Architectural Imperatives for 4th Generation IP-based Mobile Networks

Donal O'Mahony & Linda Doyle

Networks & Telecommunications Research Group
Trinity College,Dublin 2, Ireland

Donal.OMahony@cs.tcd.ie,
Linda.Doyle@tcd.ie,
http://ntrg.cs.tcd.ie

Abstract

There are many different views on what will be the form of 4th generation fixed and mobile systems. We argue that the generation change offers an opportunity to forge a new network architecture based on delivering services to large populations of mobile users through a single IP-based infrastructure. We outline some of the forces driving the development of this 4th generation architecture, as well as some of the research challenges and describe a number of projects that we are undertaking to advance the state-of-the-art.

Key words

4th Generation Mobile, Wireless IP, Ad-hoc Networking.

1. Introduction

The architecture of both the public switched telephone network and also the first three generations of mobile telephone network [1] were shaped by two main considerations. Firstly, the networks were required to effectively provide a single service – circuit switched voice communications – across a network with low-capability edge nodes and a large amount of intelligence within the network. Any additional services that could be deployed on this infrastructure were looked upon as a bonus. Secondly, they required the operators to be able to charge for network usage based on the destination and duration of the call. Mobility between network operators was only worth doing if the home network operator stood to gain financially from this possibility.

The evolution of the Internet took a different path. In the early years, it was assumed that communities of users collaborated for the benefit of all. Since no organization wanted to act as a central network operator, the interior of the network was kept as simple as possible while allowing users to offer such services as they wished at the edges. The possibility of users moving around was catered for at the edges by Mobile IP and was implemented by agents external to the network itself, or at least located within edge routers.

In recent years, we have seen the beginning of convergence between the telecommunications and Internet communities. The PSTN has become the main access network for residential users to connect to the Internet and there is a huge industry focus on re-engineering the channel structure of wireless telephony networks to accommodate data streams. In the other camp, the Internet telephony industry has proved that it is possible to build a production voice service on a packet-switched internet-based network infrastructure, and the makers of large circuit switches have conceded that there are huge economic advantages to shifting over to packet switching.

The rapid adoption of mobile telephony, most notably in the Scandinavian countries, shows that there is considerable demand from users for services delivered on mobile devices.

In the midst of the great changes being brought about by the trends outlined above, the research community is considering what form the next (4th) generation of fixed and mobile communications systems will take. A popular view in the telecommunications industry is that the 4th generation should evolve naturally from 2nd and 3rd with incremental improvements being brought about without any fundamental architectural changes being necessary.

2. Drivers for the 4th Generation Architecture

We take the view that the forthcoming 4th generation mobile systems offers us an opportunity to take a fresh approach in designing a network driven by services that people want now, and in the future as opposed to services that they wanted in the past. In the following sections, we examine some of the requirements underlying our concept of 4th Generation mobile systems.

2.1. Support for IP-based Traffic

It is clear from the rapid growth of the Internet in the late 1990s that IP forms an effective delivery mechanism for the kinds of network services that users are interested in availing of. The advent of Voice-over-IP has shown that person-to-person audio communication can easily be carried out across a packet-based IP network in spite of some difficulties in delivering consistently low end-to-end delays across the current infrastructure. This fact, when taken together with very low forecast growth rates for voice as compared with data traffic, point to a future where voice makes up a very small proportion of the traffic being carried. It is imperative then that the architecture of 4th Generation networks is determined primarily be the need to deliver a good IP service. The ability to handle voice and other streams with real-time constraints is a secondary (but essential) goal.

2.2. Excellent Mobility Support

The adoption of mobile telephony services offered by 1st and 2nd generation voice systems has surpassed many peoples expectations. In many countries, most notably in the Scandanavian region, more than 60% of the overall population make use of mobile telephones and the trend in other countries is to follow the same path. It is forecast that there will be

almost 1 billion subscribers to GSM, the most successful 2nd generation system, by 2005.

The mobility facilitated by cellular telephones is often referred to as *terminal mobility*. An additional form of mobility envisaged for both the fixed and mobile phone networks is *personal mobility,* which allows a person to move from one terminal to another and have his calls and 'environment' follow him as he *registers* on different mobile or fixed terminals. The *Universal Personal Telephony* service allows a user to be contactable via one personal number. Intelligence within the network, coupled with information on the users preferences and likely movement patterns can forward the call to the appropriate terminal or to an automated response system.
In 4th generation systems, we must assume that all users of the network are potentially mobile with a sizeable proportion of them communicating via wireless terminals. Users must be contactable via a single (albeit multi-faceted) identity. A means must be found to map from this identity to an address to which packets can be routed. Control of this mapping must lie firmly with the user who can modify the destination of the mapping and regulate the access that callers may have to it. In an environment where the path from source to destination may cross many different network domains, it would be unwise to associate this mapping with a single network operator. It is imperative that 4th Generation networks provide a single consistent means of identifying users and allow this identity to be efficiently mapped to a routable destination.

Where 4th generation nodes are actually in motion, they may need to change their point of attachment to the network while a flow of packets is in progress. The degree to which this changes the path from source to destination depends on the topology of the network and the route chosen. 2nd and 3rd generation cellular systems allow handoff where a mobile node shifts its point of attachment but stays within the same network domain. Typically, a change in domain is referred to as *roaming* and active connections cannot be maintained under these circumstances. This implicit two level tracking of a users location (i.e. current domain and current position with that domain) can be generalized into a N-level mobility tracking scheme which can allow handoff in any circumstances where the network topology renders it possible.

2.3. Support for Many different Wireless Technologies

1st, 2nd and 3rd generation mobile systems relied on the use of radio spectrum that was reserved for public land mobile use and licensed for use by a very small number of network operators in each country. Differences in allocation timing and strategy of the spectrum have led to the need for multi-mode phones capable of adapting to use the band of spectrum available in a particular region.
In 4th generation systems, it is likely that a potentially large number of different radio technologies may be used for network access. Technologies operating in the ISM band such as Bluetooth, IEEE 802.11and other *Short Range Wireless* [SRW] links may be very important in providing high-speed network access in built-up areas such as shopping malls and train stations. Similarly, satellite systems may be useful at sea, in the air, or in areas with very low population density. Existing 2nd and 3rd generation cellular may be useful in between these two extremes. A 4th generation node should be capable of adapting its radio transmission and reception capabilities to take maximum advantage of available spectrum.

2.4. Free from Unnecessary Operator Linkage

The GSM system was developed primarily by European telecommunications companies in mid- to late 1980s primarily as a mobile extension to the PSTN. The GSM model envisaged that users would subscribe to a operator who would build a cellular network infrastructure that would track the user as he moved from one location to another making every effort to maximize the availability of service. In common with the dominant PSTN billing model in Europe at the time, all usage of GSM services was to be metered and since the only way to pay for this was via the 'home' operator, every action carried out by the mobile handset is with reference to the network operator. Even when a user roams into a new domain, contact is made with the home network to establish a link with a billable entity before any calls can be made.

Two GSM handset's cannot communicate with each other directly, rather they must each first authenticate themselves to the network, be linked with their billing details and thereafter, operator mediated communication can take place. This mode of operation is consistent with the fact that the operator in effect *owns* the spectrum and is entitled to individually regulate and meter each access to it.

In an environment where spectrum ownership is much more open – such as is the case in the ISM band, such restrictions are completely unnecessary. Much of the dynamism that has taken place in wireless communication in recent years is as a result of the easy access to such spectrum and it is likely that this trend will continue in the future.

Where access to spectrum is not an issue, pairs or groups of nodes can form ad-hoc networks to allow direct communication between nodes and if appropriate, nodes can collaborate, relaying each other's traffic. Naturally issues such as the need for user authentication occur in this kind of any-to-any direct communication, but arguably, these need to be solved in an operator-independent way in any event.

Once the special position of the operator is removed, there remains the problem that a wireless node wishing to communicate with another node not within its range cannot do so unless an intermediate node relays their packets to either another wireless node or on to the fixed network.. If we could find a means of making real-time payments across a link, this would both relieve us of the need to be associated with a billable entity and also allow us to motive individual nodes to co-operate with us to relay traffic.

This motivation (financial or otherwise) could be used in a sparsely populated area to allow an individual wireless node to act as a packet relay between two out of range nodes. The payment method would allow the relaying node to be

compensated for the drain on its batteries and the usage of bandwidth that might otherwise be available to it.

In a heavily populated area, the same motivation could be used to encourage organizations to erect networks of wireless network access points in places like university campuses and shopping malls. Organizations that undertook this effort to any great scale would become the network operators of the 4th generation, but the fact that no special status is required to become an operator should ensure healthy competition.

2.5. Support for End-to-End Security

The security features inherent in 2nd and 3rd generation mobile systems are focused on two main services. Firstly, the mobile users must be authenticated to the network operator. This authentication is generally limited to associating the user with a billing relationship that is operating satisfactorily. Where accounts are pre-paid, this billing relationship often has no stored details on the identity of the user. There is no end-to-end exchange of credentials between a mobile user and their peer at the other end of the link.

The second service supported by 2nd and 3rd generation mobiles is the content encryption of information sent over the wireless path. While this does deter atacks using simple scanning devices, it is no substitute for genuine end-to-end confidentiality.

In the 4th generation, mobile and fixed nodes will interact with each other without reference to their relationship to an operator. It is imperative that protocols and procedures are devised to allow the users of these nodes to authenticate one or more facets of each others identity to a degree necessary to achieve the communication they desire.

3. 4th Generation Architecture and Research Issues

We envisage that the 4th generation mobile networks of the future will be based around an IP core network which will, over time, completely displace both the fixed PSTN and also the 2nd and 3rd generation mobile voice networks. The architecture will be based on delivering an IP transport service to a population where every user is potentially mobile and a very large proportion of the them make frequent use of wireless nodes to interact with people and services on the network.

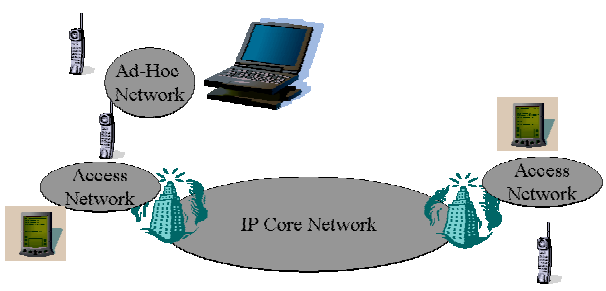

Figure 1:Architecture of 4th Generation Fixed and Mobile Network

Un-tethered users will gain access to the fixed network using a variety of different radio technologies. These will include many recent short range wireless systems that make use of the ISM band of spectrum such as Bluetooth and 802.11 as well as perhaps new access methods made available on the spectrum currently in use by 2nd or 3rd generation mobile voice systems. Individuals and organization will install and operate wireless access points making use of a real-time payment system to provide such quasi- network operators with their financial incentive.

High capability nodes will be equipped with software radios which carry out as many radio functions as possible by executing signal processing software on general purpose processors. Thus, the nodes will be capable of communicating using whichever radio technology is expedient given it's location and financial resources.

On being powered up, a 4th generation wireless node will attempt to establish communication links with its near neighbours. The radio spectrum involved and the modulation method used will be determined by scanning already extablished communications links and by probing the channels characteristics.

Ad-hoc network protocols will be used to bind together groups of nodes for local communication and also to provide a link between each node and a range of geographically close fixed network access points. The motivation for this collaborative behavior may derive from the fact that the users have authenticated each other as belonging to the same group (e.g. workers in the same office building or paramedics at the scene of an accident) or some kind of micro-payment method may be used to allow relaying nodes and network access points to profit from their activities.

Interactive links to other users will be established based on a user identity. In order to support both terminal and personal mobility, a directory service will be required to perform the mapping between an identity or identity facet and an address to which packets can be routed. This service should be standalone and independent of any particular network infrastructure. The mapping may also be quite complex allowing a user to build a profile of when and to whom he is available. In cases where confidentiality of location may be an issue, artificial relay points may be used to disguise the position of a node from the caller.

Before a call can be setup, each participant will need to authenticate themselves to the other parties within the call. Typically, users will have a multi-faceted identity and it may be inappropriate to make use of all of these facets for a any given call. Facets may also depend on one another and progressive authentication may take place. For example, to gain entry to an office building, it may only be necessary to authenticate the fact that one is an employee. If an individual wants to open a safe containing financial documents, it may be necessary to also authenticate the fact that the individuals organizational role is financial director. Such identity facets

would be inappropriate for use in a coffee shop where the same user wished to traverse a high-speed fixed network access point.

Once the parties have authenticated each other, end-to-end communication can take place with appropriate authentication and confidentiality measures being applied to the data traffic.

Where pre-built access infrastructure does not exist, or where communication is taking place in a local region, nodes may resort to the formation of ad-hoc networks. While this greatly enhances a nodes ability to communicate, it also raises a number of additional research issues. Collaboration is the essence of ad-hoc networking and will naturally take place where the members of the network belong to the same organization or have some other pre-established motivation to work together. In a public context, for instance a number of wireless nodes encountering each other in a shopping mall, the motivation may be less obvious. There is a need for security protocols to allow nodes to selectively reveal information about different facets of their identity with an aim of maximizing the level of cooperation that is possible. If nodes are to offer relaying services to others, there will be a need for some kind of real-time payment method to allow a node to be compensated for the use of its resources (battery power etc).

Once an ad-hoc network grows beyond a small number of nodes, there is a need to develop sophisticated routing protocols to allow nodes to relay for each other thus maximizing their inter-connectivity. It should also be possible for a wireless node that is within range of a fixed network access point to relay for other nodes within it's ad-hoc network.

Applications and services in the 4th generation will be built outside the network in keeping with the Internet tradition of keeping the core simple, fast and efficient. Today, many of the Internet applications rely on entities contacting each other based on a network address. This is problematic for a number of reasons. Firstly, it leads to a demand by end entities for the fixed assignment of addresses. If the address space is finite or inefficiently allocated (as is the case with IP version 4) this leads to address shortages and also causes problems when these end-entities move with respect to the network. We envisage that in the future, applications will evolve to a situation where addresses are de-emphasized in favour of the use of more abstract names. Mobile users will be contacted by name, with appropriate servers to perform the mapping between address and the necessary routing information. Clearly such a mapping must be done in such a way that the user has control over what kind of routing information is held and who may gain access to it. Where a user is in motion appropriate mechanisms must be devised to maintain a connection to a 'name' as the route to the node is changing.

4. The NTRG 4th Generation Testbed

At Trinity College, the Networks and Telecommunications Research Group(NTRG) has been investigating the form of 4th generation mobile systems since 1998. We have ongoing projects investigating many different aspects of 4th generation technology from applications though to physical layer issues. The individual projects are bound together by virtue of their individual contributions towards our 4th Generation Testbed.

Since we envisage that the 4th generation mobile nodes of the future will be constructed using commodity computing platforms, our target nodes are general purpose PC workstations, and where portability is important, laptop and palmtop variations of these. In order to keep a consistent operating environment across all platforms and to allow our work to integrate with popularly available applications, we have chosen to develop to the Microsoft Win32 environment as supported by Windows 2000 on PC and laptop and making use of Window CE on handheld, and palmtop environments.

4.1. The Layered Architecture

Components of our 4th generation environment are implemented as standalone *layers* each realized by a single main *thread*. The inter-layer interface is very simple, consisting of primitives to send information upwards or downwards through the stack and to attach a *blackboard* of parameters to each request that can be used by any layer through which the data passes.

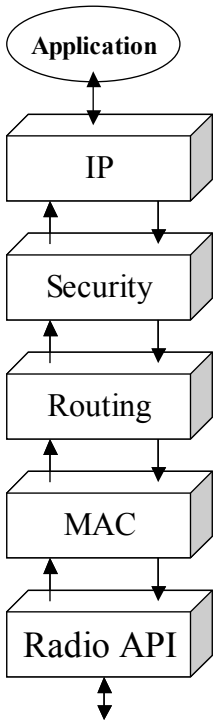

Figure 2: The Layered Structure of the 4th Generation Testbed

4.2. Wireless Alternatives

Different radio hardware can be accommodated by writing a layer that interacts directly with the hardware and presents the simple inter-layer transfer interface to whatever is above it. In this way, we have been able to

perform our wireless experiments with links as diverse as Infra-Red (IRDA), Bluetooth, IEEE 802.11 and also a very simple half-duplex radio we have constructed in-house which uses amateur radio frequencies and allows us to experiment at a very low level.

4.3. Software Radio

Ultimately, we expect to be able to replace the *real* radios at the bottom of this protocol stack with a software radio operating in conjunction with a wide-band front-end which would allow operation across a wide frequency band with the chosen form of modulation being performed in software. The individual building blocks of the software radio (such as channelization, coding, modulation, de-modulation etc) will be realized by individual layers with an appropriate *stack* being assembled to deliver the desired radio waveform. Thus far, we have receiver-only systems that implement a variety of forms of amplitude and frequency de-modulation and can rapidly switch between the two if necessary.

4.4. Routing Protocols

The subject of routing protocols enabling the formation of ad-hoc networks of nodes has been under active study by the research community[ADHOC] for some time now. Although many different protocols have been simulated, very few actual implementations exist which can test these protocols across a real radio channel under different mobility scenarios.

Using our layered architecture, we initially implemented the Dynamic Source Routing (DSR) protocol. This simple reactive protocol only begins to try to establish a route to a destination when there is data to be sent. This is in contrast to a proactive protocol which attempts to maintain knowledge of the state of the network so that it is already in possession of sufficient routing information when data needs to be sent. We are now in the process of implementing a hybrid Zone Routing Protocol (ZRP) which is proactive for nodes that are in a zone *near* and reactive for those that are further away. We anticipate that the use of these protocols across real wireless channels will give us a unique insight into their properties.

One are of active study is the integration of our ad-hoc islands with the fixed core internet.?????????????

4.5. Emulation Facilities

When routing protocols or mobility aware applications are being developed and tested, one of the major problems is to expose the evolving software to particular mobility scenarios without having to conduct all debugging on the move and out of doors. Our initial approach to this challenge was to develop a layer (which we call the *datagram* layer) that emulates the radio broadcast environment across a collection of internet links on the local area network.

Each of the emulated nodes is assigned an IP address and the emulation layer in each node is *told* what other nodes are visible to it in radio terms. When a packet arrives to be send on the emulated radio interface, it is encapsulated in an IP datagram and sent to each of the visible nodes.

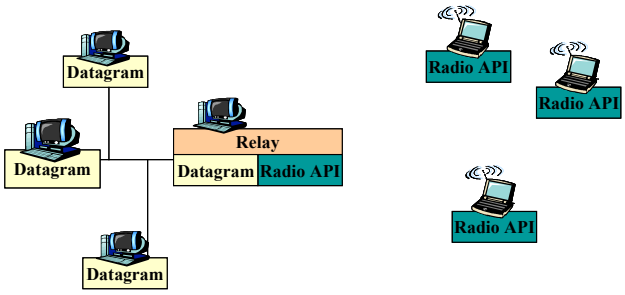

Figure 3: An Ad-Hoc Network of Genuine and Emulated Wireless Nodes

This effectively allows us to construct ad-hoc networks made up of processes running on any Internet-connected host. By constructing a *relay* layer that moves packets from one protocol stack to another, we can freely intermix nodes that are running on real wireless nodes and others that are sitting on the emulated radio layers.

While the above is an effective debug and test tool, emulated wireless nodes either see each other or they do not, and transmissions always get through without transmission errors. The ability of nodes to *see each other* or not is also statically configured.

In an attempt to improve our radio emulation environment, we have devised a system that we refer to as a *reality emulator*. In place of the radio layer in each node's protocol stack, we place a reality emulator client layer. Each of these clients connects via sockets to a server which emulates the way in which a radio channel will behave as nodes move with respect to one another and also encounter collisions in transmission.

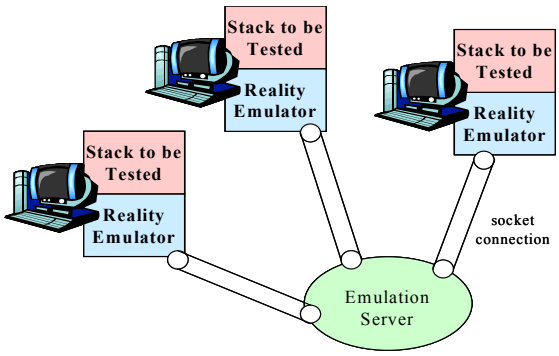

Figure 4: The Reality Emulator

When a node initializes itself, it connects to the emulation server where it is given an initial geographic position. By interacting with a graphical user interface on the server, a designer can control the transmission range of each radio and their movement with respect to one another.

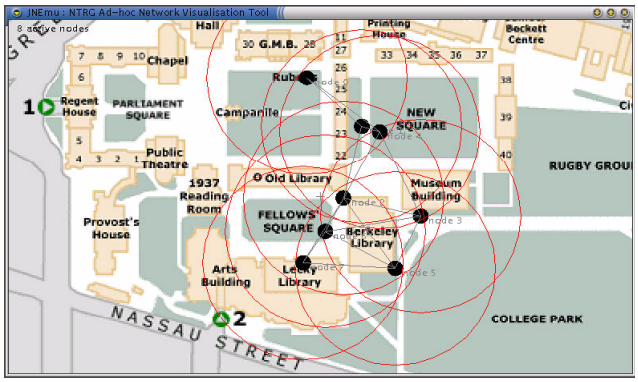

Figure 4: The Emulation Server showing 8 nodes and their radio ranges

From the point of view of the client protocol stack, the behaviour of the link is identical to that which would be experienced by the node when running across a real radio link and provided the number of emulated nodes is kept reasonably small, the emulator can support substantial amounts of traffic including real-time voice streams. The system is particularly good at exposing the routing software to particular node movement scenarios that might otherwise be hard to achieve.

4.6. The Security Architecture

We envisage that all users in the 4th generation are potentially mobile, and a large proportion of them will avail of wireless devices. This means that nodes will typically be interacting with peer nodes with little information on the identity represented by that peer.

Authentication in the physical world is typically achieved by resorting to a multiplicity of cards that people carry around in their purse or wallet. These may contain basic information about a person, such as might be present on a driving licence. Other cards may contain information on an individuals affiliation with a bank or perhaps more personal information such as a card detailing their blood type or information on a medical condition.

Individuals typically produce cards detailing different facets of their identity only as the need arises. For example, to make a cash withdrawal at a bank, they may need to produce both a driving licence and a bank card. They might only hand over details relating on their medical record to someone who had already proved they were a medical professional.

The above exchanges are characterized by individuals entering into a negotiation dialogue where credentials are exchanged in order of increasing importance as trust is progressively built between the individuals. This process concludes either when the shared trust has reached it's highest level, or it has exceeded that required by the communication exchange.

At present, we are designing a system to do this kind of credential exchange between nodes in an ad-hoc network. When a node wishes to avail of a service on another system, this will cause the credential exchange agents to enter into a dialogue where details on different facets of a users identity are exchanged according to a Credential Release Policy (CRP).

Nodes may authenticate neighbouring nodes up to a level in which they are happy to relay traffic through them. A considerably higher level of authentication may be needed before a pair of nodes may be willing to enter into a specified application dialogue.

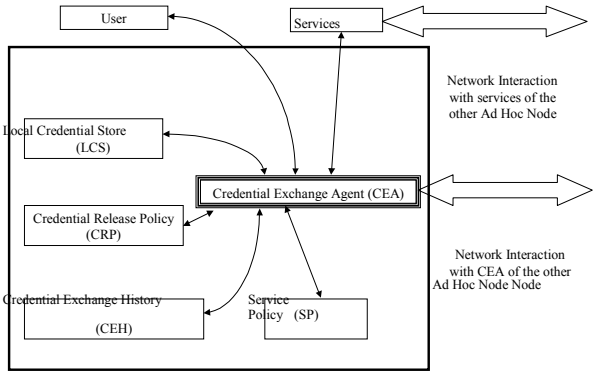

Figure 5 The 4th Generation Security Agent

Based on the authenticated identity facets, which are shared between the two nodes involved in the trust negotiation, groups can be formed to share a common purpose. This will allow shared keys to be negotiated which can be given to all member of the group to allow shared access to resources. Similar group formation systems have been developed for internet multicast environments that can serve as a model for this[VERSAKEY]. This can be used for such applications as a walkie-talkie type system where all users hear all traffic or indeed any other form of data-based group collaboration.

4.7. Real-Time Payment

Most of the flaws in 2nd and 3rd generation mobile systems can be traced to the fact that the major technical decisions defining the architecture have been made based on one overriding concern, namely that of generating as much revenue as possible for the network operators.

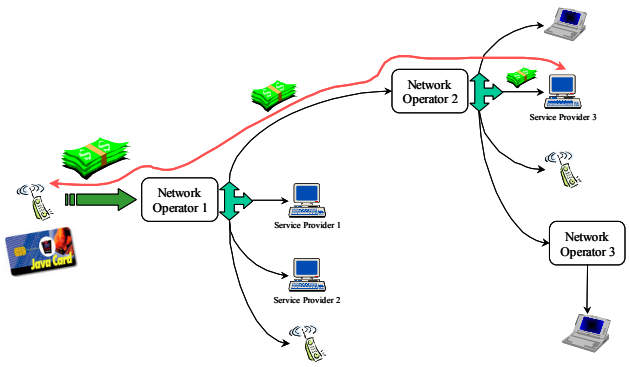

Figure 6: Real-time, Multi-party Micropayment

If the promise of the 4th generation is to be fulfilled, the focus must be on enabling connectivity between network users without tieing this connectivity to a user-operator subscription. Clearly some other way needs to be found to allow providers of network infrastructure (even if this is just a single relay node) to be rewarded for making this available to the general population of nodes.

Figure 6 depicts a multi-party micropayment system [PEIRCE] developed within our research group that allows a stream of cryptographic payment tokens to be travel interspersed with the normal data packets that make up the end-to-end flow. Before this can commence, a pricing phase was undertaken where each party (or network operator) along the path offered relay services for a particular price. Once the contract is agreed to by all, the payment tokens can be captured by nodes along the path and redeemed for an agreed proportion of the overall payment for the traffic. We plan to adapt this system to support ad-hoc operation where the end-to-end route may change over the lifetime of the communication.

4.8. Applications

The principle application we have used on our 4th generation network test bed have been simple web access where pages containing HTML as well as complex multimedia data have been delivered to hand-helds moving in a wireless network. We have also used the performed some experiments using the wireless link for video image data[ANIL]. More recently, we have developed a simple point-to-point telephony application using the Session Initiation Protocol (SIP) as the signaling protocol. In the future, we will modify this application to integrate elements of our security architecture, in particular, to incorporate the notion of multi-facetted user identities. There are a whole host of issues to be resolved before we can support mobility thorough the mapping of this identity information to addressing information that works well with both fixed and ad-hoc networks.

5. Conclusions

We have outlined above some of the problems inherent in the architectural design of 2nd and 3rd generation mobile systems. Arising from this, we have enumerated some of what we believe are imperatives for the design of a new 4th generation architecture. In order to progress our ideas for 4th generation systems, our research group has embarked on a number of distinct projects dealing with different aspects of the overall system. Each of these projects contributes to the construction of a test bed based on the use of commodity PC and PDA hardware running common operating systems. The test bed supports wireless mobility via a range of technologies from

Infrared to software radio and allows the experimentation with real ad-hoc networks that are engineering to integrate with populations of fixed nodes. Security concerns are also catered for both in terms of node authentication and also as a means of payment for resources consumed. The utility of the 4th generation systems will be demonstrated with non-interactive data-based applications as well as those, such as telephony, that involve continuous multi-media streams.

6. References

[1] O'Mahony, D., Universal Mobile Telecommunications Systems: The Fusion of Fixed and Mobile Networks, *IEEE Internet Computing*, 2(1), 49-56, January/February, 1998

[2] [SRW] Leeper, D.G., A Long-Term View of Short-Range Wireless, *IEEE Computer*, 34(6), 39-44, June, 2001

[3] [ADHOC]Royer Elizabeth, Toh C-K, "A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks ," *IEEE Personal Communications Magazine*, April 1999, pp. 46-55.

[4] [VERSAKEY] Waldvogel M., Caronni G., Sun D., Weiler N., Plattner B., The VersaKey Framework: Versatile Group Key Management, *IEEE Journal on Selected Areas in Communications*, Special Issue on Middleware, 17(8), August 1999

[5] Peirce, M. & O'Mahony, D., Flexible Real-Time Payment Methods for Mobile Communications, *IEEE Personal Communications*, (6) 6, 44-55, December 1999

[6] [ANIL] Kokaram, A., Doyle, L. & O'Mahony, D., *Error-resilience in Multimedia Applications over Ad-hoc networks*, Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2001, May 7-11, 2001 - Salt Lake City, Utah