussed above; the slope of the body of the function is due to the small scale CDV effect of the ARQ mechanism and the shoulder is caused by the handovers.

It may be seen that even with the WT changing AP eight times per second, the delay at the ninety-ninth percentile is only a little more than 1.5ms greater than the median delay. This reflects the fact that our procedure causes very little disruption to cell transfer.

When a particular WT performs sixteen handovers per second, the time taken for the signalling system to set up the new connection segments exceeds the period between handovers. The speed of the signalling system places a fundamental upper limit on the maximum per-WT frequency of handover achievable by all handover schemes except for the virtual connection tree approach. We believe that the latter is the only scheme proposed to date that remains viable for networks with per-WT handover rates in excess of approximately 10Hz.

B. *Interpretation*

Our procedure introduces very little cell delay variation even at handover rates which are unlikely to occur in a real-world situation. The amount of play-out buffering required for CBR streams to remove this jitter is on the order of a maximum of 10ms; even 1ms of buffering would imply that at eight handovers per second, only just over 1 per cent of cells would arrive too late to be usable.

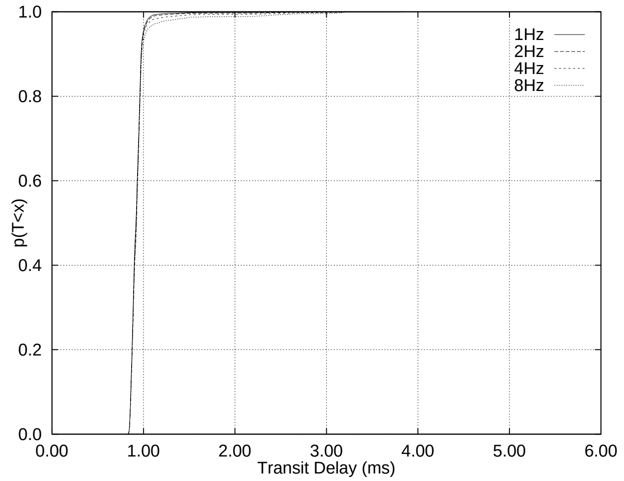

Fig. 12. Measured probability density functions for end-to-end delay of cells at a data rate of 1.5Mbit.s^{-1} with handovers occurring at 1, 2, 4 and 8Hz.

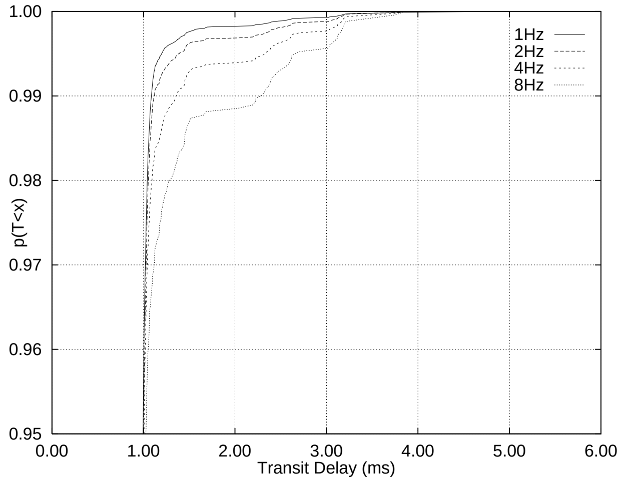

Fig. 13. Close-up of a section of Figure 12.

We are able to guarantee that cells are never duplicated or re-ordered but a side effect is that is not possible to guarantee zero cell loss. Consider a cell which has been transmitted by the old AP and received correctly by the WT. If the acknowledgement is not received by the AP, the two entities temporarily have an inconsistent view of which cells have been successfully transmitted. If handover occurs at this point, the AP must decide what to do with this cell—it can discard it, assuming (correctly, in this case) that the acknowledgement was lost. Some authors [4], [14] propose that the cell should be forwarded to the new AP and retransmitted. It can be seen that in the scenario described, this will lead to a duplicate of the cell being received. We deem that modifying the standard cell header used in the wired network to include a sequence number to allow the WT to detect this duplication is not an acceptable solution and therefore choose to discard such cells. In the converse situation where the cell had not been received correctly by the WT, we will therefore lose a cell as a direct result of handover. As evidenced by the lack of such an event in the data above, such an event will be rare enough that it will have a negligible effect on the CLR of a WT's VCs.

The very small effect on both CLR and CDV introduced by our handover procedure makes it suitable for both delay-sensitive and loss-sensitive traffic.

IX. Concluding Remarks

In this paper, we described a handover protocol which we believe meets the requirements detailed in Section II. The novel use of path rerouting from a static COS enables a low latency handover which is scalable to many WTs with many VCs. The use of a two-phase backward handover which decouples the rerouting of a WT's connections from the radio handover minimises the interruption to data transport. The handover protocol requires no modifications to existing ATM signalling or routing protocols. This means that an existing wired ATM network can be mobility-enabled simply by updating the software in one, or a limited number of switches. Furthermore, applications on both wireless terminals and fixed terminals require no changes to operate with a wireless ATM network. We believe that this is a major benefit in heterogeneous systems and will speed uptake of wireless ATM systems. We introduced the concept of the use of user-plane connections for the low-latency transport of handover messages between the mobility-supporting software entities. As well as obviating the need for modified signalling protocols, this is significantly faster as the overhead and delay of control plane signalling processing at each hop is avoided. We have summarised five other handover schemes from the literature and examined the similarities and differences between those schemes and the scheme we have proposed in this paper. Finally, we presented quantitative results from an implementation of this protocol on a first-generation wireless ATM LAN. The results indicate that the protocol is well-suited to supporting both realtime and non-realtime traffic.

Acknowledgements

The authors would like to thank all their colleagues within the ORL wireless ATM group for their contributions to the project described here and their valuable comments during the preparation of this work.

References

[1] J. Porter, A. Hopper, D. Gilmurray, O. Mason, J. Naylon, and A. Jones, "The ORL radio ATM system, architecture and implementation," Tech. Rep. 96-5, ORL, 24a Trumpington Street, Cambridge CB2 1QA, England, 1996, available on-line from http://www.orl.co.uk/radio.

[2] A. S. Acampora and M. Naghshineh, "An architecture and methodology for mobile-executed handoff in cellular ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 8, pp. 1365–1375, Oct. 1994.

[3] R. Yuan, S. Biswas, and D. Raychaudhuri, "Mobility support in a wireless ATM network," in *Proceedings of the WINLAB Workshop*, New Brunswick, NJ, 1995.

[4] K. Y. Eng, M. J. Karol, M. Veeraraghavan, E. Ayanoglu, C. B. Woodworth, P. Pancha, and R. A. Valenzuela, "A wireless broadband ad-hoc ATM network," *Wireless Networks*, vol. 1, no. 2, pp. 161–174, July 1995.

[5] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M. B. Srivastava, and J. A. Trotter, "SWAN - a mobile multimedia wireless network," *IEEE Personal Communications*, vol. 3, no. 2, pp. 18–33, 1996.

[6] B. A. Akyol and D. C. Cox, "Rerouting for handoff in a wireless ATM network," *IEEE Personal Communications*, vol. 3, no. 5, pp. 26–33, Oct. 1996.

[7] A. Acharya, S. Biswas, L. French, J. Li, and D. Raychaudhuri, "Handoff and location management in mobile ATM networks," in *Proceedings of the Third Internation Conference on Mobile Multimedia Communications (MoMuC-3)*, Sept. 1996.

[8] D. Niehaus, A. Battou, A. McFarland, B. Decina, H. Dardy, V. Sirkay, and B. Edwards, "Performance benchmarking of signaling in ATM networks," *IEEE Communications Magazine*, vol. 35, no. 8, pp. 134–143, 1997.

[9] The ATM Forum Technical Committee, "ATM user-network interface (UNI) signalling specification version 4.0," June 1996, document number af-sig-0061.000.

[10] The ATM Forum Technical Committee, "ATM private network-node interface (PNNI) version 1.0," Mar. 1996, document number af-pnni-0055.000.

[11] J. Porter and D. Gilmurray, "Tunnelled signalling for the support of mobile ATM," *ATM Forum contribution number 96-1699*, Dec. 1996, available on-line from http://www.orl.co.uk/radio.

[12] R. Steele, Ed., *Mobile Radio Communications*, chapter 8, pp. 677–765, Pentech Press, 1992.

[13] B. A. Akyol and D. C. Cox, "Signaling alternatives in a wireless ATM network," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 1, pp. 35–49, Jan. 1997.

[14] H. Mitts, H. Hansén, J. Immonen, and S. Veikkolainen, "Lossless handover for wireless ATM," *Mobile Networks and Applications*, vol. 1, no. 3, Dec. 1996.

John Naylon received the B.A. degree in computer science in 1994 from the University of Cambridge, Cambridge, U.K., where he is now working toward the Ph.D. degree.

He has been working with the Olivetti & Oracle Research Laboratory (ORL) wireless ATM group for the past three years, investigating QoS issues and the implementation of handover and location management schemes.

Damian Gilmurray received the B.A. degree in computer science from the University of Dublin, Dublin, Ireland in 1990.

Since 1991 he has been a research engineer at the Olivetti & Oracle Research Laboratory (ORL) where he is currently a member of the wireless ATM group. His research interests include networking, mobility and distributed systems.

John Porter received the B.A. degree in computer science in 1983 from the University of Cambridge, Cambridge, UK.

He has worked with the Olivetti & Oracle Research Laboratory (ORL) since 1987 developing wired and wireless ATM networks. Since 1992 he has been running a project which is developing a high speed broadband radio for mobile and fixed operation.

Andy Hopper received the B.Sc. degree from the University of Wales, U.K. in 1974 and the Ph.D. degree from the University of Cambridge, Cambridge, U.K. in 1978. He was elected a Fellow of the Royal Academy of Engineering in 1996.

He has recently been appointed Professor of Communications Engineering at the University of Cambridge Department of Engineering and is a Fellow of Corpus Christi College. He is Vice President of Research of Ing. C. Olivetti & C. SpA, Italy, Director of the Olivetti & Oracle Research Laboratory (ORL), Chief Technical Officer of Advanced Telecommunications Modules Ltd., Chairman of Telemedia Systems Ltd. and a Director of Acorn Computer Group plc. His research interests include networking, multimedia, and mobile systems.