Special issue of IEEE Personal Communications –
"European R&D on Fourth Generation Mobile and Wireless IP Networks"

An Open Service Architecture for Adaptive Personal Mobile Communication

Author: Theo Kanter

Ericsson Radio Systems AB, SE-16480, Stockholm, Sweden
theo.kanter@ ericsson.com

Abstract. Currently, 3GPP is reinventing the telecom model of services in wireless Internet — a model that is not well suited to (1) meet the demands for new modes of personal mobile communication, enabled by wireless packet services and multimedia devices, or (2) the ability to deal with increasingly heterogeneous wireless infrastructure (one interpretation of 4G). This article characterizes the properties of service architectures in relation to the steps taken in successive generations of wireless communication networks for personal communication. The article then continues to present a novel service architecture for open communication in wireless Internet, describing its necessary properties and evaluating its merits. Finally, we present our experiences from building application prototypes based on our service architecture, in an urban wireless testbed consisting of WLAN extensions to a Gigabit-Ethernet network.

I. INTRODUCTION

This section describes the background to this research and what motivates it. The model for personal communication has been mobile telephony with fixed amounts of bandwidth reserved for voice communication between two parties and GSM has been successful at implementing this model and achieving a world-wide penetration of 250 million users before the year 2000[1]. The enormous popularity of the Internet has caused us to look for access to Internet through any network and the GSM industry has responded with the Wireless Application Protocol, which placed itself between the user and the Internet, and thus leaving the user at the mercy of the operator, who controls the WAP-gateway and who determines what services are available to the user. To make things worse, users where connected over circuit switched wireless access, incurring excessive costs for very limited service. These limitations have been removed in iMode, where users are connected to the Internet by means of packet-switched network access, and where third-party content is not published in the operator's gateway, but can be sent directly to the mobile as compressed HTML. A general-purpose packet radio service (GPRS) is now becoming available in GSM, and the bandwidth will increase in successive generations towards that of Universal Mobile Telephony System (UMTS). Simultaneously, Wireless LAN (WLAN) technologies are available at increasing speeds, which can deliver magnitudes greater bandwidth within a shorter range. Consequently, we can build mixed systems, where wide-area wireless access (GPRS/EDGE/UMTS) is interspersed with WLAN access (so-called hot spots). However, in light of the fact that we can deliver both asynchronous and isochronous multimedia with end-to-end IP connectivity over all these wireless links, we need to reconsider what services we really will be delivering and also reconsider the basic design of the architecture.

II. NEW MODELS OF COMMUNICATION

Mobile devices are becoming extremely computationally capable, thus they are able to not only generate and process the multimedia, but also to generate different visualizations of the multimedia, gather information about the user's context (e.g., using position, direction, acceleration, or other sensory information from either the user or the environment) or communication conditions, and make decisions to adapt its own behavior. User carried mobile devices, resources, and (potentially intelligent) virtual objects should be able to connect to each other and auto-configure their communication. Thus, we need mechanisms and an open service architecture that enables these entities to accomplish this without having to rely on unique services that are present in a single type of access network, nor should we have to rely on a single operator or even the presence of a network to setup these service. Thus, we allow users to use ad-hoc applications. Ultimately, we wish to create a plug & play Internet, allowing users to connect to virtual spaces that can be correlated to physical spaces and be able to jointly observe and manipulate objects, while conversing with others in real time. For example: a user A who joins a mobile group visits some physical locations, where he or she leaves location-dependent voice annotations with notification triggers for others in the group to discover. Subsequently, User B passes the same location and finds the note & plays it out.

[1] Source: GSM Association.

III. DESIGN GOALS

In light of this new model of communication, the design goals for the communication network and its components are to support negotiation of these new services between end-user and service provider, without requiring the assistance of the network operator or network access provider. Even more importantly, we should allow any user to become a service provider whenever this applicable (e.g., live broadcast from an event). This overall design goal can be divided into the following parts:

1. Deregulated Access
   End-users are allowed to connect to the Internet via *any* network. Trust relations between an end-user and *any* visited network can be negotiated via a trusted third party, thus enabling anonymous Internet access [3].
2. Device Mobility
   End-users are allowed to roam within a network or between networks, implying that any on-going sessions must be retained. This requires the automatic assignment of a network address to the mobile device and redirection of the on-going communication to the new address, either by making this redirection transparent as a network service (e.g., Mobile-IP), or combinations of a transparent network service and mobility awareness on the application level (e.g. SIP and Mobile-IP) [30].
3. Service Mobility
   A service that is available to the end-user in one network can follow the user to another network. Not only should this service remain reachable in terms of addressing, but it should follow the user to the new network and operate in the new context (e.g., a print service should discover available printers).
4. Personal Mobility
   End-users are allowed to move to other networks and remain reachable using the same address, not only for voice communication (e.g., personal numbers in fixed or mobile networks), but also for any other application.
5. Adaptive Communication
   The end-user's mode of communication adapts itself based on events (e.g., not just calls, but also any type of sensor input, such as spatial data, bandwidth, content availability) to user context or communication conditions. This implies that users can benefit from storing and reusing service data in order to learn how to respond optimally to communication events (see section VI.E).
6. Ad-hoc Application Negotiation
   The negotiation of communication between users, resources, and virtual objects should require only minimal shared service knowledge, implying the necessity for a general-purpose service negotiation protocol (see section VI.G).

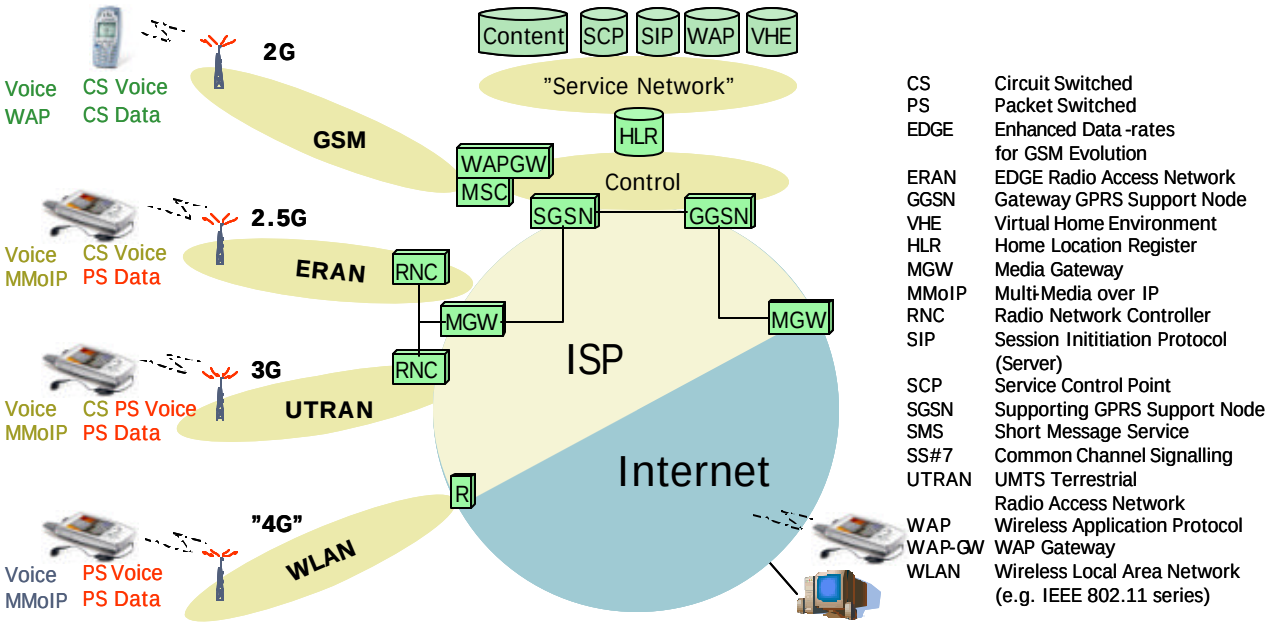

Fig. 1 3G Evolution and beyond

IV. SERVICES ARCHITECTURES (PREVIOUS WORK)

This section characterizes the service architectures that are used in 2G, 2.5G, 3G, and beyond. In addition it characterizes the properties of wireless networks beyond 3G. Fig. 1 provides an overview.

A. 2G

In 2G, mobile devices authenticate themselves and the identity of the user while reporting their location to the Home Location Register (HLR). Speech or data sessions are based on circuit switching of radio channels. A very limited packet data service is provided by SMS. Except for SMS, all services are mutually exclusive. Additional client software in the mobile device (e.g., for Personal Information Management) may be used to invoke the services resulting in so-called Smart-Phones. WAP-clients in the mobile device offer a simple interface to Internet content that can only be accessed through a WAP Gateway, which translates between IP and WAP protocols. A web server on Internet can eliminate the need for an HTML filter by publishing pages with Wireless Markup Language (WML) tags. Through the use of WML Script content, other services can be invoked (e.g., sending short messages, invoking calls). By following a specific URL, the user can download and play a video from a media server. Using WAP in the mobile terminal causes user services to be strictly dependent on the functionality of the WAP-GW, and thus dependent on the network operator. In addition, circuit switched network access disallows asynchronous application events; this greatly limits the type of services that can be offered to users in a meaningful way.

B. 2.5 G

GPRS and EDGE will remove some of these limitations, by offering packet data service. Mobile terminals authenticate themselves to the GGSN and report the location of the user to the HLR through the SGSN. The Mobile Device obtains an IP-address from the GGSN. There are different traffic classes allowing for combinations of switched GSM and packet-switched GPRS traffic. The current standard

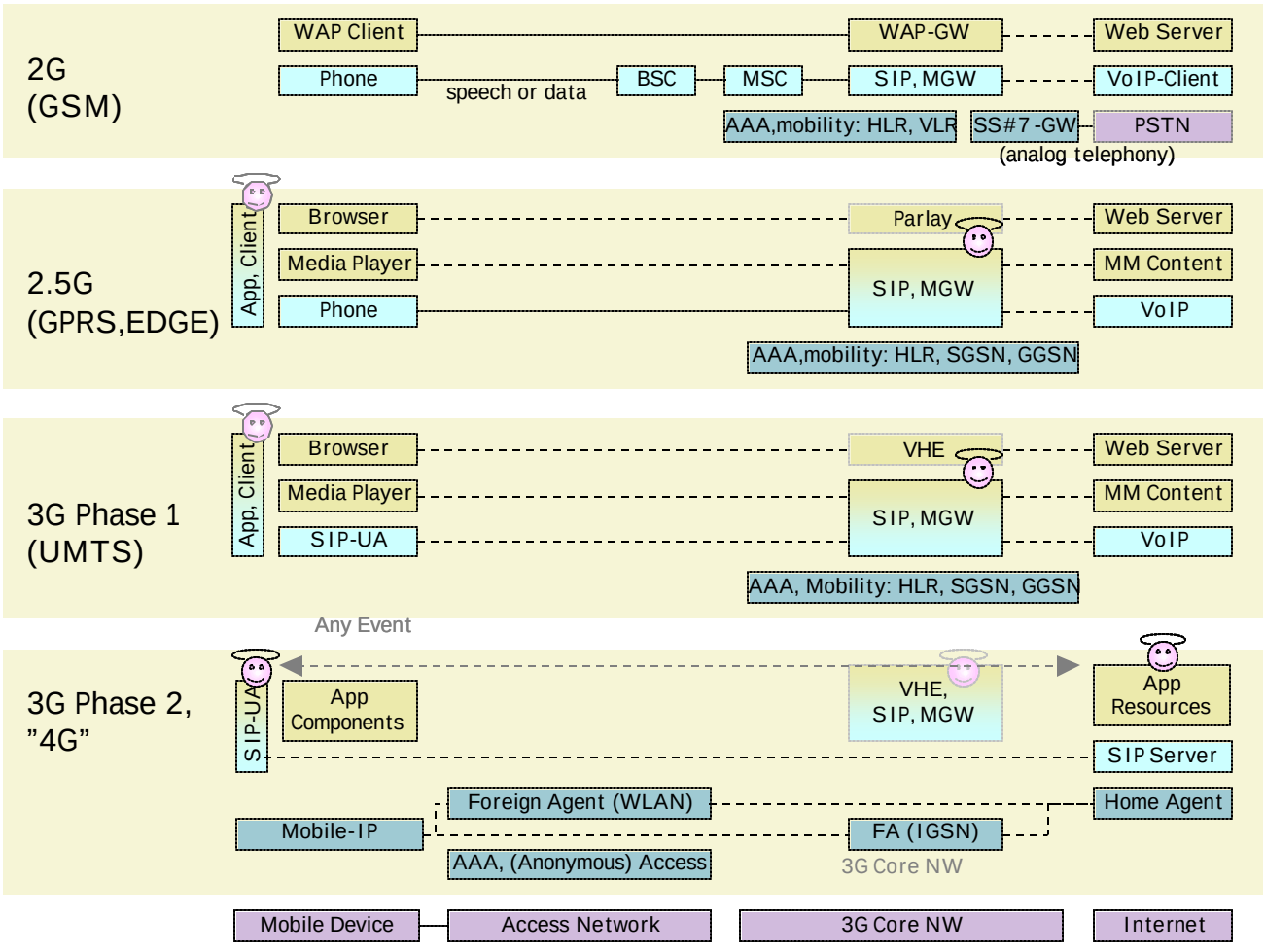

Fig. 2 Service Architectures

for GPRS data traffic incurs considerable latency by interleaving data (in order to increase the reliability of data transfer) and to allow for per packet establishment of radio bearers (in order to optimize utilization of radio resources). The operator is still in the position to encourage, if not require that the mobile device be configured to use servers in the operator's service network to setup multimedia sessions.

A SIP server can be used to setup multimedia communication between end-points. This can be further enhanced by adding a Parlay-API [12] to the SIP server in order to execute servlets via a Corba interface on a web server. A web browser can be used for customer control of the services — in what can be regarded as a Virtual Home Environment (VHE), with integrated interfaces to a Service Control Point (SCP), in order to be able to control legacy services.

Scripted mobile code can be sent to and executed by agents that are co-located with an application client in the mobile device [6]. Moving the execution of code to the mobile device has various advantages, e.g. performance, and allows the device to report local states back to the server.

Parlay — The Parlay Architecture is based on Corba interfaces that enable hosting of applications outside of specific networks while accessing resources in other networks, through gateways that are installed by the network operator, making these applications and services available to the user irrespective of what network the user is located in. The Parlay API specifications are open and technology-independent, so that anyone can develop and offer advanced telecommunication services.

Clearly we can move services between different networks but only within Parlay domains, but this process is entirely controlled by the network operators. Fig. 3 shows an example of how a simple service using these interfaces can be built. This example was used to prototype wake-up calls and location-dependent information push services in a mobile network.

What is particularly important about this example is that the controlling web interface and the application are only synchronized through network-based servers across a network boundary. Mobile code can be sent to the device to enhance user interaction, but the process must be carried out under the supervision of the application servers and require synchronization across network boundaries. Furthermore, the Parlay interface must be changed each time to reflect capabilities that are present or introduced in SIP [4]. Parlay has these two limiting properties in common with other network-centric service architectures, such as WAP, VHE (see IV.C), or TINA-C [18].

C. 3G Phase 1

Mobile terminals authenticate themselves and report the location of the terminal to the HLR through the combined SGSN and GGSN (which also assigns an IP address to the mobile terminal). 3G Phase 1 supports real-time and isochronous multimedia (e.g., voice calls) using end-to-end connectivity over wireless links, which is set up using servers in the operator's service network, to negotiate session parameters regarding quality of service levels (QoS). A Virtual Home Environment ensures that user access to services is independent of the location of the terminal, and that the user interface is independent of the terminal, for instance using (as in 2.5G) a web interface (HTTP) and Java for customer control.

In summary, the service architecture does not

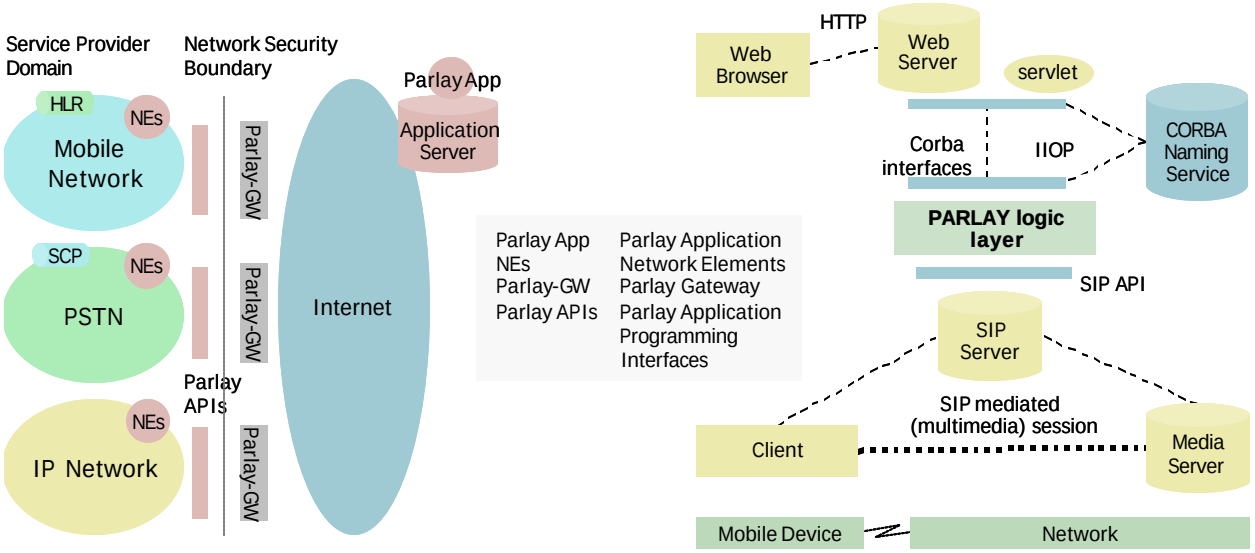

Fig. 3 Parlay Architecture and Prototype of Parlay Access to SIP resources

differ in principle from the one in 2.5G, and the service architecture offered by an operator of a GPRS, EDGE, or 3G Phase1 network requires any communication, beyond simple browsing of web pages to be *mediated by servers in the operator's Service Network*. Negotiation of services and levels of QoS linked to network specific AAA and mobility mechanisms effectively blocks any possibilities to import or export services to/from Internet ad-hoc. While this service architecture makes perfect sense from an operator's point of view (and follows an established business model), it disallows or in the best case makes it extremely complicated to support the types of communication that we propose. Service mobility between this and other networks can in principle be solved on a per service basis, with adaptations to deal with the specific requirements for mediating functionality in the Service Network. However, we believe this is makes services harder to deploy rather than easier!

Virtual Home Environment (VHE) — is a concept for providing personalized service portability across network boundaries and between terminals. The concept of the VHE is such that users are consistently presented with the same personalized features, User Interface personalization, and services in whatever network and whatever terminal (within the capabilities of the terminal) and wherever the user may be located. For UMTS phase 1, VHE consists of GSM services & roaming principles and Service capabilities — see Fig. 4.

The service capabilities offered are call control, location & positioning, PLMN information & notifications, and bearer establishment. It is clear that any services created in this platform are unique to this platform and network, and cannot be moved outside of a 3G network to the Internet, and this architecture is not open to executing random services on the Internet.

D. 3G Phase 2

Mobile terminals authenticate themselves to AAA-servers via the integrated SGSN and GGSN node (IGSN), which also acts as a foreign agent for mobile-IP, thus assigning an visiting IP address to the mobile terminal. The IGSN reports the location change to the home agent of the mobile terminal and forwards the AAA information to the HLR for charging purposes. This AAA and mobility scenario enables the mobile to negotiate communication with resources outside of the 3G networks without intervention of servers in the operator's Service Network. Naturally, the operator can offer support for different levels of QoS and even differentiated charges, but the fact remains that the services are negotiated end-to-end, and not simply inside the operator's Service Network. However, mobile terminals are required to have detailed knowledge of such support services, which may differ between networks and may change over time. Thus, we need a means to describe shared knowledge of support services and also means to automatically obtain such knowledge in order for services and mobile devices to migrate between networks. This is one of the design goals of the extensible Service Protocol (see section VI.G).

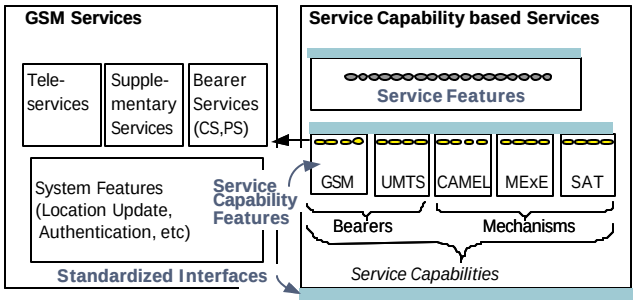

Fig. 4 VHE Realization

E. 4G[2]

Since multimedia can be delivered with end-to-end IP connectivity over wireless links, this allows us to extend all existing voice services to these networks. So-called 'hot spots' equipped with wireless LAN (WLAN) extensions to the Internet are becoming available, and today provide us with even higher bandwidths (e.g. 11 Mbps in IEEE 802.11b), for example Telia's HomeRun system [31], corporate WLANs, and "semipublic" WLANs.

This is particularly important, since broadband Internet access is being provided in a rapidly increasing number of public locations (hot spots) and even homes in urban areas. The provisioning of broadband Internet access is being installed / provisioned by power companies, transportation companies, housing co-operatives, joint-ventures of municipalities, etc., all of whom have a radically different business model than traditional telecom vendors and operators of cellular networks. Extending this packet-switched infrastructure with wireless access points, such as IEEE 802.11b Wireless LAN is straightforward. In addition, mobility solutions, such as Mobile-IP, and IPv6 are available to provide the necessary scalability that accommodating millions of users and devices will require.

Furthermore, solutions for direct access to Internet (not requiring an existing subscription, but rather a direct settlement, e.g. with Ecash) are available [3]. In fact, such operators simply provide IP-access, they

[2] Different interpretations exist of "4G", e.g., new high data-rate radio interfaces. Here, "4G" refers to heterogeneous wireless packet networks.

do not necessarily even need to do authentication, authorization, and accounting (AAA), since they get paid directly or indirectly.

Consequently, users with mobile devices can, in principle, use any service from any third party, without any intervention by the operator that provides the network access. It should be noted that attempts to limit the customer's choice by incumbent operators have been found to violate the EU's competition laws.

Thus, the properties of 4G are such that it provides users with (1) multimedia over end-to-end IP (wireless) links with (2) high-bandwidth, between (3) multiple, heterogeneous, access networks, and with (4) direct access to the Internet and thus end-to-end IP-connectivity to (5) third-party mobile multimedia services, without the need for prior subscription for Internet access with these access network operators.

In section IV, different service architectures for 2G (WAP), 2.5G (Parlay), and 3G Phase 1 (VHE) were presented, which share a common property of making network based services available to mobile devices over a network boundary. Removing this boundary, i.e. locating our services on Internet, also allows us to move these services out to the mobile device, to the resources, or to any virtual object on the Internet. Therefore the service architecture — which is presented and discussed below — should not only be *open* in the sense that the previous ones said they were (making the interfaces public) [11], but anyone or anything can at anytime publish a service for or use a service from anybody else.

V. Internet Service Architectures

A. UPnP and JINI

Universal Plug & Play (UPnP) and JINI [16,17] enable devices to connect to and use each other's services dynamically. JINI uses a server, whereas UPnP relies upon a control point that in principle can be co-located with the resource it represents (e.g. a printer). Resources and profiles are registered during a discovery phase, and then the server/control point assists in connecting devices during the lookup phase. Furthermore, the server/control point provides event services to send notifications of changes, e.g. when a resource leaves. Registrations are time-limited.

B. Tuple Spaces

In order to provide even greater flexibility, researchers have investigated tuple-space based architectures (e.g. IBM T-Spaces [24]), in which devices and resources are able to store and share common application knowledge, by connecting to a tuple-space server. The advantage of this architecture over JINI and UPnP systems is that we more easily can add shared knowledge. In all three cases, we encounter potential scaling problems, as JINI, UPnP, and tuple-space architectures require the assistance of a server. Although in the case of UPnP, so-called control points could in principle be co-located with the agent in the device.

C. Adaptive Communication

Examples of creating adaptive communication are Cognitive Radio [23] or the Oxygen project [33] regarding local and global properties for communication devices and systems.

D. Summary

What we are looking for is an *open* approach, which *scales* to the size of mobile networks while retaining the flexibility of tuple-spaces in order to be able to *adapt* our communication. JINI and Tuple – Spaces do not scale being server centric. UPnP protocols allow locating a control point per device, but services register in and send events to any control point that is listening. Consequently, control points cannot be regarded as presenting a device, where we are looking for an entity that we can co-locate with any object and let it be the object's representative.

The next section introduces a novel service architecture that combines plug & play capabilities with support for adapting the communication to arbitrary events, and which scales well.

VI. An Open Service Architecture

A. Communication Space

The Service Architecture constitutes a *framework* for creating application level connectivity networks of communicating entities in a *communication space* [1]. These entities can be mobile artifacts, resources, sensors, (potentially intelligent) software objects, or users, each of which is represented by a mobile agent that respond to events from other entities, thus

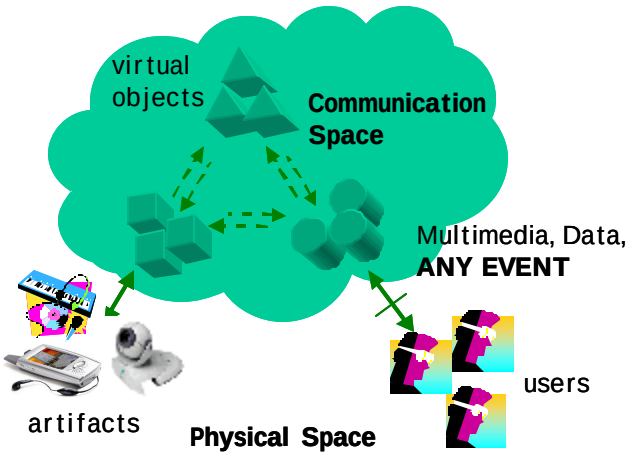

Figure 5 Communication Space

enabling applications where virtual and physical spaces are intertwined.

B. Overview

The user is represented in this communication space by a Personal Agents, which can act on the user's behalf if the user is off-line, and migrate into the user's mobile device, once the user becomes logically on-line. The user can interact with this personal environment using a mobile device or any of the sensory equipment that is available in the communication space.

The user may "visit" his or her Personal Agent, using a web interface. Naturally, in this case, user interaction is limited to managing user-preferences and information items, due to the limitations of a web browser interface.

Personal Agents act in the communication space through events. Events are transported between entities using a novel eXtensible Service Protocol (XSP) that is described in section VI.G.

The service architecture (Fig. 6) has mandatory and optional components and therefore varies between application areas. Mandatory parts are the agents and the eXtensible Service Protocol (XSP). The Active Context Memory and SIP (naming, localization, session initiation) are both optional (e.g., in the case of sensors only the mandatory parts are needed).

Service components such as VoIP, Chat, MP3 streaming, or three-dimensional virtual spaces (3D) are also optional and can be incorporated by adding actions with corresponding session description parameters for invocation via SIP.

The remainder of section describes the logical elements of the service architecture and the context in which this architecture and XSP operate.

C. *Naming, Localization, Registration*

Each agent has a unique identifier (SIP URL), which enables it to be located, invite other agents, or be invited to communication using SIP. In addition, XSP allows agents to discover and register with each other, provide (implicit) subscription to events, and exchange service profiles.

Furthermore, the SIP URL provides service mobility because agents can move to other SIP-servers and use SIP to redirect communication to their new location.

Personal Agents are locatable through a SIP-URL, thus providing personal and service mobility.

D. Agents Profiles

Agents have an external communication profile that specifies a number of properties:

```
<object>
   <type>           // ontology
      <id/> <rdf/>
      {[<space>]    // optional
       [<db>]
       [<sensor>]}
   </type>
   <actions>
      <action>
         <event>    // trigger templates
         [<event>]*
      </action>
   </actions>
   <states>
      <public>      // visible states
         <state>
      </public>
      <state>       // invisible states
   </states>
</object>
```

Fig. 7 Agent Profile

<type> has a reference to a Resource Description Framework (RDF) file [26], which acts as an XML [27] equivalent of an ontology[3] and contains resource metadata modeled as object relationships between resources. Other agents look for semantic equivalence to one of the related resources during registration and are thus able to enter an ad-hoc application network, without requiring user intervention..

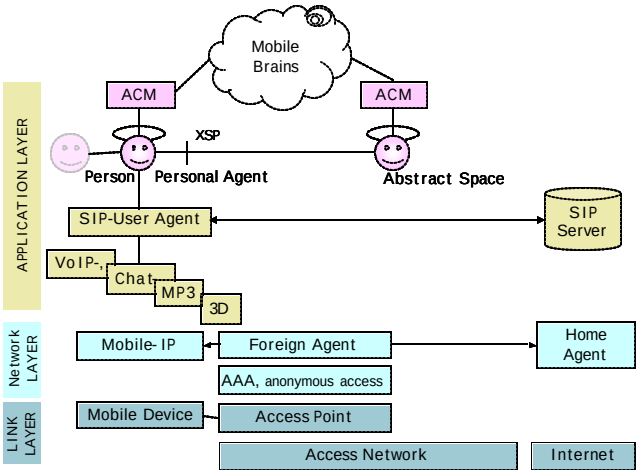

Fig. 6 Service Architecture and Protocol Stack

[3] An ontology states the concepts in the universe of discourse, similar to a data dictionary.

<*space*>*:*** Agents can enable a *space* tag (optional), indicating that it will propagate information about all registered agents to other agents. The enabled tag allows an agent to act as an *abstract space*. Other agents registering with this abstract space obtain information about other registered agents (i.e., objects that are present), and can register with these other agents. Agents with the space tag enabled will forward notifications of events that have been received from one its registered agents to all other registered agents; while agents with a disabled *space* tag will not.

<*db*>: Agents can also volunteer to act as a repository for background service knowledge for other agents, by enabling the (optional) *db* tag. In principle, agents can volunteer to act *both* as a space and a repository, but there is no requirement that they should. XSP offers a limited default set of actions for knowledge storage and retrieval with regular expressions.

<*sensor*>: Sensors can be monitored by agents either via an internal interface, or via an external API (e.g., Java RMI). Stimuli from the sensor cause the agent to act. Details are outside of the scope of XSP. However, by setting the *sensor* tag agents will understand that this agent has only mandatory functionality, and its actions are limited to monitoring and setting properties. The *sensor* tag is mutually exclusive with both the *db* and *space* tags. Agents of this type allow us to build and incorporate sensor networks in our applications.

Actions are XML representations of rules that are triggered by a logic expression of event patterns, allowing it to contain variables or not even to be fully instantiated. The action repertoire can be *extended* as a result of internal reasoning in the ACM (see below) or more likely as the result of a negotiation with another agent resulting in it sending the mobile code for this new action to add to the receiving agent's repertoire.

States can be either public or private. Actions cause state-changes as side effects. Therefore, exporting actions also results in exporting states to the receiving agent's profile.

E. Active Context Memory

Agents contain an optional component, the Active Context Memory (ACM), the purpose of which is to respond to XSP events and maintain the agent's service profile. To this end, the Active Context Memory contains a rule based knowledge structure and inference engine. The ACM is usually not present in simple artifacts such as sensors, since their functionality cannot be extended, then XSP is limited to establishing a network of such devices.

F. Service Knowledge Repository and Data-Mining

All information regarding the *services* in this communication space is transported using XSP formatted as XML. XSP allows service knowledge to discover an agent, which acts as a repository, and moves service knowledge to it for later retrieval. The importance of this is that external applications can be used to data-mine service knowledge, which can provide important feedback to application service providers or operators about how there services are used.

G. An Extensible Service Protocol (XSP)

This section further details how the agents use the XSP protocol. Fig. 6 shows an overview of the protocol stack and the location of XSP in it.

The mobile agent handles both SIP and XSP. The agent handles the (de-) registration and keep-alive messages, but then hands the XSP based communication over to the Active Context Memory, following these general steps:

1. Find XSP enabled entities: broadcast, multicast requests on a well-known XSP-port. Knowledge of the existence of entities may be stored and be present in the Active Context Memory. In this case the agent merely re-registers.
2. Register with implicit subscription to events.
3. Query profile: type, actions, states, (and neighbors)
4. If the entity that is found is a sensor, just receive events, otherwise, if it is a repository (and not a space), invoke foreground/background knowledge handling with this object.
5. Analyze the RDF and select a matching object profile. The ACM looks for familiar semantic patterns. If there is a match then the querying agents requests the actions that it does not have in its repertoire.
6. If the ACM is unable to instantiate the received profile in its own knowledge, the profile will be presented to the user for input (if there is one), or else store this for later instantiation due to new user input or events. The ACM can also respond to the presence of agents (e.g., detection of a public display and speaker then triggers a video conferencing dialog).
7. If the agent that has been found is a space, then try to register with the other agents that it knows (repeat steps 2-7).
8. Send keep alive messages to other agents with which this agent is registered.
9. Extension of the agent's action repertoire results in sending an event to its registered agents. These

agents may or may not request to be sent this action.

H. Extension of Mobile Service Knowledge

In the previous steps there are two different contexts in which Mobile Service Knowledge in agents is extended, that need to be discussed further. In the first case, agents discover that to act as a related agent (as specified by the RDF template), they need to extend their action repertoire and ask the other agent to forward any missing actions. This *automated* extension of the repertoire of capabilities is a key feature of XSP. Secondly, agents can apply a paging strategy [1] to move mobile Service knowledge to other agents, which act as repositories, using exactly the same messages in XSP. In the RDF files that agents refer to in their type fields we can specify inheritance. This way we extend the *context* in which the agent specifies that it may be used.

I. Event Routing

A key feature of XSP is that its agent communication is based on events. The routing strategies for events differ between types of agents. Sensors do not route events, nor do repositories. An repository agent that also acts as a space, is a space from an event routing point of view.

Through XSP we can enable different modes in agents for different event-forwarding routing strategies, the default being that an agent (except sensors or repository-only ones) route events to all its *registered* agents (i.e., not *all* agents), which still is not optimal from a performance point of view, since an agent may receive the same events from several neighbors. Other modes can be more judicial in their use of available resources (e.g., energy, bandwidth), see also [28].

J. Merits of the Solution

This service architecture offers several improvements in comparison to the ones in 3G and proposed for the Internet. Not only are we able to move the services out to the mobile devices, we have also eliminated the artificial distinction between a controlling interface and the application interface, by removing the requirement to synchronize the two. Thus, services *can* but no longer *must* be hosted in the network. The eXtensible Service Protocol enables us to scale ad-hoc application networks, in which entities can adapt to their changing communication context.

VII. Application Prototypes Based on the Open Service Architecture

In this section, we present our experiences from building application prototypes based on our service architecture, in an urban wireless testbed consisting of WLAN extensions to a Gigabit-Ethernet network, exemplifying the new model of personal mobile communication and summarizing what has been in enabled as opposed to what was possible in previous work.

A. Experimental Network

We have built an experimental fourth-generation wireless testbed by extending Internet42, an existing Gigabit-Ethernet IP-network (Fig. 8). The project involved several parties: Ericsson Radio, Royal Institute of Technology (KTH), Telia, and Brf Bågen. Besides points of presence at research facilities (Ericsson Radio, KTH, and Telia) in the Stockholm suburbs of Kista, Älvsjö, and Farsta; Internet42 also has a point of presence in the center of Stockholm where it provides, at low cost, 100 Mbps network Internet access to each apartment in a large housing co-operative, Brf Bågen. The housing cooperative acts as an operator with the following distinguishing characteristics: Users get real IP-numbers, either statically (for servers) or dynamically through DHCP, there is no firewall to the Internet, and there is no restriction on traffic, neither between the users nor to the Internet.

We have extended the services of Internet42, by adding 11 Mbps wireless packet data access points (IEEE 802.11b), agent servers, media servers and content management, voice gateways (VGW) with

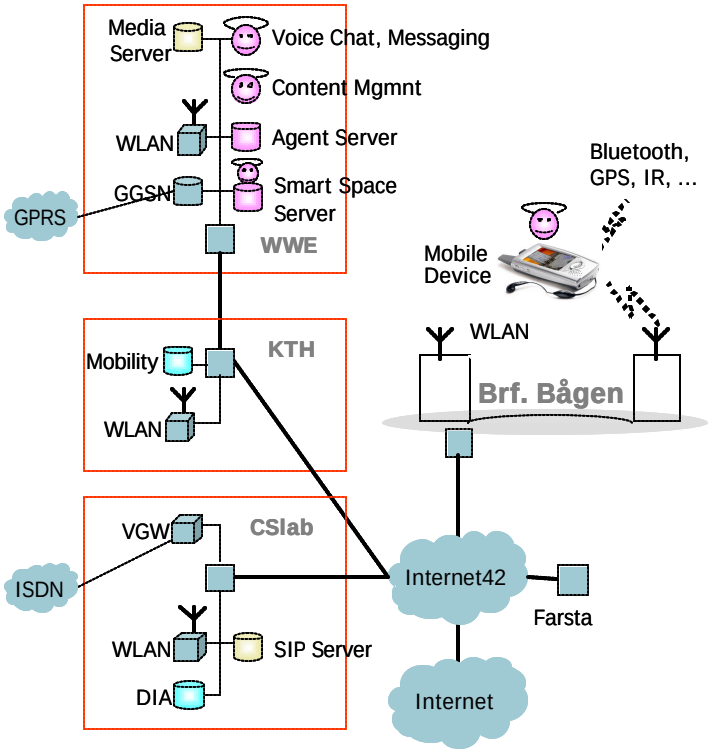

Fig. 8 Experimental Network Overview

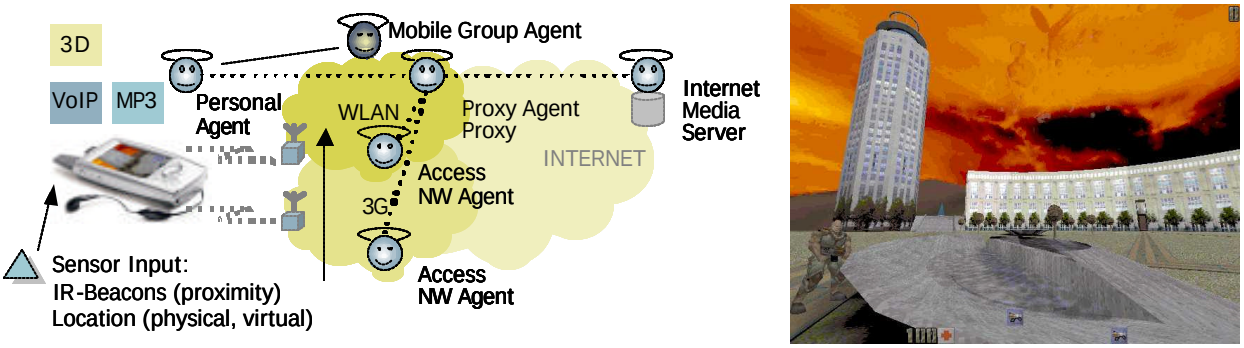

Fig. 9 Context- & Mobile-Aware Media Player Prototype and the Mobile Interactive Space

anonymous direct access to Internet (DIA), support for device mobility (Mobile-IP) and service mobility (SIP). We will are currently extending our testbed with GPRS access.

Direct Internet Access — We used Direct Internet Access (DIA) [3], which provides anonymous authentication and allows the access provider to charge via eCash. This approach thus makes access authentication keys redundant, and allows simple roaming access. Consequently, there is no reason to do accounting or administration of users. Additional security is ensured using IPsec and IKE [29].

Device Mobility — We used the Mosquito Net Mobile-IP stack [32] to enable our devices to do handoffs. Handoffs between GSM-data and WLAN, proved to be unsuccessful due to long GSM-data session setup times, which is fixed by using GPRS.

Naming, Localization, and Session Invocation — A SIP redirect server allows end-uses to register with a SIP URL and enables others to send them invitations to multimedia communication (enabling personal service mobility). Thus, assigning these identities to Personal Agents allowed us to leverage its functionality to easily implement remote customer control of personal messaging (via web pages) and Internet Telephony (e.g. diverting calls when in a meeting), where the voice gateway allows us to locate agents locally or remote as SIP URLs by identifying telephone numbers and vice versa.

B. Mobile Applications

Within this testbed, we have created a Context- & Mobile-Aware Media Player, comprising a Personal (mobile) Agent on a mobile device. Using the eXtensible Service Protocol for the necessary flexible negotiation between entities (agents representing Internet media stations, access network support proxies, sensors, shared spaces, and end-users) the Personal Agent is able to:

1. Play personalized media content which is negotiated with an Internet Media Server and support agents in the access network, which enable the player to deal with the varying communication conditions in the heterogeneous wireless infrastructure, and plan download and playing of media content.

   The Personal (mobile) Agent connects to an Internet Media Station with MP3 content, which in turn diverts communication to a Content Proxy Agent in the access network. The Content Proxy also extends the functionality of the Personal Agent by sending a protocol object for streaming and playing out MP3-audio using RTSP when the user is on-line. Multimedia delivery is redirected to an optimal point of access from a user (price/performance) perspective, based on user context information: e.g., Access Network Agents notify Content Proxies in the access network of available bandwidth, and the Location Agents provide location prediction information, on the basis of which the Content Proxy the Personal Agent decides to receive content in a hot spot with 802.11b WLAN.
2. Take into account user context, i.e. proximity (relative IR-beacons), physical position, and virtual position (relative a shared virtual space), thus play media content based on the user's context, and in addition to this, connect to a shared virtual space (Fig. 9), where users are able to share multimedia objects and be aware of each other's presence.

   Our agents can recognize resource URLs from IR beacons [15] to invoke the automatic playout of multimedia content that was associated with this device. Thus we can attach beacons to various locations at Brf Bågen and demonstrate multimedia that is associated with different locations in that area to visiting mobile users. Furthermore, the Personal Agent can register with

a Mobile Interactive Space (Group) Agent which not only forwards voice messages and real-time communication between registered participants, but it also multicasts their respective virtual positions. Fig. 9 shows the 3D user interface to a scale model of Brf Bågen with on-going communication with another participant. As the Personal Agent can negotiate to receive physical location information (from GPS) and virtual location information (from the Mobile Group Agent), we have created a mix reality system [20], in which visitors visiting the area physically can meet virtual visitors, and vice versa.

VIII. CONCLUSIONS

In this paper, we presented the service architectures that are or will be used in successive generations of wireless networks (2G, 2.5G, 3G Phase 1), that reinvent the telecom model of services, and discussed their limitations in terms of their ability to deploy an open service model, supporting a heterogeneous wireless environment, in which anyone is allowed to use a service by anybody else.

A few key Internet service architectures were discussed, followed by a more detailed account of how these architectures can be further extended and improved for adaptive personal mobile communication, allowing us to use ad-hoc negotiation of services. This was further exemplified by discussing our wireless testbed and illustrated by our mobile- & context-aware media player prototype in this testbed.

Clearly, the service architectures up until 3G Phase 2 have considerable deficiencies according to the criteria that were put forward in the introduction. They make it hard for anyone but the network operator to install a new service, and they are extremely complicated to program and costly (in terms of man power) to develop advanced applications in, in comparison to the Internet Service Architectures. The service architectures that we will have to live with, until 3G Phase 2 or 4G arrives (whichever is first), are thus entirely unsuitable for developing advanced end-point based multimedia communication that is becoming commonplace on the fixed Internet, and will soon be when the equipment manufacturers and application developers have discovered that wireless bandwidth is rapidly increasing already today with the proliferation of wireless LAN. Therefore multi-mode devices (2.5G, WLAN) will soon be commonplace. The wireless industry must deliver optimal Internet access via these devices to third party services, in order to achieve the same success as iMode, and thus avoid 3G going the same way as WAP.

Current Plug & Play service architectures for the Internet are interesting starting points, but they do not scale well in mobile networks. Additionally we need better support for automatically being able to negotiate our communication. We cannot await the establishment of an ultimate protocol that deals with all known capabilities nor one which stifles the system as is the case in the so-called Open Service Architecture for 3G Phase 1. Instead we need an open protocol, which allows entities to obtain service knowledge, and learn about service capabilities. We presented the eXtensible Service Protocol as such a means to communicate metadata about services between entities, allowing them to negotiate services starting with a minimal set of shared knowledge. We concluded by showing the feasibility of building a prototype that provides advanced multimedia communication between users and content in our wireless testbed. We believe that it build an important starting point for further investigations regarding services architectures for adaptive personal mobile communication for 3G Phase 1 and beyond.

It should be added that the limitations in the service architectures are reflected in how far Internet is allowed to make inroads in successive generations of 3G. Therefore we conclude that the wireless industry should go for 3G Phase 2 directly, after installing GPRS, thus skipping Phase 1.

IX. FUTURE WORK

The increased diversity and flexibility in the infrastructure demands that we develop metadata describing resources, user profiles, device capabilities. Furthermore, we foresee that for these new powerful modes of communication we need a better understanding, and support to collect, dissimenate and reason about service knowledge. In the course of 2001 we will further investigate how this can be applied in our service architecture.

X. ACKNOWLEDGEMENTS

This research was conducted with the support from the Swedish Foundation for Strategic Research' Personal Computing & Communication (PCC) research program and Ericsson Radio Systems AB (Research). Many thanks to Prof. Gerald Q. Maguire Jr., Royal Institute of Technology, Stockholm, Sweden, for his comments on earlier drafts of this article.

REFERENCES

1 T. Kanter, “Adaptive Personal Mobile Communication”, Licentiate Thesis, Royal Institute of Technology (KTH), March 2000.

2 T. Kanter, P. Lindtorp, C. Olrog, G. Maguire, "Smart Delivery of Multimedia Content for Wireless Applications", to be presented during the 2nd International Workshop on Mobile and Wireless Communications Networks (MWCN'2000), May 2000.

3 P. Jokela, “Wireless Internet Access Using Anonymous Access Methods”, Proceedings of the Sixth IEEE International Workshop on Mobile Multimedia Communications (MoMuC’99), November 1999.

4 S. Desrochers, R. Glitho, K. Sylla, “Experimenting with PARLAY in a SIP Environment: Early Results”, Proceedings of the IP Telecom Services Workshop 2000 (IPTS 2000), September 10-11, Atlanta.

5 R. Glitho, Advanced Service Architectures for Internet Telephony: A Critical Overview, IEEE Network, July/August 2000

6 R. Glitho and A. Wang, A Mobile Agent based Service Architecture for Internet Telephony, International Switching Symposium 2000 (ISS00), Birmingham. May 2000.

7 M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg - RFC 2543 on SIP: Session Initiation Protocol, IETF/Network Working Group – March 1999

8 WAP Forum (Wireless Application Protocol Forum Ltd.), Proposed 1.1 technical documents, http://www.wapforum.org/docs/technical.htm

9 Ericsson Review 3G Evolution article

10 3GPP, Virtual Home Environment, Technical specification 22-21, http://www.3gpp.org/

11 3GPP, Open Services Architecture, Application Programming Interface, 3G TR 29.998, http://www.3gpp.org/

12 PARLAY Forum: http://www.parlay.org/

13 Internet42: Public Gigabit-Ethernet IP-network connecting housing co-operative BRF Bågen in Stockholm (http://www.fatburen.org)

14 T. Kanter, "An eXtensible Service Protocol for Adaptive Personal Mobile Communication", Accepted for the International Workshop on Smart Appliances and Wearable Computing (IWSAWC'2001) 2001.

15 T. Kindberg et al, “People, Places, Things, “Web Presence for the Real World”, Proceedings of the 3d IEEE Workshop on Mobile Computing Systems and Applications, December 2000, Monterey USA.

16 Universal Plug and Play Device Architecture, Version 1.0, June 2000. – http://www.upnp.org

17 Sun Microsystems Inc., “Jini Architectural Overview”, Technical White Paper, — January 1999.

18 TINA-C Service Architecture Version: 5.0, Date: 18 June 1997

19 White Paper: The SyncML Initiative, http://www.syncml.org

20 H. Tamura: "Mixed reality: Merging real and virtual world," Journal of the Robotics Society of Japan, vol.16, no.6, pp.759-762, 1998.

21 1993 KQML Specification, by The DARPA Knowledge Sharing Initiative External Interfaces Working Group.

22 Knowledge Interchange Format (KIF), draft proposed American National Standard (dpANS), NCITS.T2/98-004.

23 J. Mitola, "Cognitive Radio: an Integrated Agent Architecture for Software Defined Radio", Tekn. Dr. dissertation, Royal Institute of Technology (KTH), 2000.

24 T. Lehman, S. McLaughry, P. Wycko, “TSpaces: The Next Wave", published in the Hawaii International Conference on System Sciences (HICSS-32), January 99.

25 Object Management Group, Inc., CORBA/IIOP 2.2 Specification, 98-02-01

26 World Wide Web Consortium (W3C), Resource Description Framework (RDF) Schema Specification 1.0 – http://www.w3.org

27 World Wide Web Consortium (W3C), XML 1.0 Second Edition, Oct 2000 – http://www.w3.org

28 W. Heinzelman, J. Kulik, H. Balakrishnan, Adaptive Protocols for Information Dissemination in Wireless Sensor Networks, Proc. 5th ACM/IEEE Mobicom Conference, August 1999

29 F. Moioli, "Security in Public Wireless LAN Networks", Master Thesis, Royal Institute of Technology, June 2000

30 E. Wedlund, H. Schulzrinne, "Mobility Support using SIP", Proceedings of the 2nd ACM International Workshop on Wireless Mobile Multimedia (WoWMoM'99)

31 Telia HomeRun: http://www.homerun.telia.com/

32 X. Zhao, M. Baker, "Flexible Connectivity Management for Mobile Hosts", Technical Report: CSL-TR-97-735, September 1997

33 J. Guttag, “Communicating chameleons”, Scientific American, July 1999.