The Architecture of the BRAIN Network Layer

Robert Hancock[1], Hamid Aghvami[2], Markku Kojo[3], Mika Liljeberg[4]

[1] Siemens/Roke Manor Research Ltd., Romsey, U.K., email: robert.hancock@roke.co.uk
[2] King's College, London, U.K., email: hamid.aghvami@kcl.ac.uk
[3] University of Helsinki, Finland, email: markku.kojo@cs.helsinki.fi
[4] Nokia Research Center, Finland, email: mika.liljeberg@nokia.com

Abstract: *This paper describes the architecture of the network layer used in terminals and the access network of the BRAIN project. It describes the design principles that have been applied, the basic access network structure that results, and the way in which this structure fits with other components of the complete mobile system.*

Introduction

The IST project BRAIN (see [1]) is a wide ranging research activity to develop an IP-based mobile wireless network complementary to current 2nd and 3rd generation systems. The initial focus is customer premises applications evolving from WLAN systems; however, it extends naturally to public metropolitan networks as the demand for broadband multimedia increases, and thus is a first step beyond 3G networks. The project encompasses user applications, through middleware, all the way to the air interface; the focus of this paper is the network layer architecture which supports and unifies the entire system. The key problems here are seen as the interactions between mobility and quality of service, the adaptation of applications and protocols to a wide variety of air interfaces with varying QoS, and the unification of a disparate set of Internet protocols into a coherent mobile network.

The BRAIN network layer provides a universal IP-based foundation for mobile wireless networks. It encompasses both the terminal and the infrastructure of the access network. The scope of the BRAIN network layer is shown in figure 1.

- In the terminal, it consists of an Internet protocol stack with backwards-compatible optimisations for mobile applications, and a lower convergence layer interface towards the selected radio technology.
- In the access network, it provides support for local mobility which is optimised for transport of IP application data, and makes the assumption of a direct interconnection with fixed IP backbone networks with a standard routed interface.

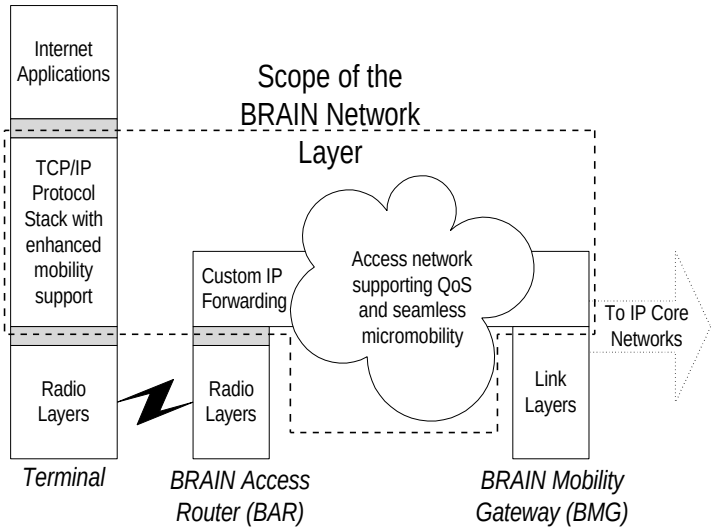

Figure 1: Scope of the BRAIN Network Layer

This paper describes the architecture of the BRAIN network layer. Firstly, the overall design approach is discussed, with comparison to the current paradigms of GSM/UMTS and the Internet. Secondly, the top-level architecture is presented, with attention to the specific problems of link layer integration, mobility, and quality of service. Throughout, the main open problems for the primary components of the architecture are mentioned.

Design Principles for the Access Network

General Approach

Up to now, current 2nd and 3rd generation mobile networks have used an architecture which is highly integrated and tightly specified, providing a complete definition of the complete network, all the way from air interface coding to application support. This provides for optimal performance and a high degree of user convenience in areas such as roaming and interoperability. However, the resulting system is slow to evolve in the face of new technological innovations, and seamless integration with alternative network types is a challenge.

An alternative approach has guided the evolution of the Internet under the auspices of the IETF. The emphasis is on the development and standardisation of individual protocols according to certain basic principles, which can then be implemented by manufacturers and deployed by service providers according to their specific needs. While this provides for rapid network evolution and great flexibility, sacrifices are made in performance and seamlessness of integration.

The BRAIN network layer architecture is built on a unification of these two approaches. It adopts the 'protocols as building blocks' model of the Internet, recognising that this is the best approach to cope with rapid advances in telecommunications engineering. Indeed, wherever possible, existing Internet protocols are used unchanged, in particular within the core network where no BRAIN-specific functions are assumed. However, it is not sufficient to consider protocols simply as independent, free standing components. We also require a framework within which the interactions of these components can be considered and controlled. This allows the determination of overall BRAIN network performance and verification of system completeness. Crucially, it also allows for the evaluation of alternative 'plug-in replacement' solutions for specific parts of the problem. Adopting this method allows us to approach the level of system optimisation found in traditional public mobile networks.

Fundamental Principles

During the initial architectural investigations of the BRAIN project, the following critical design guidelines have been followed.

Obey the 'End-to-End Principle': Classically, this means that the network should offer some kind of minimal service to end systems – in other words, it should be 'stupid'. In the mobile environment, the term is unfortunate since it is hard for a high performance mobile network to be truly stupid. Nevertheless, the underlying concept still applies and in the BRAIN context it is refined concretely as follows:

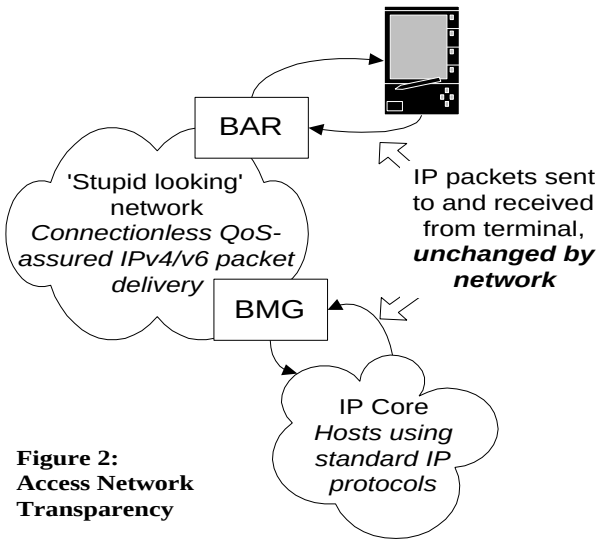

Figure 2: Access Network Transparency

- Be independent of specific transport layers and applications.
- Provide only a generic connectionless IP service, which offers (with varying degrees of performance) to get packets between the terminal and core network.
- Be independent of what packet type is being transported, and assume simply that packets are forwarded according to their IP header. In particular, don't assume that any transport layer or mobility encapsulation is used above them.
- Minimise the number of special functions that are provided in the access network.

These concepts are summarised in figure 2.

Obey the Layer Model: As discussed above, the access network should limit its functionality to providing IP packet forwarding, independent of upper layer applications or specific link layers.

- The network layer should have a generic interface towards the link layer, such that new (and old) radio technologies can be exploited without wholesale redesign of the network infrastructure.
- Where applications require optimised support, this should be invoked in a generic way – typically via some sort of QoS aware service interface.

This combination of generic upper and lower layer is fundamental to the BRAIN network layer architecture. Indeed, while the BRAIN project as a whole initially uses HiperLAN Type 2 as a starting point, the expectation is that the BRAIN network layer can ultimately be implemented efficiently over any air interface technology or even combination of technologies.

Minimise Barriers to Evolution: This applies to applications, link layers, and indeed components of the network itself. This is especially true of the public mobile environment, where a radical upgrade may involve hundreds of organisations and hundreds of millions of terminals. Likewise, it should be easy to deploy a new system incrementally, for example, starting initially with limited performance.

- Components within the access network should be modular, so that different parts can be evolved and upgraded independently. Later enhancements can be carried out transparently to the end users.
- Logical interfaces between the terminal and network should not enforce the use of particular internal protocol, but should allow network providers to choose solutions appropriate to their own circumstances.

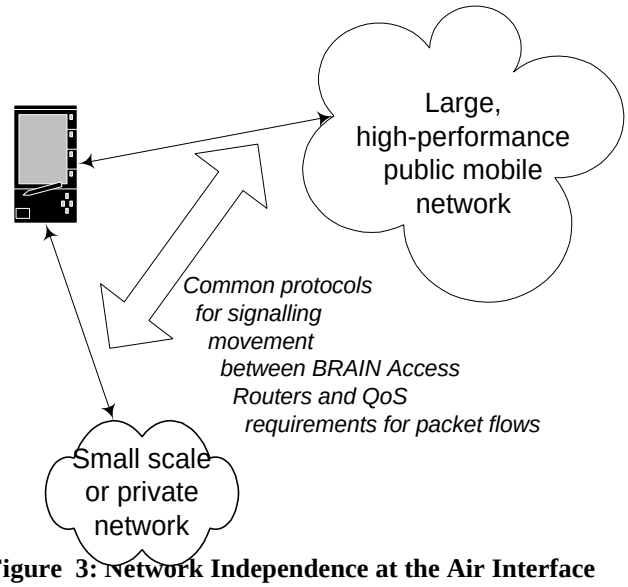

Figure 3: Network Independence at the Air Interface

The BRAIN Network Layer Architecture

In this section, we present the key features of the access network layer architecture that arise from these principles.

Addressing and Scaling

The basic goal of any given BRAIN Access Network (BAN) is to make mobile wireless Internet access look like 'normal' access through wired infrastructure. Thus, a BAN must allow a terminal to get an IP address to use in communicating with correspondent hosts in other networks; the BAN routes packets to and from this address in a way which externally looks the same as any other IP network.

The mechanism of address assignment has not been fixed, although solutions such as DHCP are one typical option; in any case, this is a function of the link convergence layer, which is discussed below. One assumption for BRAIN is that the address is unique to the terminal, rather than shared (e.g. as would be the case for 'foreign agent care-of addresses' of Mobile-IPv4). This is a consequence of the requirement for a clean, unified solution that applies to both Mobile IPv4 and IPv6 (and indeed many other higher layer protocols), recognising that shared addresses are simply one mechanism for IPv4 address space conservation, which is often ruled out because of security and other considerations.

Once an address has been assigned, the fundamental role of the BAN is to support seamless mobility of the terminal as it moves between access routers. In consequence, the allocated address must remain valid throughout the entire BAN, so there is a direct relationship between access network scalability and address allocation. There are essentially two options:

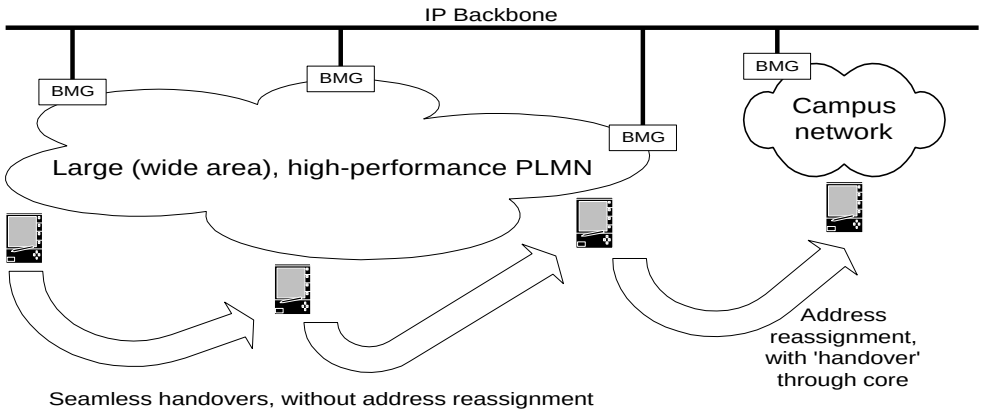

Figure 4: Address Assignment and Access Network Scalability

- If seamless mobility within a single area only is required, a BAN is allowed to interconnect with the core network at a single point, corresponding to a single BMG.
- If seamless mobility over a very wide area is required, the performance of the Internet prevents us relying on BAN-BAN handovers to support this. Therefore, the combination of wide area support and seamless terminal mobility forces the use of multiple interconnects with the core.

This is one example of using the option for different protocols within the BAN depending on service provider requirements, since achieving very high scalability for a terminal mobility and QoS protocols is a hard problem and not relevant to (for example) a campus network operator. In either case, it is assumed that a BAN is under single administrative control, and seamless handovers between administrations are not catered for. The combination of these scenarios is shown in figure 4.

Inter-Layer Interfaces

In an activity concerned only with interoperability, there is no need for inter-layer interfaces since these can be considered as implementation issues. However, abstract interfaces play a valuable role in partitioning the mobile networking problem, and clarifying the behaviour expected from or supported by particular network components. By extension, they provide a framework for research into the operation of particular functions (for example, header compression or TCP performance). In the Internet world, service interfaces have traditionally been minimalist; however, enriching the functionality of these interfaces is one mechanism for allowing network performance enhancements towards the level of traditional PLMNs while preserving layer separation.

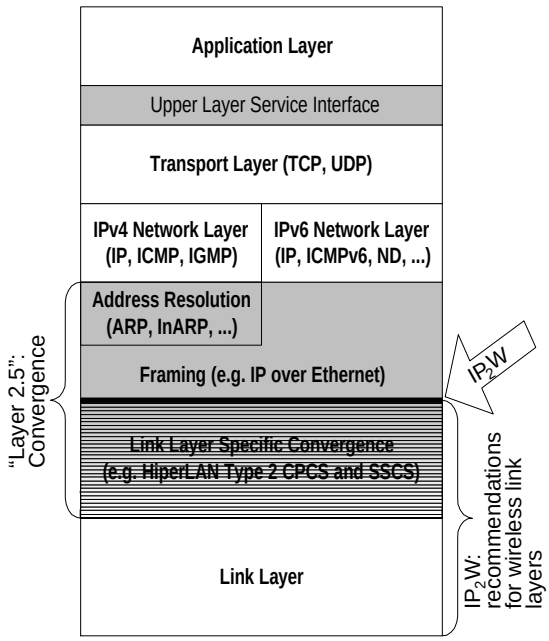

Figure 5: Inter Layer Interfaces

The BRAIN network layer relies on two inter-layer interfaces for this purpose. The first lies above the basic network and transport protocols and provides the enhanced application support that is necessary in the mobile environment. Broadly, it allows for extended negotiation of QoS information between the application and lower layers, including renegotiation during active sessions; further details are given in [2]. It exists only in BRAIN terminals The second is a specialised interface for matching the IP layer to wireless layers, hence the name 'IP_2W', and is common to terminals and access routers. The interfaces are shown together in figure 5.

The combination of these two interfaces is the key to allowing the development of a re-usable, QoS capable TCP/IP stack which offers advanced facilities to applications, yet is efficiently integrated into link layer. The main issues at this interface are link setup / release, layer 2 and 3 address assignment, link layer QoS negotiation and re-negotiation, and the interaction between this and buffer management (scheduling) in the network layer. In particular, IP_2W enables the use of layer 2 procedures which are much more efficient than equivalent IP protocols operating over a generic data interface. The performance of these operations and the level of control that upper layers have over them has a direct impact on the performance of handovers at the IP layer and on the QoS received by a mobile user.

In detail, the IP_2W interface is separated into a Data and Control part, each offering access to some functionality at the link layer. Several distinct functions have been identified under the interfaces, shown in table 1; some are optional, and the link layer advertises which it supports through a configuration interface. The control interface is also used to control the operation of some of the user plane parts such as buffer sizing and error control characteristics. The model mandates no specific structure within a given link layer, and indeed, some functions may be inherent in a particular link type, while others may have to be added by a convergence layer. Where an option is not supported, the TCP/IP stack can fall back to a layer 3 protocol instead.

	Interface	
	Control	**Data**
Core	Configuration Management	Error Control
	Address Management	Buffer Management
Optional	QoS Control	QoS Support
	Handover Control	Segmentation & Reassembly
	Idle Mode Support	Header Compression
	Security Management	Multicast

Table 1: Functionality Visible at the IP2W Inter-Layer Interface

Mobility and Quality of Service Protocols

We have already defined the key responsibilities of the BRAIN network layer as being to support mobility and quality of service for mobile terminals. But what exactly does this mean? For mobility, we can distinguish three broad classes of requirement.

1. Users can advertise reachability at a given (IP) address, e.g. using Dynamic DNS or SIP. The registration procedures are assumed to be transparent to the access network address assignment.
2. User can maintain a ‘permanent’ IP address as they move between networks. This function is the domain of classical Mobile IP, and is often loosely referred to as ‘macromobility’.
3. Users can move rapidly between wireless access points, without needing to repeat registration or macromobility procedures, and preserving the illusion of a seamless connection to a fixed network. This function is loosely referred to as ‘micromobility’.

Assuming colocated care-of-addresses only, use of Mobile IP affects only the terminal, since the access network offers a transparent packet delivery service which is aware only of the network-assigned care-of-address – this situation is shown in figure 6 below. Therefore, it is only the third of the mobility requirements that is directly the concern of the BRAIN access network. However, as already discussed, the network must be prepared to provide a *complete* solution to this problem even over a wide area, since Mobile IP within the core cannot be assumed to support this performance.

There is already a family of ‘IP-based’ micromobility protocols that have been proposed to solve this problem, such as Cellular IP, HAWAII, tunnel proxying solutions, and ad hoc routing, and evaluation of these is currently in progress. What is clear is that while several of them fit into the basic model of support for very fast local handovers for terminals using a semi-static IP address, none of them satisfy all the requirements for QoS integration, scaling, or efficient link layer integration, and this will be one of the areas of future study.

The second key functional area for the BRAIN access network is quality of service support. QoS-aware packet processing within the terminal is handled by implementation in accordance with the service interfaces described earlier; the focus here is on how the terminal negotiates QoS requirements

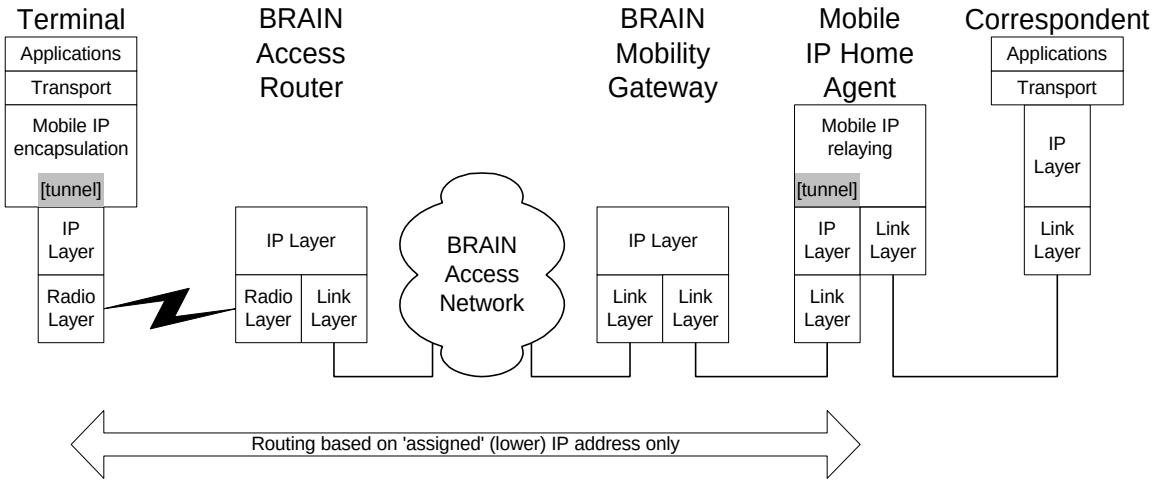

Figure 6: Transparent Mobile IP Support in the Access Network

for its data flows to the network, and how the network manages itself to support this. The interactions between QoS and mobility are severe, with the major problems being

- The need, in an Integrated Services approach, to move an existing reservation to a new path;
- The fact that cell-local congestion may require QoS renegotiation after a handover event;
- The inability of current fixed-network QoS protocols to deal with aspects of the mobile environment such as payload BER criticality and packet level behaviour during handover.

Again, evaluation of current IETF approaches is in progress, and again it is clear that existing IntServ and DiffServ solutions do not meet all the requirements of the environment under consideration.

In traditional 2nd and 3rd generation systems, both handover and QoS are considered as aspects of radio resource management, and the BRAIN architecture builds on this. The complete radio resource management strategy for a BRAIN network includes cell selection, handover initiation (intra and inter access network), admission control according to required QoS and current system load, bearer control, and dynamic channel allocation. The BAN can also interact with external resource management entities to obtain other IP network resources for the session; alternatively, the user can signal requirements transparently through the BAN to a remote network, for example using RSVP. Efficient and fast interactions between radio resource management, and micromobility and QoS management entities are essential and this support and integration is provided primarily by the IP_2W interface.

The close interaction of these areas of micromobility, QoS support, and radio resource management is notable and provides a guide and unifying theme for further investigations. This will be to develop a common protocol framework supporting the full range of radio resource management functions at the air interface, which allows for the integrated operation of a variety of QoS and mobility protocols within the BAN. This will be the focus of the next stage of the project.

Acknowledgement

This work has been performed in the framework of the IST project IST-1999-10050 BRAIN, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues from Siemens AG, British Telecommunications PLC, Agora Systems S.A., Ericsson Radio Systems AB, France Télécom – R&D, INRIA, King's College London, Nokia Corporation, NTT DoCoMo, Sony International (Europe) GmbH, and T-Nova Deutsche Telekom Innovationsgesellschaft mbH.

References

[1] "Broadband Radio Access for IP Networks (BRAIN)", D. Wisely, W. Mohr, J. Urban, IST Mobile Summit, October 2000

[2] "BRENTA – Supporting Mobility and Quality of Service for Adaptable Multimedia Communication", A. Kassler, L. Burness, P. Khengar, E. Kovacs, D. Mandato, J. Manner, G. Neureiter, T. Robles, H. Velayos, IST Mobile Summit, October 2000