# Research Direction for Developing an Infrastructure for Mobile & Wireless Systems: Consensus Report of the NSF Workshop Held on October 15, 2001 in Scottsdale, Arizona

Birgitta König-Ries[1], Kia Makki[2], S.A.M. Makki[3], Charles E. Perkins[4], Niki Pissinou[2], Peter Reiher[5], Peter Scheuermann[6], Jari Veijalainen[7], Alexander Wolf [8], Ouri Wolfson[9]

1: Universität Karlsruhe, Germany
2: Florida International University
3: Queensland Univ. of Technology, Australia
4: Nokia Research Center, Mountain View, CA.
5: University of California at Los Angeles
6: Northwestern University
7: University of Jyvaskyla, Finland
8: University of Colorado
9: University of Illinois at Chicago

## 1 Introduction

The recent NSF Workshop on Infrastructure for Mobile and Wireless Systems, held on Oct. 15, 2001 in Phoenix had a goal of defining and establishing a common infrastructure for the discipline of mobile wireless networking. This consensus-based paper is the outcome of that workshop. The paper provides a foundation for implementation, standardization, and further research and discussion on the issues of what should constitute such an infrastructure. Workshop participants came from many different wireless communities, including those of communications, operating systems, core networking, mobility, databases, and middleware. The workshop presented various research directions in the field and included substantial discussion on the role of an infrastructure for wireless mobile networking and the desirable components of such an infrastructure. The outcome of the workshop was not a crisp and clear-cut definition of the infrastructure and its components, rather a step towards a better understanding of the infrastructure requirements of the mobile wireless environment.

Not all participants agreed fully on whether particular features and services belong in this infrastructure, but the discussions helped clarify the issues. Another role that this paper serves is to guide research in the area of mobile wireless infrastructure, in part to flesh out the infrastructure requirements all participants agreed upon, and also to cast light upon the areas where no agreement was reached. Relevant funding agencies and companies interested in research in this area should consider these unanswered questions when they define new programs and projects in the mobile wireless area.

B. König-Ries et al. (Eds.): IMWS 2001, LNCS 2538, pp. 1-37, 2002.

Section 2 discusses terminology and principles concerning a wireless mobile infrastructure that were developed during workshop discussions. The rest of the paper is divided into sections describing infrastructure services in horizontal slices, which are: network layer services, transport layer services, and middleware layer services. This division is somewhat arbitrary, but provides a simple organizing principle. The next three sections of the paper describe sets of possible infrastructure services at each layer. The sections are divided into subsections that discuss the requirements of particular kinds of services in that layer, along with arguments for why these services belong in the infrastructure. Concluding this paper is a summary of the consensus recommendations of the workshop participants on what the composition of an infrastructure for a mobile wireless environment should be. Appendix A lists the papers presented at the workshop, and appendix B lists the participants in the workshop.

## 2 Terminology and Principles

The bulk of the discussion in the workshop dealt with the boundaries and definitions of the key terms, and on ideas inherent in the workshop's theme. The fact that a group of leading researchers in the area had so much difficulty agreeing on what constitutes infrastructure, middleware, and wireless mobile systems suggests that this central issue is far from settled. This section captures the core of this discussion and sets the stage for more detailed presentations in individual technical areas.

The scope of the infrastructure to be provided was an important issue for discussion. There are many kinds of wireless networks, existing and proposed. Is the infrastructure meant to handle all of them, and all kinds of mobility? The general agreement was that the infrastructure should handle both wireless cellular networks that rely primarily on single hop communications to a fixed base station that is connected to a wired network and ad hoc networks that might communicate via multihop wireless networks before reaching a wired segment (or perhaps without any participation by wired segments). In the latter category, the infrastructure should at least include 802.11, Bluetooth, Jini and ad hoc IP-based systems currently under development, but ideally it should be flexible enough to handle many other such networks.

There was less agreement on whether the infrastructure being defined here should support sensor networks, particularly those that use diffusion-based methods to transmit their information. The needs of such networks are substantially different than those of more conventional wireless networks. Whether a sufficiently general infrastructure could suitably service both styles of networks requires further research and discussion.

Another major question of scope was precisely what was meant by "infrastructure." The workshop participants wrestled with different definitions. Here are the two most popular candidates:

1. Infrastructure is the collection of system components, including middleware, network layers 1-5, and hardware, that services a large class of applications in the mobile wireless environment.

2. Infrastructure defines a set of assumptions that application developers can make about the components and behaviors of a wireless mobile network (It defines the minimum requirement for the establishment of such a network).

Clearly, either definition suggests that there is a common base of hardware, software, and protocols widely deployed for the purpose of servicing the common needs of many applications.

Another way of grappling with the problem of what constitutes infrastructure for the wireless mobile environment is to think by analogy of what constitutes infrastructure for other network systems, particularly the Internet. By considering the differences of the wireless mobile environment and determining where Internet infrastructure will be insufficient to handle such differences, one can perhaps identify the set of components required for the new infrastructure.

While some of the above definitions mention hardware, the purpose of the workshop was not to address hardware issues. Thus, the remainder of the paper does not deal much with transmitters, routers, antennae, or other hardware. Neither does it much discuss issues of the physical layer of networking, such as waveforms or modulation schemes for wireless networks. Little is said here about the MAC layer, or even the link layer. Nevertheless, this paper does touch upon protocols and software at multiple layers in the protocol stack. One way of thinking about the required infrastructure components is to consider what features are required at each layer. Still a different approach is to decide what high level demands will be placed on the infrastructure, and then to consider what components are necessary at each layer to support those demands.

The workshop participants agreed that the infrastructure must support multiple computing paradigms. In addition to the widely used client/server paradigm, the infrastructure should provide support to the emerging peer-to-peer and agent paradigms of computing.

Much of the infrastructure is likely to be provided by middleware. Like the word “infrastructure,” “middleware” is subject to many definitions. Certainly it implies that the software in question is not compulsory (as, in practice, the use of IP is compulsory in the Internet), but also that the software is ubiquitously available for applications that need it. Middleware should be generally useful. Anything that is useful only to a single or small class of applications can more properly be provided in those applications. Middleware typically acts as an abstraction layer to hide complexities of the system from program developers and users. Wireless mobile networks are rich in complexity, therefore common middleware for this environment needs to address the key areas of complexity that are most commonly used and are most likely to cause difficulties.

Middleware and infrastructure already exist for the wired environment, and one position is that what is good for wired networks is good for wireless networks. There is some truth in that observation, but only to the extent that the two types of networks are actually alike. There are certain fundamental differences between the fixed network environment and the wireless environment. First, wireless terminals exhibit communication autonomy (C-autonomy) towards the network components and other terminals, meaning that they are normally detached from the network from time to time. There are many reasons for this behavior, but the main reason is that the laptops and especially small, portable telecom terminals are highly personal and mediate data and voice communications of individuals who simply cannot or do not want to

communicate all the time. People have the right to choose when to communicate and with whom to communicate over the wireless network. This behavior has a profound effect on the design of the infrastructure and its applications.

The main attraction of wireless communication is that it makes "untethered" communication possible and also allows free movement of the terminal while communication takes place. Thus, an issue for infrastructure is the support for the mobility of the wireless terminals. Roaming - or mobility-in-large - support should be global so that the terminal can have unrestricted movement while still being able to access communication services in its immediate environment. It should also be able to use other services connected to the Internet anywhere in the world, as well as communicate directly with other terminals. Mobility support also requires that the terminal be allowed to move while communicating over a wireless network. This mobility-in-small feature requires hand-over (or hand-off) support from the network infrastructure. These mobile devices also have energy resource constraints, thus making the handover (Hand-off) procedure necessarily power aware. A security problem inherent in all wireless communication environments is that third parties can capture the radio signals while they are in the air. This problem cannot be avoided, because the signals must propagate in all directions from the base stations, terminals, and communicating components of mobile ad hoc networks (MANETs) that support mobility. The only way to protect messages is to encrypt them. Thus, encryption and decryption support are an inherent part of the infrastructure. The infrastructure should also address other security issues, because mobile terminals are more vulnerable to loss or theft than fixed terminals. The Wireless Public Key Infrastructure suggested by WAP Forum [83] is one possible solution to many security issues in wireless cellular networks. For MANETs, the security issues are largely open. The Wired Equivalent Privacy (WEP) for 802.1x standards, which was considered to be insecure, has been improved recently. The improved WEP no longer has a shared key for the packets; instead it now has a unique key for every packet. This new WEP has not been cracked at this point, but also has not withstood the test of lengthy attempts to find new flaws [65].

The telecom industry estimates that in a few years there will be 1-2 billion wireless terminals in the world, of which hundreds of millions will be "Internet-enabled", i.e. can access data in Internet. The infrastructure must therefore be highly scalable.

The development of mobile terminals has been very rapid. In general, this development can be seen as finding a compromise between portability and usability. Portability increases with the decreasing size and weight of a gadget, and it decreases with the number of gadgets. Usability increases with the size and weight of a device, but the particular application determines the optimal size of the device (watch-size video-phones are realistic, but taking notes with them is difficult). The majority of wireless terminals will be resource-poor (small energy resources, displays, keyboards, memories) due to the small physical size and weight, at least compared to their non-mobile equivalents. Wireless links will similarly have less capacity than fixed-network links at any particular point of time. Thus, while the infrastructure should handle increased power of mobile devices and wireless networks, disparities in relative power between mobile and non-mobile components, and between wireless and wired networks must be expected to persist.

Although it seems that the future telecom networks and compatible terminals will be most important economically, there are other wireless environments emerging.

These include MANETS and sensor networks. Wearable or ubiquitous computing and personal area networks (PANs) can also be connected to a larger infrastructure. Bluetooth [7] and JINI (a typical MANET technology) have already reached the marketplace, embedded in hand-held terminals, PDAs, and also fixed devices, like printers, air conditioners, doors, and cash registers; cars are likely to soon follow. These networks typically configure themselves whenever and wherever necessary, without base stations or other central components; and with the allocated radio spectrum used in a genuine broadcast mode. The maximum transmission distance of two components in these networks is typically 1-10 m. The infrastructure must be able to recognize that these diverse wireless networks can function as an access network of personal communications to wireline backbone networks, or can feed data into the nodes of a wireline network (temperature sensors, "health" sensors, movement control sensors, etc).

Although the GSM infrastructure already provides roaming, mobility, and security (there are over 700 networks in over 170 countries) and has proved to be feasible and scalable, it is dedicated for wireless voice traffic. It does not work as well with best-effort packet delivery. While operational and relatively mature, the GSM infrastructure still requires improvement. For example, QoS in GSM data or Short Message Service is not well defined. The GSM infrastructure needs to be more reliable, beyond making handoffs somewhat more transparent. Furthermore, GSM is going to be augmented by new technology (3G, 4G) thus making research on more general infrastructure issues necessary. The recent development in the marketplace seems to indicate that the global wireless world is moving towards an open mobile environment based on open communication and contents standards [50] and [48].

Given these directions and differences, what are the desirable properties of an infrastructure for a wireless mobile environment? The workshop participants discussed many important properties, but decided that five of them were key to successfully defining a good infrastructure for a wireless mobile environment:

1. The infrastructure should be minimal. The less burdensome the infrastructure, the more likely that providers will deploy it in its entirety. Furthermore, making the infrastructure smaller increases the chances that its implementations will be correct. Finally, if infrastructure features are only added when deleting them will inconvenience most users, there is less chance that infrastructure providers will have to pay a high cost to support largely unused features.
2. The infrastructure must be complete. While the Internet is largely a success, assumptions were made in the definition of its infrastructure, which are no longer valid. That infrastructure lacks security, for example, and is difficult to manage. There are many useful features (such as multicast and quality of service guarantees) that are hard to provide within the constraints of the Internet infrastructure. The mobile wireless infrastructure should include those things whose absence from the Internet has been badly missed. The infrastructure designers should similarly learn from the experiences of other communications networks. One common lesson from telegraph, telephone, television, and the Internet is that adding new infrastructure features after the system has become popular is very difficult, so researchers should expend great effort to foresee any needs now.
3. Infrastructure features should be generally useful. In some ways, this is a corollary to property 1. Features should not be added to the infrastructure to service one

application, one provider, or one manufacturer. Only things of general use to the entire community belong in the infrastructure. Studying the commonalities of applications likely to flourish in the mobile wireless environment should help achieve this property. Finding the right models for interaction will help.

4. The infrastructure should be secure. As property 3 is a corollary of property 1, this property is a corollary of property 2. Security is perhaps the greatest deficiency of the Internet infrastructure, and everyone on the Internet now suffers for that mistake. This mistake should not be repeated in new mobile wireless infrastructure, particularly because that environment adds new security challenges not commonly present in the Internet. A secure mobile wireless infrastructure will never be achieved solely by including some security features in the infrastructure, such as cryptography and authentication. Rather, all infrastructure components must be designed with security in mind. In addition, the security of their interactions must be considered.
5. The infrastructure must be compatible with a realistic economic model. Someone must pay for the infrastructure. Any service to be provided will require some amount of resources, and someone will need to provide them. If no one is likely to actually pay the money to provide the resources, that infrastructure component will never actually be deployed. In some cases, services can be spread across all participants, with each paying a trivial amount to provide the collective service. In other cases, some entity can make a profit or improve their business model by choosing to provide the infrastructure service. Still, generally someone must pay, and infrastructure designers should keep that requirement in mind.

There are other desirable properties for the wireless mobile infrastructure, such as scalability and reliability. The five listed above are those that to the workshop participants seemed the most key and the most likely to be overlooked if not emphasized.

# 3 Network Layer Infrastructure

## 3.1 Alterations to IP

Because mobile computers using a wireless infrastructure will want access to the same services as wireless computers, they will need to interoperate with the Internet. However, the Internet's critical protocols do not handle mobility well, especially the fundamental network layer protocol, the Internet Protocol (IP). IP provides end-to-end delivery of datagrams between devices. To achieve this, IP requires that routers forward packets using routing tables indexed by the IP addresses of the destination devices. These tables must be of manageable size, so IP addresses with the same prefix are aggregated in these tables.

Aggregation is vital to achieving scalable router tables, but it requires that large blocks of addresses be reachable by the same path, since the router tables associate address blocks with path components. Device mobility works against this requirement, since a device with any IP address prefix could pop up anywhere in the world. One way to overcome this problem is for a mobile device to obtain a new IP

address at each location it visits. However, higher-level protocols (particularly TCP) assume that a device does not change its IP address throughout a communication, so changing IP addresses every time a device moves would require shutting down all communications before moving and rebuilding them afterwards. Furthermore, the Domain Name Service (DNS), which provides bindings between more human-meaningful names (like www.ieee.org) and IP addresses (e.g. 140.98.193.38, the actual IP address for www.ieee.org) works on the assumption that mappings change infrequently. Solutions that frequently change a device's IP address would also require a redesign of this core Internet service.

Mobile IP [55] provides smooth mobility without breaking existing Internet components. The basic idea behind Mobile IP is to provide care-of-addresses for mobile computers. Whenever a mobile device moves to a new location, it informs its home agent (an entity in its home network) of its new location. Packets for the mobile node will first be routed to its home address, where they will be intercepted by the home agent. The home agent then resends these packets using encapsulation to the care-of-address of the mobile node.

Mobile IP was originally developed as an outgrowth of IPv4. The newer IPv6 [18] allows a redesign of Mobile IP. In Mobile IPv4, packets are sent from the home agent to the mobile node using tunneling. In Mobile IPv6 [34], packets can be sent to the mobile node directly, without traveling to the home network (thus typically not needing the services of a home agent). Mobile IPv4 care-of-addresses are often the addresses of a foreign agent in the visited network and can be shared by a number of mobile devices, while Mobile IPv6 care-of-addresses are always collocated on the mobile computer, i.e. each mobile computer has its own care-of-address at each location. IPv6's tremendously expanded address space (128 vs. 32 bits) allowed this improvement. Another enhancement in version 6 is considerably enhanced security. In IPv6 all nodes must provide strong authentication and encryption features, a feature Mobile IP can be build on. Nevertheless, security is one of the main topics still being addressed in the further development of Mobile IPv6. In particular, key distribution for secure update to care-of address information is receiving a great deal of attention. A new technique (return routability) has been developed that enables establishment of enough security data between end points to verify that the IP addresses are not being used for any denial of service or masquerading attacks. This new technique indicates a much brighter future for Mobile IPv6 and along with it mobile commerce.

The wireless infrastructure must provide IP support. The main question is whether it need only support Mobile IPv6, or whether it should also support Mobile IPv4. While some parts of the world have enthusiastically embraced IPv6, in other areas (significantly in the United States of America), adoption has been slower. Any infrastructure designed for the near future must therefore support both Mobile IPv6 and Mobile IPv4. Even if IPv6 comes to dominate the total population of Internet wireless mobile computers, IPv4 will nevertheless remain important, running on many millions of network nodes for the foreseeable future. This concept opens up interesting avenues for applications in a mobile environment. While the problem of resource management could be difficult to handle, new enhanced applications on the device could be designed and developed. An example of such an application is a Smart Grocery Assistant. A device could come into the store and get a new mobile IP from the local network and use the services it provides.

### 3.2 Routing Protocols for Mobile and Wireless Infrastructure

The Internet uses several routing protocols to build the routing tables mentioned in the previous section. The existing routing protocols (BGP, OSPF, RIP, etc.) are designed for fairly static situations where changes tend to be caused by failures, rather than by mobility. The use of Mobile IP finesses the issue of updating routing tables when a single mobile device moves from place to place, but it assumes that infrastructure routers are available nearby the mobile environment. This assumption is true for environments where a single wireless hop takes a packet to the wired infrastructure, but is not necessarily true for some ad hoc wireless environments that must operate without the assumption of fixed base stations. In ad-hoc networks with dynamically changing topologies, the classical approaches to construction and maintenance of routing tables do not work well. Thus, the development of new routing protocols has become necessary.

The IETF MANET working group [43] covers most of the ongoing standardization effort in this area. Within the group, a number of proposals for routing protocols in ad-hoc networks have been developed. These protocols can be divided in two main groups: table-based routing protocols, and demand-driven routing protocols.

Table-based ad hoc routing protocols like DSDV [57] and OLSR [12] are adaptations of classical routing protocols. Each node stores a routing table whose entries contain the interface used to reach each destination node or sub network and some measure of the distance to the destination via that link. Reachability and distance change more frequently in ad hoc networks, so these protocols include mechanisms to cope with those differences. One such optimization is to use a cluster-based protocol. In fixed networks, routing tables typically do not contain the address of each individual node, but summarize nodes in the same part of the network (i.e. with the same prefix of the IP address) into one entry. In ad-hoc networks, a similar effect can be reached by clustering nodes. For each group of nodes, one dedicated node, the cluster head, is determined to be responsible for all communication and routing outside of the cluster. The cluster head also routes information to destination nodes in the cluster. All the cluster heads together form a backbone of the ad-hoc network. However, it is not yet clear under which circumstances the overhead for hierarchy maintenance provides an overall savings in bandwidth utilization.

Demand-driven protocols are the other major alternative for routing in ad hoc wireless networks. Such protocols do not aim at storing complete routing information. Instead, whenever a message needs to be sent from one node to another, a route is solicited. The easiest way to achieve this is to simply flood the network, i.e. broadcast the message from the original node to all the nodes in its radio range, have all of them rebroadcast the message, and so on. When the message reaches the destination, it carries enough information to determine a reasonable route from the source, which will be used by subsequent messages between the two nodes. This solution is not very efficient, so a number of protocols have been developed that reduce the message overhead incurred by selecting appropriate nodes to which messages can be forwarded and by caching information about known routes. Examples for this type of protocol are DSR [33] and AODV [56].

No consensus has been reached on which ad hoc routing protocols are best, nor on which style is better. If the wireless infrastructure being discussed here is an analog of the infrastructure that supports the Internet, routing protocols play a central role.

But the lack of consensus on ad hoc routing protocols suggests that the best option for defining the infrastructure today is to avoid choices that may preclude the routing protocols that are eventually shown to be best. For this area, flexibility (with the assumption that eventually a more definite choice will be made) is the safest choice for the infrastructure.

The infrastructure should be agnostic to the choice of routing protocols, but ensuring this is true may prove difficult. For instance, certain choices related to security may restrict the flow of information that some routing protocols require. Recent work has provided a better understanding of how ad hoc networks should be connected (when possible) to the fixed Internet infrastructure. It is an interesting question whether mobile gateway nodes between an ad hoc network and the Internet should have authorization or responsibility for changing the routability for all nodes within the ad hoc network. Thus, the best that is currently possible is to consider carefully whether choices made about other aspects of the infrastructure affect ad hoc routing protocols. This issue should be revisited when the network research community has achieved better understanding of ad hoc routing protocols.

### 3.3 Multicast Protocols for Mobile and Wireless Infrastructure

Future mobile and wireless networks are expected to support group-based communication such as teleconferencing, multimedia, collaborative work, real-time workgroup, and distributed database access. Additionally, several fundamental services in mobile and wireless networks are likely to be built upon multicasting, especially in light of plans for widespread deployment of IPv6. These applications are usually characterized by their large bandwidth requirements, stringent delay bounds, and multi receiver connections. Therefore, to support these types of applications, efficient multicast routing algorithms for mobile and wireless environments must be developed [42].

The unprecedented growth of the Internet community has created a strong demand for a new class of services, particularly those providing means for groups of users to collaborate and share information over mobile and wireless networks in an efficient and real-time manner [2]. These pose several challenges if efficient multicast protocols are to be provided. A multicast routing scheme should reduce the hand-off latency and optimize the multicast tree for stable regions that do not experience frequent changes in group dynamics. It should handle frequent join and leave requests efficiently, without disturbing the ongoing multicast connections.

Multicasting in a mobile and wireless network is substantially more complex than in a purely wired network, because the mobile and wireless environment adds several twists to multicasting in wired environments by allowing for node mobility and by low-bandwidth, unreliable wireless links, etc [37]. The importance of reliable multicast protocols has been gaining recognition in recent years [13,51].

There are several difficult scalability challenges associated with reliable multicasting. Reliable multicast protocols operate on multicast delivery trees constructed by multicast routing protocols. They ensure reliable end-to-end delivery of unreliable multicast datagrams for group members. It is a challenge for reliable multicast protocols to satisfy the simultaneous requirements of efficiency and scalability while ensuring reliability. One problem is feedback implosion. As the

number of group recipients grows, the number of feedback messages increases dramatically. This leads to a heavy burden on data sources, and causes more severe congestion and packet losses. Reliable multicast in ad hoc networks adds the dimension of host mobility within the scope of reliable multicast [32]. Reliable multicast in the mobile and wireless environment deserves serious attention.

Another challenging problem in this area is to provide QoS in mobile and wireless multicast [82]. Due to the dynamic nature of such networks, guaranteeing QoS is a difficult task. Most multicast protocols still do not provide support for QoS.

### 3.4 Content-Based Networking

Multicast is appropriate for many group communication scenarios. There is, however, an emerging alternative to multicast that is more flexible in its segmenting of the message space, and can be less complex to deploy. This alternative is called *content-based networking*.

A content-based network is a communication network based on a connectionless service model [11]. The idea is to deliver messages to a host based on the interests declared by that host, rather than using a destination address specified by the sender. The term "content-based" refers to the fact that a host's interests are expressed by conditions over the entire content (or payload) of messages.

We can describe a content-based network by comparing it to a traditional connectionless, address-based network like the Internet. The traditional network relies on an addressing scheme whereby hosts (or interfaces) are assigned a unique network-level address, and messages are sent to specific network-level destination addresses. The network service model consists of delivering each message to the host associated with the destination address of the message. By contrast, a content-based network does not use explicit addresses. Each host is associated with a *receiver predicate*, which is a logical expression over message content that defines the messages that the host intends to receive. Messages are simply injected into the network. The network service model consists of delivering each message to all the hosts whose associated receiver predicate matches the message.

The service model of a content-based network is motivated by the needs of several classes of distributed applications. It grew from a scalable publish/subscribe event notification [9,10]. However, content-based networking is not limited to wide-area publish/subscribe middleware, and can be exploited by numerous other applications. These include distributed auction systems, ad hoc information sharing communities, distributed multi-player games, personalized news distribution, service discovery, and sensor networks. Note that what all these applications have in common is a style of communication in which the flow of information---from senders to receivers---is more naturally determined by the specific interests of the receiver, rather than by explicit knowledge of one or more destinations by the sender. In general, the content-based service model is particularly well suited to support applications that require seamless, many-to-many communication among loosely-coupled, autonomous participants.

As mentioned above, a content-based network is more flexible than multicast. This is due to the fact that receiver predicates are not limited to a strict partitioning of the

message space, as are multicast groups. This is important because, although for every individual receiver's interests there exists a mapping of information to multicast addresses satisfying that receiver, there is no general mapping that satisfies all receivers.

As it turns out, there are two opposing strategies for associating multicast addresses to the interests of receivers: one could either define a large number of specific multicast groups, or one could define a small number of generic groups. Both solutions have significant limitations. With specific groups, receivers would be able to select information of interest with high accuracy, but at the same time senders would be forced to send to multiple groups whenever they produce information that spans multiple specific selections. Moreover, the multicast routing infrastructure would have a hard time efficiently serving a large set of very sparse groups, and would have a hard time dealing with highly dynamic changes in interest that would lead to highly dynamic restructurings of the groups. The case of a small number of generic groups has opposite advantages and disadvantages: senders could send to a few groups and the multicast routing infrastructure would benefit from a lower number of dense groups, but receivers would receive, and therefore would have to process (i.e., filter out), a large volume of uninteresting information.

Content-based networking has another important advantage that makes it an attractive alternative, especially in constrained environments, and in the absence of an underlying network layer. This advantage is given by the absence of addresses in the service model. In principle, a content-based network does not use network-level addresses, and therefore can be deployed with minimal global configuration. In practice, addresses are necessary, since some form of name is needed to identify nodes, and some form of link-level address is needed to establish direct communication between nodes. However, neither one is used for routing, so addresses can be reduced to physical addresses and can be maintained locally, while names need only be unique identifiers. In particular, the advantage is that the assignment of names or addresses to hosts does not need to obey any hierarchical structure to mimic the physical topology of the network, which in turn reduces the need to establish a coordinated configuration. This gives content-based networking a significant administrative advantage, and makes it suitable for ad hoc network environments and other kinds of emergent networks in which establishing appropriate network-level address configurations may not be practical or cost effective.

At the architectural level, a content-based network is very similar to an address-based network. A content-based network transports messages from senders to receivers through an interconnection of routers, each one performing a *content-based routing* function and a
*content-based forwarding* function. The content-based routing function is responsible for propagating receiver predicates throughout the network, so as to establish (loop-free) forwarding paths that realize the content-based delivery model. As a result of the content-based routing function, each router compiles and maintains a forwarding table. The forwarding table is essentially a map between neighbor routers or hosts and logical conditions, and it is used by the forwarding function to determine where to forward each incoming message.

Note that content-based networking is not intended as a replacement for multicast or unicast address-based networking. Instead, its content-based service model is intended to provide a complementary, value-added communication service, possibly

well integrated with existing networks. In fact, for practical reasons, initial prototypes should be implemented as application-level overlay networks.

In summary, the main benefits of content-based networking are its advanced ``content-based'' service model and its lack of network-level addresses, which in turn leads to an inherently self-configurable network topology. When the interests of receivers can be easily partitioned, and when that partitioning remains relatively stable, then multicast offers an appropriate communication service. But when more flexibility is necessary, particularly in a dynamic or ad hoc situation, then a content-based network could offer a better approach.

It should be clear, however, that there is no magic in the concept of content-based networking and, not surprisingly, its attractive features introduce costs not found in traditional networks. For example, the forwarding function may have to match the entire content of a message against large sets of predicates---certainly a more time-consuming task than matching address prefixes in a traditional router. Also, the routing function must deal with predicates instead of network addresses, and in doing so it must compensate for the lack of imposed structure in the distribution of predicates over network nodes. Despite these and other conceptual and technical obstacles, we believe that content-based networking represents an interesting architecture that could provide a valuable communication facility.

### 3.5 Other Network Services

One of the beauties of the Internet is that it provides tremendous utility while offering relatively few services. Thus, there are not many services beyond the basic network layer protocols and routing protocols that could require alteration. Transport layer protocols are discussed in section IV. One other key Internet service that should be considered is the Domain Name Service (DNS). DNS is a key piece of glue that allows translations of names familiar to human users to IP addresses managed by routers. If the wireless infrastructure uses a routing solution akin to Mobile IP, neither mobility nor the use of a wireless network should change the relationship between a name and an address. Intermittent connectivity of a mobile device does not mean that the mapping between its name and IP address should change.

DNS operates well at the Internet scale because of its hierarchical nature and because of caching of intermediate results. In this respect, some adjustments for the special circumstances of wireless networks and mobility may be useful. Altered protocols to resolve names could allow fewer messages between battery-constrained mobile nodes, for example. With luck, only changes in how DNS is used will be needed, rather than fundamental changes in its design.

Other services have been introduced into the Internet over the years. Many are layered on top of IP. Others are provided via IP options or through augmentation of routers. Examples include multicast, provisions of quality of service guarantees, secure IP (IPSec), and queuing disciplines applied at routers. To the extent that they rely on standard IP services that can be equally well provided by Mobile IP, these advanced services will work in the wireless arena. However, they may be built on assumptions that are not true in a mobile wireless world. For example, an optimal multicast tree may quickly become sub-optimal if the constituent nodes move around a lot, requiring excessive amounts of Mobile IP forwarding. On the other hand, a

quality of service guarantee made when a node was in an area with ample free wireless bandwidth may need to be adjusted when the node moves into an area with less free bandwidth. In many cases, these services enjoy only moderate popularity, which has the advantage of making it easier to change them for the wireless environment, if necessary. Defining a new wireless infrastructure also offers a chance to prune out any failed services that were introduced at some point into the wired Internet. Critical thought should be applied to each service that has at some point made its way into the Internet infrastructure to decide if its value warrants supporting it in the wireless infrastructure.

### 3.6 Network Management Issues

The Internet lacks sufficient infrastructure facilities to allow configuration, monitoring, and control of the network. Relatively little functionality is reliably available to perform these services except in single local networks. The wireless mobile environment will be more dynamic and difficult than the wired Internet, and even simple single-hop models of a wireless network add further complexity. Thus, we require more management functionality then any wired environment for handoffs and other complex issues [59]. Proper infrastructure services will be vital for network management.

To the extent that ad hoc networks, personal area networks, and ubiquitous computing networks are included as being part of the mobile wireless world (and not considering all of these possibilities is an unrealistically simplistic view of the world of tomorrow), these networks will require new management solutions that can handle much more complex situations, such as fault tolerance [58], self initialization, and high dynamism. In many cases, researchers in specific sub-areas are tackling these problems themselves, but because their solutions are usually expected to interoperate with the greater network world, and with each other, there is a role for a common set of network management services provided by the infrastructure. For example, there is work on different management protocols for ad hoc management (ANMP), and work on making ad hoc networks fault tolerant. The wireless mobile infrastructure requires support for these new capabilities that is compatible with the network management protocols used for wired networks, such as SNMP and CMIP.

Some candidate management services (all of which would be useful today) for the infrastructure could include:

- adding and removing mobile nodes into a network,
- determining the status of particular nodes,
- probing routing tables and other shared network data structures,
- determining link conditions at various points in the network,
- interfaces to security features, such as an ability to filter or trace attack packets,
- diagnosis tools to pinpoint network problems,
- ability to configure and manage network overlays,
- managing the network based on the resources available in particular nodes (power, processing speed, available memory, etc.)

Determining the proper set of network management services to include in the infrastructure is a question for research. Choices made for other infrastructure

components will impact the choice of network management services. Most other components under consideration require either management of their interactions with other components, or internal management of their own operations, if not both. Thus, defining the proper set of network management services for a mobile wireless infrastructure must be done in conjunction with defining the other elements of the infrastructure.

### 3.7 Adaptation Services

Networks supporting mobile wireless use often have links and devices with limited capabilities that are not suited to normal data flows. For example, a wireless link can have too little bandwidth, or could fail to provide sufficient security. Likewise, a mobile device's battery may be low, requiring special treatment. Various adaptations can be made to data flowing over the network to transform the data stream into a form appropriate for current conditions. Support for these services could be provided in the infrastructure for a mobile wireless network.

In its simplest form, such an infrastructure could consist of proxy nodes at the boundary between the wired and wireless network [25]. In its most complex form, it could include active networks [74]. There are many points in between, and research is required to determine how powerful and completely deployed the adaptive infrastructure services should be.

Proxies are single nodes designed to provide services to mobile clients with limited capabilities. The popularity of this technology suggests that it is the least that would be required for an adaptive infrastructure service. Proxies have varying capabilities, including transforming data to match the display of the portable device, performing compression or encryption, filtering the data, or scheduling transmissions. If the proxy is to be a widely used service for networks with large numbers of nodes, the proxy can be supplied by a powerful, highly available cluster of machines [26].

Other simple versions of an adaptive service include protocols with adaptive capabilities [5], single link services [22], or gateway services [39]. The various alternatives are captured in a general conceptual framework created by several of the leading researchers in this area [4].

Assuming that the infrastructure supports peer communications between wireless nodes, at minimum there can be two troublesome links connecting to a relatively trouble-free wired network. Proper handling of such circumstances requires some cooperation between the adaptive services near the endpoints [63]. If one considers multihop wireless networking, the reality of troubles in the wired network, or other network complexities, adaptations must be chosen based on varying conditions and must be placed at various points in the network to achieve the best possible behavior. These latter models of adaptivity require more sophisticated infrastructure that deals with issues of security [15], reliability [86], and composition of services [16].

# 4 Transport Layer Infrastructure

## 4.1 Alterations to Transport Protocols

A transport service in the OSI sense offers a reliable end-to-end connection-oriented transfer of data between endpoints. Thus, the upper layers in the protocol stack can view the connection offered by this service as a pair of one-way pipes that transfer chunks of data from one endpoint (host) to the other, without any errors. Transport services also offer (global) addressing mechanisms for hosts. Other aspects of the error-free data delivery are that the service offers flow control, buffering, and disassembly and assembly of the data into/from packets used during transmission. The flow control aspect of transport services ensures that data is not lost due to differences in the processing speeds of the hosts. If the receiving endpoint is not able to process the data fast enough, the sending party is asked to slow down or to stop sending new data until the receiving end is again able to accept new data.

The abstract transport service can be implemented using various existing transport protocols, such as the Transmission Control Protocol (TCP) [53]. TCP was originally designed for the Internet, and assumes IP as the network protocol and wired physical links at the bottom of the stack. Checksums or other additional bits are used to detect and correct corrupted bits.

TCP is not perfect for wireless networks because it makes assumptions about the behavior of the underlying packet network that are not true in the wireless case. If the sending host notices that some sent packets are not positively acknowledged, it concludes that they are lost due to network congestion or corruption. This decision is made based on a transmission time-out value. This value is of crucial importance for the performance of the individual connections, as well as for the whole network; if it is too small, unnecessary retransmissions are generated; if it is too large, the applications at both ends tend to stop while awaiting retransmission of lost or corrupted packets. Because the situation can vary during the single connection, the parameter is adjusted based on round-trip delay samples. Because wireless links and terminals tend to be slower than PC-level terminals accessed through a wired network, this parameter should be larger for connections using wireless links and terminals.

A retransmission event also launches congestion avoidance measures: “When a TCP experiences a retransmission timeout, it is required by RFC 1122, 4.2.2.15, to engage in “slow start” by initializing its congestion window, cwnd, to one packet (one segment of the maximum size)” [52,2.2]. So immediately after a timeout, each packet is ACKed individually, slowing down the transmission speed on the single connection. The window then grows with time. In this way, TCP eventually senses the maximum capacity on the path.

This makes sense in a packet network where the links are reliable, because it helps mitigate network congestion. The same phenomena can, however, occur due to errors on the wireless link; the corrupted packets are not acknowledged by the receiving host and should thus be retransmitted. Fixing this problem quickly calls for fast retransmission of the corrupted packets, not a backoff that slows down further transmission, which is how TCP behaves in the face of this behavior. If the sending

host has no way to determine what the real trouble is (congestion or transmission errors), it cannot react properly. This is the situation with basic TCP [17].

Modifications to TCP of relevance here include the Selective ACK (SACK) option [44,24], and the Congestion Control [3]. Using the first option, in an ACK, the receiver can indicate which blocks are missing from the receiving window. Thus, the sending host only needs to retransmit those missing packets. This saves network capacity, but the sending host is still not able to determine why the packets must be retransmitted, whether they were corrupted or lost. The latter modification suggests that each packet received out of order generates a duplicate ACK and requires sending ACKs for at least every second received block. The latter proposal [3] also discusses starting the transmission after a longer idle period. These proposals have relevance for mobile terminals relying on wireless links because they help distinguish between congestion and corruption, as well as save bandwidth on the wireless links. Still, these solutions have been shown to have flaws [61]. It has been shown that despite the enhancements, TCP is not yet well-suited to the wireless world. A link layer to perform adaptive forward error correction can be a definite enhancement.

Another problem remains: a connection built over diverse links can exhibit both problems (corruption and congestion) at the same time, in which case there is no known solution for the problem combination, even if the sending host could determine that they both occur simultaneously.

This problem has led to a suggestion [8] that an end-to-end TCP connection that uses wireline and wireless physical links when transmitting data should be composed of two separate TCP connections, bridged at a Performance-Enhancing Proxy (PEP). One connection is over the error-prone wireless link, and the other is over the wireline links(s). The former is aware of the fact that the link used is wireless, and the latter can assume a wireline network to be used with its typical behavior. The two connections exchange data at a (transparent) gateway, and the gateway must be able to break the connection into two parts and manage them correctly. The addressing between end-systems must not change. There are several critical points against this suggestion in [17]. One central point is the end-to-end argument, that data must not be acknowledged unless it is, in fact, received. This condition is very difficult to maintain with existing PEP designs.

This arrangement makes it possible for wireless links to use a completely different TCP– or at least use a different “profile”. One can also design different transport protocols or profiles for different wireless links (WLAN, 2G, 3G links). At least the timers can be adjusted in an appropriate way for the transfer speeds in wireless and wireline networks.

It has been known that TCP is not an optimal transport protocol for multimedia data such as video and audio streams, which have various real-time requirements. Therefore, several streaming protocols have been suggested. Video and voice streams require real-time end-to-end delivery guarantees to work properly. A transport protocol that supports streaming must guarantee two separate but related things: the transfer speed must be high enough in average between the end systems so that the data stream can be replayed correctly and without interruptions at the terminal; and the variance in transfer speed must be small enough to make the need for buffers small and initial latency small.

A Transport Protocol for Real-Time Applications (RTP) has been designed for the Internet environment to provide a streaming service for real-time streams, like audio

and video [68]. There is also a control protocol for streams called the Real Time Streaming Protocol (RTSP) [69]. The former is not a closed specification but can be enhanced by different media types by describing payload formats. There are currently several of them for voice and video, including RTP Payload Format for MPEG1/MPEG2 Video [29], for MP3 [23], for MPEG4 [36], for DTMF Digits, Telephony Tones and Telephony Signals [70], ITU G.722.1 audio signals [41], and for PureVoice(tm) Audio [45].

When IP networks are deployed as backbone networks, and wireless terminals also become IP –enabled, the above protocols become interesting for various network operators. For example, NTT DoCoMo is currently offering video calls in its 3G network (called Freedom of Multimedia Access, FOMA), operational since October 2001. They use MPEG4 as the video payload format for the RTP, as specified in [36]. The audio format used is AMR (Adaptive Multi-Rate) developed by 3GPP [1]. The service is called V-Live and was launched commercially in the spring of 2002 [49]. Although the above real example shows that the streaming protocols developed for Internet can be applied in a wireless network, there are still problems with the future infrastructure. RTP, for instance, does not attempt to provide end-to-end real-time guarantees, but rather assumes that the lower network layers provide them. Therefore, layer-by-layer Quality-of-Service (QoS) is an important issue.

### 4.2 QoS Support

In the communication context, Quality of Service (QoS) refers to four things: transfer capacity (or bandwidth), jitter, delay of the communication path, and bit error rate (BER). Transfer capacity is measured by bits/sec transferred between hosts or between two components (routers). For batch data transfer this property of the network is often of secondary importance, unless delay really matters. For streams though, end-to-end transfer capacity is vital. Should the transfer capacity on a path (between client and server) fall below the speed that is required by a continuous presentation of the stream, the presentation is stopped for that period of time. This is very bad for the perceived quality. Applications could ask for a sufficient transfer capacity for a stream and the network should guarantee this. Assuming that the stream can be transmitted with a constant transfer rate R, the capacity offered by the network should be roughly R over the entire path on average. The less the buffer space available at the receiving end, the less the real capacity can deviate from R. The more buffer space and the longer allowable preloading time at the receiving end there is, the more average capacity can deviate from R.

End-to-end jitter is also important for stream QoS. Jitter is the variation of the transmission speed over time and is typical in packet networks that use independent routing of individual packets. It can be measured as the standard deviation of the transmission speed. One can eliminate the impact of jitter to a certain extent by using larger buffers that are filled when there are a lot of packets coming, and then emptied whenever fewer are coming. Jitter and transfer capacity are related in a tricky way. If there is a link (e.g. a wireless link) on the path to the terminal (or sink) whose maximum capacity is slightly larger than the required R, jitter decreases the capacity below R. This happens because jitter causes idle intervals for the wireless link and during the busy periods it is not able to fill the buffers at the receiving host.

For traditional host-to-host data transfer, BER is very important because the data (e.g. a file) should arrive exactly in the same form as it was sent. Bit errors are detected and corrected using additional redundant data. Correction is usually based on Forward Error Correction (FEC), which adds enough redundancy before sending the data to allow the receiving host to detect and correct bit errors to an acceptable level. FEC keeps the BER at a pre-specified level, provided that the raw bit error rate on the path does not exceed the highest expected value. This method is used on wireless links e.g. in GSM networks, because they tend to be rather error prone. More generally, FEC is used for data streams with stringent end-to-end real-time requirements, because resending corrupted data would be unpredictable and probably takes too long, halting its presentation at the receiving host (video or audio). Besides, quality of audio and video streams suffers little from a few bit errors here and there, or even from some missing data packets, in contrast to alphanumeric data, whereas stopping the stream deteriorates the video and/or audio quality very badly.

Required BER could be a QoS parameter because different streams have different error tolerances. Furthermore, in a wireless environment, BER can vary abruptly and greatly during a single stream transfer (varying signal strength due to various factors like shading, multi-path effects, rain, etc). The required BER for particular applications should be known by the wireless (and perhaps wireline) infrastructure and it could also be statically, or even dynamically, adjusted.

Capacity is a typical QoS parameter that should be fulfilled at every point on the path the stream is passing. This is a difficult requirement for IP networks, where the network components like routers are not actively policing individual streams or connections and individual packets of a stream might follow different routes. The Multiprotocol Label Switching (MPLS) Architecture [31] has been proposed to help provide end-to-end capacity and other QoS guarantees. MPLS makes it possible to attach a label and a specific route to (IP) packets and also introduces precedence or service class concepts. The latter have a role in offering end-to-end real time guarantees for data streams. During the last couple of years, MPLS has been successfully deployed within proprietary IP networks of some telecom and Internet operators. It remains to be seen whether the approach is suitable for the “common” Internet, and whether all QoS issues can be reasonably tackled with it.

# 5 Middleware Layer Infrastructure

## 5.1 Service Discovery

Mobility and wireless networks lead to frequently changing environments. Thus, we cannot rely on the user or the computer to know which services are available in the network, where they are located, and how they can be accessed. To a certain degree this is true even in fixed networks, however, due to user movement, the problem is much more pronounced in mobile networks, where the user (and the computer) will be much less familiar with the environment. This environment will also change much more frequently than in the fixed network case. Consider for example a speaker at a workshop. Ideally, she would be able to use a notebook to control different display

mediums in the room, like maybe a whiteboard for her presentation, a second notebook for a video, or the monitors of the audience's computers, or also to control different systems in the room, like lighting, heating, microphones, and to access the network. All this should be possible without her needing prior knowledge of the environment. Other examples include, computations offered by more powerful computers, information about nearby real-world services (e.g., ATMs, restaurants, etc.).

Thus, what is needed is an infrastructure component that enables computers to find services in an unfamiliar network. This need has been recognized by researchers both in academia and industry, with a plethora of different service discovery architectures being proposed over the last few years. Typically these architectures consist of a dedicated directory agent that stores information about different services, a set of protocols that allows services to find a directory agent and to register with it, and a naming convention for services. Examples are the Service Location Protocol (SLP) [27], Jini [20], HAVi [30], Web Service Description Language [84], and UpnP [76].

Another context in which this problem is addressed is the Semantic Web. There, too, service description and discovery mechanisms are developed, (e.g. DAML Webservices) [14,71].

The large number of approaches results in a situation where either service needs to be registered with different kinds of directory agents, or users need to query different directory agents in order to make reasonably sure that the most appropriate service for a request gets discovered.

Open Mobile Alliance [50] is currently addressing the issue of global service discovery for wireless telecom networks.

In ad-hoc networks the situation is even more complex. Here, we cannot rely on any one component to be always available. Therefore, approaches that build on a dedicated service directory are not feasible. What is needed instead, is a set of protocols that allows nodes to discover services offered by their peers. These protocols need to be able to deal with the high dynamics of ad-hoc networks. Until now, this problem has not been extensively investigated. A first, agent-based approach is described in [54].

## 5.2 General Authorization Service

### 5.2.1 Authentication

Many network services require some form of access control to allow approved users to perform certain tasks while prohibiting others. In some cases, these checks are most appropriately performed in the application itself, but in others it is desirable to provide authentication at a lower level. Some wireless standards already ensure terminal authentication of the wireless device to the base station, and visa versa. However, this level of authentication is also not always sufficient, since it merely ensures that the wireless device is properly authenticated, not that higher-level issues of users, software components, and privileges are checked. When higher-level middleware services wish to interact with authentication among themselves, they cannot rely on the applications they serve to provide security, nor can they rely on the lower level authentication of the wireless terminals.

Authentication over networks is most typically provided through cryptographic methods. This report will not repeat widely available material on these methods, but will simply note that public key cryptography is a popular alternative for such authentication, but symmetric key cryptography (e.g., Kerberos [47]) is also in use. Certificate systems and other forms of public key infrastructure have been developed to allow secure distribution of the keys required for cryptographic authentication. The middleware infrastructure should not define new alternatives, but integrate existing useful systems and allow for the addition of future developments in these areas.

A primary requirement for authentication in the wireless middleware infrastructure is the allowance of integrated authentication between different cooperating middleware components. Thus, it would be desirable for the middleware infrastructure to offer standard methods for authentication that would allow application and middleware designers to perform such authentication. However, it would be undesirable to build any particular method of authentication into the middleware services. Instead, in emulation of the approach taken by IPSec [35], the infrastructure should provide the ability to plug in different authentication services, along with methods for middleware and application components to negotiate the authentication schemes they want to use.

Care in implementation is particularly important for security software. All too often a supposedly secure scheme fails because the cryptography was used improperly or other details were bungled. Thus, the details of how authentication services plug into the infrastructure and are used, is not a matter for casual design. It must be designed carefully by knowledgeable security professionals and made available for public scrutiny and comment before it is included in the infrastructure. The lessons learned while defining IPSec [35] and its associated authentication components [46,28] should be remembered in this process.

### 5.2.2 Inter-system Roaming and Security Support

As was stated above, mobility of the terminals can be divided into inter-system mobility and intra-system mobility. The former requires roaming support. GSM networks, the number of which exceeds currently 700, were the first to deliver a truly global roaming support. They provide access to the wireless network resources for foreign, roaming terminals. The key components of the infrastructure for roaming are the Home Location Register (HLR) and the Visitor Location Register (VLR). The former is located at the home network of the terminal (i.e. the network that is operated by the operator that granted the SIM card in the roaming terminal) and it keeps track of the location of the terminal by storing the current Location Area Identity (LAI). Whenever a terminal is roaming and is logged on to a foreign network, VLR forwards a (IMSI, MSRN) pair to the HLR, where IMSI = International Mobile Subscriber Identity and MSRN = Mobile Subscriber Roaming Number. Based on the latter, the HLR is able to forward the incoming calls to the VLR that has supplied the latest location update. VLR contains information about the terminals currently roaming in its particular network. This information includes the current location, the resolution of LAI, the Temporary Mobile Subscriber Identity (TMSI)-IMSI mapping, as well as a ciphering key $K_c$ and a key number $K_{n.}$ used to generate $K_{c.}$ TMSI is used in the communication with the terminal, instead of IMSI and both VLR and the terminal store it. A new TMSI is allocated whenever the terminal roams to a new location area.

An important part of the security guarantees that the authenticity of the terminal is checked by exchanging keys and by a random number. The message completing the location update is encrypted on the radio path, so that the new TMSI is sent encrypted [72, ch.8]. The functionality of HLR resembles that of the Home Agent and the VLR likewise shares some features of the Foreign Agent discussed earlier in the context of Mobile IP. In GSM networks the ciphering key $K_c$ is generated for use in the subsequent communication over the wireless link.

These mechanisms are necessary for roaming support and they belong to the current network infrastructures at layers 1-3. In the global network with over one billion subscribers, such mechanisms seem indispensable. From the future infrastructure point of view, it is worth considering how the authentication and encryption of the terminal – and the user - will be performed. One can especially ask, should the roaming support, authentication of the terminals, and encryption of the subsequent communication be combined in a similar fashion as in the GSM case, or should these functions be somehow separated? In mobile phone networks, the services offered are voice and data calls and SMS services, and the network relies on the terminal authentication and authorization while granting the services.

Currently, and in the future, there will be more higher-level services (e.g., Mobile Commerce services) offered to the roaming customers. These services often require their own authorization and identification mechanisms at the application level [78,79]; it is not conceivable that these services would only trust the terminal authentication, performed at the network level, without also ensuring that the right user is using the terminal. This requires a separate identification and authorization scheme for the user level for those services. Apart from this, a terminal authentication and authorization scheme must still be offered at the network level for many reasons. These include charging correctly for the network resource usage by domestic and roaming users, maintaining a certain security level (otherwise e.g. stolen devices could easily be used unnoticed), protecting privacy, as well as providing ease of use for the customer.

Voice and data traffic between a wireless terminal and another component must also be encrypted in the future. Thus, the infrastructure should offer this also for the roaming terminals. As the example of GSM shows, this is by no means trivial due to the key exchange problematics.

### 5.3 Location Management

Location-based services are likely to be very important future applications of mobile and wireless systems. Emerging commercial location-based services fall into one of two categories.

1. Mobile Resource Management (MRM) applications, including systems for mobile workforce management, automatic vehicle location, fleet management, logistics, and transportation management and support. These systems use location data combined with route schedules to track and manage service personnel or transportation systems. Call centers and dispatch operators can use these applications to notify customers of accurate arrival times, optimize personnel utilization, handle emergency requests, and adjust for external conditions like weather and traffic.

2. Location-aware services that use location data to tailor the information delivered to the mobile user to increase relevancy. Examples include delivering accurate driving directions, instant coupons to customers nearing a store, or nearest resource information like local restaurants, hospitals, ATM machines, or gas stations. Analyses Ltd. estimates that location based services will generate $18.5B in sales by 2006. (Location-awareness alone is not enough for a large number of mobile applications. Here, location is only one aspect of a more complex context that should be taken into account when delivering services. As an example, services should adapt to the user's computer and the status it is in, e.g. with respect to battery power. Services should adapt to environmental conditions (a tourist guide should not suggest outdoor events during a major thunderstorm) and user profiles. Thus, nodes should not only be able to figure out where they are, but also other environmental conditions.)

*Location management*, the management of transient location information, is an enabling technology for all these applications. Location management is also a fundamental component of other technologies such as fly-through visualization (the visualized terrain changes continuously with the location of the user), context awareness (context of the user determines the content, format, or timing of information delivered), augmented reality (location of both the viewer and the viewed object determines the type of information delivered to viewer), and cellular communication.

Location management has been studied extensively in the cellular architecture context. The key problems are finding which cell a user is in (point queries), and updating a user's location when he moves to a new cell (point updates). Typical research issues in cellular architectures are how to distribute, replicate, and cache the database of location records. Related questions are how frequently to update and how to search the database [60], [6].

The location management problem is much broader, however. The main limitations of the cellular work are that the only relevant operations are point queries and updates that pertain to the current time, and they are only concerned with location to within the resolution of a cell, which can be quite large compared to what is needed by many applications. For broader wireless mobility, queries are often answerable by a set of related data objects. Location of a finer resolution is necessary. Queries may pertain to the future or the past, and triggers are often more important than queries. Some examples of queries/triggers are:

- "During the past year, how many times was bus #5 late by more than 10 minutes at some station?" (a past query)
- "Send me a message when a helicopter is in a given geographic area." (a trigger)
- "Retrieve the set of trucks that are expected to reach their destination within the next 20 minutes." (a set-oriented future query)

In other words, two things should be kept separate from each other in the location of the terminal 1) for the purpose of call/packet forwarding, and 2) for the purpose of location-dependent applications. In 2G networks, the former is solved and functions globally (using HLR and VLR registers); Mobile IP is expected to scale up in IP networks. The latter is a different issue and boils down to determining the coordinates of the terminal. The structures used for the former purpose are not well-suited for the latter task. The intersection is cell-id based positioning, but notice that

this also requires additional information (e.g., the coordinates of the base station) in order to be usable.

The current approach for developing MRM and location-aware applications is to build a separate independent location management component for each application. However, this results in significant complexity and a duplication of efforts, in the same way that data management functionality was duplicated before the development of Database Management Systems. To continue the analogy, we need to develop location management technology that addresses the common requirements, and serves as a development platform just as DBMS technology has done by extracting concurrency control, recovery, query language and query processing, in order to serve as a platform for inventory and personnel application development. We believe that a Location Management System (LMS) needs to be part of the infrastructure of mobile and wireless systems. For an example of an LMS, see [85].

The capabilities required of an LMS include support for modeling of location information, uncertainty management, spatio-temporal data access languages, data mining (including traffic and location prediction), location dissemination in a distributed/mobile environment, privacy and security, and fusion and synchronization of location information obtained from multiple sensors. Moreover, indexing and scalability issues need attention in order to enable widespread deployment. A large number of publications have addressed indexing for moving objects (e.g. [66,73]).

**Modeling.** A fundamental capability of location management is modeling of transient location information, particularly the location of mobile devices such as cell phones, personal digital assistants, laptops, etc. These devices are carried by people, or mounted on moving objects such as vehicles, aircraft, or vessels. The location information is updated by positioning technologies. Some example technologies are:

1. GPS transmitted from the device to the location server via a wireless network.
2. Network-based positioning that computes the location by triangulation of transmission towers.
3. Fixed sensors in the environment (e.g. at a toll booth) that identify the moving object by proximity.
4. Cell-id that identifies the cell in which the moving object is located (a low resolution method).
5. In the future, digital cameras may be used to identify the location of an object by scene analysis.

The challenge will be to model and provide an efficient access language for the location information obtained from many mobile devices (see for example [77]).

**Uncertainty Management**. The location of a moving object is inherently imprecise. Continuous motion and difficulty in reliably obtaining the object's location at any point in time mean that the object's location stored in the database cannot always be identical to its actual location. Systems that do not manage this uncertainty delegate to the user the responsibility of understanding and taking into consideration its implications. The objective of uncertainty management is to assist the user in this task, or to totally relieve one from it. An approach to uncertainty management was published in [75].

**Distributed/Mobile Environment**. It is often impractical or simply impossible to store the location database in a centralized location. For example, in the cellular database discussed earlier, a centralized architecture would create an intolerable performance problem. The question becomes how to allocate, update, and query

location information in a geographically distributed environment. Another complication for middleware infrastructure components arises when the location database is not only distributed but is also mobile. This is the case, for example, in a Mobile Ad-hoc Network.

**Location prediction**. This capability is critical in a digital battlefield, or security and anti-terrorist applications, where the owner of a mobile device does not provide her location to the LMS server. Instead the server needs to infer this information from historical, potentially incomplete and noisy location data. This capability is also important in mobile electronic commerce. For example, if at 8 A.M. it is known that at 9 A.M. a customer will be close to a store that has a sale on merchandise that matches her profile, the system could transmit a coupon at 8 A.M, allowing the customer to plan a purchase stop. Location prediction is important in other applications such as wireless bandwidth allocation.

### 5.4 Resource Discovery when Users and Resources Both Move

The techniques for roaming support at the network level are currently in 2G and 3G networks, as well as in WLANs. That is, a roaming terminal is able to get access to the network resources and can communicate with other terminals or servers using them.

Higher-level services, like location-based services or other locally available M-commerce services are still not automatically accessible. The terminal should find them first and subsequently be able to use them. What hinders this is a phenomenon, sometimes called *roaming heterogeneity*. This problem is potentially present whenever a non-resident terminal begins to use network resources in a foreign network. The heterogeneity can be exhibited at various system levels, ranging from the basic network level, up to the application protocols, mark-up languages, service semantics, and natural languages used as part of the services. A typical example is GSM terminals that do not operate in Japan or in many places of USA, because they cannot communicate with the network infrastructure.

Assuming that the terminal is able to communicate with the visited network, finding a suitable service is still a problem. This is a general service discovery problem discussed in subsection 5.1. In the roaming case one must, however, remember two things: the global scale of the problem, and the need for "local services" whose scope ranges from room level to city level. The first aspect means that there are millions of services that should be listed in a huge global directory (conceptually). The second aspect requires that each separate service description should also contain its "area of validity". The latter can range from a room to the entire globe.

Those services that are valid on the whole earth (such as the Amazon.com book shop) are of less interest here; they can be provided whenever the terminal gets access to the Internet and is able to run a suitable Web browser, whether it is roaming or not. The more difficult cases are the "local services" that are in the vicinity of the user and that can only be consumed on the spot. A typical example is a taxi service, local restaurant, hotel, gas station, museum, department store, local newspaper, etc. In order to find these services, the terminal should be able to determine its location and then ask for the services relevant for that location. Several questions immediately

arise, including the location determination, how the services should be described in the global directory, how the latter is effectively organized, how the concrete services can be accessed by a foreign terminal (mark-up language, protocols supported by the terminal and the server offering the service, language of the service, business rules obeyed by the service), etc. In abstract terms, the problem can be seen as a special location-related query: "Find the (nearest, cheapest, …) service of type T with attributes (A, B, C, …) for my position (X,Y)". That is, the directory service must map the current coordinates and other predicates provided by the user to the appropriate service of type T that is offered in the vicinity of the user. Position (X,Y) can be the current one, a future/past one, or even a hypothetical one.

Positioning of the terminal has been recently introduced within wireless networks as a new basic service. This service is necessary for the operation of the dynamic location-dependent services, both in the service discovery phase, as well as in the service execution phase. For instance, a taxi service can be requested based on the location of the customer (taxi company in the city or a suburb where the customer is located while ordering) and the actual taxi fetching the customer can further be allocated based on the coordinates of both. Payments can even be made in the taxi using the capabilities of the terminal if it contains the credit card information and is able to communicate with the taxi over a local wireless connection (see e.g. [80] for further elaboration).

Design of a global scale service discovery infrastructure is largely an open research issue. Partial solutions are presented by Location Forum, which is standardizing the protocols and formats for LBS [40], SLP [27], OSGI (www.osgi.org), and the protocols for service discovery presented above, like UDDI. This list must be considered incomplete, and new mechanisms are expected to be developed in the future. The directory structure, mark-up language used, service description DTD etc., can perhaps be adapted from existing work on W3C, other similar forums, or standardization bodies.

Apart from service discovery, many issues related to roaming heterogeneity still need work. It must be addressed on all protocol levels (especially in the presentation layer), and in the middleware. The highest level is the natural language level, because local services tend to use the local language (cf. Japanese I-Mode services).

### 5.5 Persistent Storage

Enabling persistent storage of data is particularly important in a system comprising mobile devices due to the inherent vulnerability of these devices. Mobile devices break more easily than stationary hosts; maybe even more importantly, they are at a much greater risk of being stolen or otherwise lost. Thus, being able to persistently store data on a more reliable device is a necessity.

While the workshop participants agreed on this, they did not agree quite as unanimously on whether it should be the responsibility of the infrastructure to offer persistent storage. While one group of participants strongly argued that this kind of service will be constantly needed by everybody everywhere and should thus be part of the infrastructure, another group of participants argued that this should be just one service among others that users should be able to find using a service discovery mechanism should they need it. Another counterargument against providing

persistent storage as an infrastructure service is that the format in which the data is stored might not be suitable for the needs of all applications. Perhaps it is better for applications to provide customized storage services suited to their needs.

Assuming persistent storage is regarded as part of the infrastructure, a number of open questions need to be addressed. The first decision that needs to be made is what data model to use for storage. Should data be stored as raw-byte streams, in a file system, as web pages, or would it make more sense to rely on the services of a relational, object-oriented or semi-structured database system? Should the storage facility provide a per-user store, a single common store shared by all users, stores for particular applications, or some combination of these choices? The second decision is where to store the data. Would a centralized location be preferable? If (as is likely) a distributed solution is used, should data move around with the user? Closely related to this question is the third decision, that is, should data be replicated and if so, which replication strategies should be used? Should persistently stored data be usable by one user only? If multi-user access should be supported, transactional mechanisms are needed to ensure data consistency.

Past work that has been done on file systems for mobile and wireless environments (e.g., CODA [67] and Rumor [62]) is a good basis for new efforts. Newer Coda models such as the import/export and the session server can be considered. Recent work from the database community on mobile databases should be considered when determining the answers to these questions [21]. Important work has been done in particular on replication strategies [81] and transaction models [19] for mobile and wireless environments.

### 5.6 Caching/Hoarding/Pre-fetching

While some participants of the workshop argued strongly for the inclusion of caching/hoarding/pre-fetching components into the middleware, others felt as strongly that while these services are without a doubt useful in mobile and wireless environments, and should not be considered as part of the infrastructure but as the application programmer's responsibility. However, all of these techniques may use some kind of fixed resource such as memory, server side or client side, which may be based on the infrastructure.

The main argument in favour of infrastructure-based caching was that efficient caching might only be possible if independent nodes cooperate, requiring a coordinating infrastructure component. While this sounds reasonable, there is very little prior work in the area that would allow specification of exactly how cooperative caching in wireless networks should be if realized. There is also a need for some fixed standards/specifications and formal models in the caching mechanisms. Among the open questions are: Which model of cooperative caching should be used? Alternatives are cooperation between client and cache nodes, between different cache nodes, and between server nodes and cache nodes. Should push caching be included? Should predictive semantic caching be used along with a semantic model that fulfils all kinds of user specifications [38]? Should cached items be distributed using broadcast? What should the unit of caching be? Should semantic distances be used to determine the caching priority of data? Should location-dependent data be considered by caching mechanisms? [64]. Different types of data items could be cached, such as

web pages, files, database relations or fragments of relations, objects; alternately, the system might provide a general caching approach allowing a combination of these alternatives.

As infrastructure elements, hoarding and pre-fetching are even more questionable than caching. As a rule, hoarding is performed by a mobile device before disconnection, and is largely a matter of communication between that device and the sites that permanently store its data. It is unclear that infrastructure has any useful role to play here, other than providing network connectivity and reasonable bandwidth. Pre-fetching is most commonly performed by the using node, so again the role of the infrastructure is unclear, though perhaps pre-fetching by caching infrastructure nodes might be useful. On the other hand, many issues in hoarding such as server-side caching mechanisms can depend on the infrastructure. Another area where hoarding might depend on the infrastructure is resource allocation. The more resources available for allocation the smoother the resource allocation will be in the system.

In this area the open questions outnumber the answers by far.

## 6 Conclusion

This report reflects the authors' broad consensus with respect to defining and establishing a common Infrastructure for the discipline of Mobile and Wireless networking. It is the direct result of presentations and discussions that took place during the NSF workshop on Infrastructure for Mobile and Wireless Systems, held on October 15, 2001 in Phoenix, Arizona. The purpose of the workshop was to identify and characterize the major research issues associated with the design, development, and deployment of future mobile and wireless networking and to provide specific suggestions for focused research activities. The participants at the workshop embraced several different disciplines and points of view, numbering 26 experts from many different mobile and wireless communities, including communications, operating systems, core networking, mobility, databases, middleware, and networking. The report provides a foundation for implementation, standardization, and further research and discussion on the issues of what should constitute such an infrastructure to investigators and funding organizations.

Prior to the workshop, attendees submitted position papers that described their recommended priorities for research in the field. The position papers were discussed, and this discussion formed the basis for the research priorities identified by the workshop. Whenever possible, we have organized the topics into broad classes, attempting to impose order and structure on the wide variety of research issues and problems brought up during the workshop.

One issue for discussion in the workshop was the scope of the infrastructure. There are many kinds of wireless networks, existing and proposed. Is the infrastructure meant to handle all of them, and all kinds of mobility? There was general agreement that the infrastructure should handle both wireless cellular networks that rely primarily on single hop communications to a fixed base station that is connected to a wired network, and ad hoc networks that might communicate via multihop wireless networks before reaching a wired segment (or perhaps without any

participation by wired segments). In the latter category, the infrastructure should at least include 802.11 wireless LANs, Bluetooth, JINI and ad hoc IP based systems currently under development, but ideally it should be flexible enough to handle many other such networks (including cellular networks).

There was less agreement on whether the infrastructure being defined here should support sensor networks, particularly those that use diffusion based methods to transmit their information. The needs of such networks are substantially different than those of more conventional wireless networks. Whether a sufficiently general infrastructure could suitably service both styles of networks requires further research and discussion.

Some major currents emerged from the workshop and are taken up in this report. First, participants identified specific, high-priority topics in several areas that require the attention of researchers. These priorities are grouped into different categories and are discussed at length in this report. Next, the workshop participants acknowledged the desirability of providing themes for research that would serve to promote mobile and wireless infrastructure. Finally, fourteen focused research initiatives were suggested, having clearly defined objectives and deliverables. These initiatives would demonstrate advanced concepts in mobile and wireless networks and would provide critical components and software for use in experimental research in the area.

### 6.1 Recommended Components of Wireless Infrastructure

Three broad areas of research priorities, which are discussed in more detail in the foregoing workshop report, include the following:

1 Network layer Infrastructure

1.1 *Alterations to IP*

Because mobile computers using a wireless infrastructure will want access to the same services as fixed computers, they will need to interoperate with the Internet. The wireless infrastructure must provide IP support.

1.1.1 *Routing Protocols for Mobile and Wireless Infrastructure*

The infrastructure should be agnostic to the choice of routing protocols, but actually ensuring that this is true may prove difficult. For instance, certain choices related to security may restrict the flow of information that some routing protocols require.

1.1.2 *Multicast Protocols for Mobile and Wireless Infrastructure*

Already significant in wired networks, multi-casting is expected to assume an equal role in mobile and wireless infrastructure, especially in support of multimedia transmission. Several fundamental services in mobile and wireless infrastructures will be built upon multicasting. Multicast routing in a mobile and wireless infrastructure should be able to handle multicast groups of varying densities, depending on the capacity of the terminal devices.

1.1.3 *Other network services*

Other services have been introduced into the Internet over the years. Many are layered on top of IP. Others are provided via IP options or through augmentation of routers. Defining a new mobile and wireless infrastructure also offers a chance to prune out any failed services that were introduced at some point into the wired Internet. Critical thought should be applied to each

service that has at some point made its way into the Internet infrastructure to decide if its value warrants supporting it in the wireless infrastructure.

1.2 *Network Management Issues*

Determining the proper set of network management services to include in the mobile & wireless infrastructure is a question for research. Choices made for other infrastructure components will affect the choice of network management services. Most other components under consideration require either management of their interaction with other components, or internal management of their own operations, if not both. Designers of the middleware infrastructure should consider the network management needs of higher-level components and incorporate useful services for them, where such services can be provided with suitable privacy and security.

1.3 *Adaptation Services*

Networks supporting mobile and wireless use often have links and devices with limited capabilities that are not suited to normal data flows. Various adaptations can be made to data flowing over the network to transform a data stream into a form appropriate for current conditions. Support for these services could be provided in the infrastructure for a mobile and wireless network. Adaptations must be chosen based on varying conduits and must be placed at various points in the network to achieve the best possible behavior. This in turn necessitates discovery and negotiation protocols.

2 Transport Layer Infrastructure

2.1 *Alterations to transport protocols*

The abstract transport service can be implemented using different transport protocols such as Transmission Control Protocol (TCP), which was originally designed for the Internet. TCP is not perfect for wireless networks because it makes assumptions about the behavior of the underlying packet network that are not true in the wireless case. It does not supports audio and video streams currently emerging in the wireless arena. A great deal of work appears to have been done in identifying the important research issues here, but no consensus has emerged on the best approach for a solution.

2.2 *QoS support*

Wired and wireless networks must both provide the QoS support required by applications. QoS in mobile and wireless infrastructures is substantially more complex than in a purely wired infrastructure because there are more factors involved, some of them being highly variable. Research in mobile and wireless infrastructures is necessary to ensure that advanced applications will be supported and that wireless and wired network QoS principles will be compatible and interoperable.

3 Middleware Layer Infrastructure

3.1 *Service discovery*

Mobility and wireless networks lend to frequently changing environments. Therefore, we cannot rely on the user or the computer to know which services are available in the network, where they are located, and how they can be accessed. The large number of approaches results in a situation where either service needs to be registered with different kinds of directory agents or users in order to reasonably make sure that the most appropriate service for a request gets discovered.

3.2 *General authorization service*

Provisioning of authentication and encryption needs to be addressed as part of the development of the mobile and wireless infrastructure. A general method of permitting authentication of user requests is a vital infrastructure component. Not enough work has been done to identify the important research issues here. Results should be oriented towards the practical needs of network operators.

3.3 *Location management*

Location management functions make it possible to exploit and cope with dynamically changing location information in applications. Location management can work at several layers and can be a complex process. Location based services are perhaps the most important future application of mobile and wireless systems. It is critical that researchers analyze and develop experiments that determine a scalable overall architecture for location management in a mobile and wireless infrastructure.

3.4 *Resource discovery when users and resources both move*

The techniques for roaming support at the network level currently work in 2G and 3G networks, as well as in WLANs. That is, a roaming terminal is able to get access to the network resources and consequently can communicate with other terminals or servers. Assuming that the terminal is able to communicate up to the application level, finding a suitable service is still a problem. Apart from service discovery, many issues related to roaming heterogeneity still need work on all protocol levels, and in the middleware.

3.5 *Storage*

Due to the inherent vulnerability of these devices, enabling persistent storage of data is particularly important in a system comprising mobile devices. Mobile devices break more easily than stationary ones and maybe even more importantly, are at a much greater risk of being stolen or otherwise lost. Thus, being able to persistently store data on a more reliable device is a necessity. A number of open questions need to be addressed, such as what data model to use for storage, where to store the data, should data be replicated, which replication strategies should be used, etc.?

3.6 *Caching/hoarding/pre-fetching*

The main argument in favor of infrastructure based caching was that efficient caching might only be possible if different nodes cooperate, requiring a coordinating infrastructure component. However, while this sounds reasonable, there is very little prior work in the area that would allow specification of exactly how cooperative caching in mobile and wireless networks should be if realized. In this area the open questions outnumber the answers by far.

These topics are not exhaustive, and one expects that other topics will come to the forefront as the field develops.

## 6.2 Areas Requiring More Research

6.2.1 Network management for wireless
6.2.2 Security design for wireless infrastructure
6.2.3 Resolving controversy on some infrastructure components

### 6.3 Need for Integration and Testing of Various Infrastructure Components

The workshop participants agreed that the infrastructure support must support multihop computing paradigms. In addition to the widely used client/server paradigm, the infrastructure should provide support to the emerging peer to peer and agent paradigms of computing. The increasing heterogeneity of future communication systems and network, both in terms of network architectures and terminal capabilities, requires efficient design, implementation and operation. This is particularly in the case that mobile and wireless infrastructure is compared to a wireline backbone networking infrastructure and is most evident in the transport of evolving services such as multimedia. The heterogeneity can perhaps be addressed through use of some form of scalable delivery approach where the source material is represented in terms of hierarchical layers with each layer handled differently by the network.

## References


1. 3rd Generation Partnership Project (3GPP). Adaptive Multi-Rate Speech CODEC., Oct.-Dec.1999. Available through http://www.3gpp.org/specs/publications-partners.htm.
2. A. Acharya, and B. R. Badrinath, "A Framework for Delivering Multicast Messages in Networks with Mobile Hosts," ACM/Baltzer Journal of Mobile Networks and Applications, Vol. 1, No. 2, pp 199-219, October 1996.
3. M. Allman, V. Paxon, W. Stevens, TCP Congestion Control, April 1999. RFC2581. http://www.ietf.org/rfc/rfc2581.txt
4. B. Badrinath, A. Fox, L. Kleinrock, G. Popek, P. Reiher, and M. Satyanarayanan, "A Conceptual Framework for Network Adaptation," Mobile Networks and Applications, Vol. 5, No. 4, pp. 221-231, 2000.
5. H. Balakrishnan, S. Seshan, E. Amir, and R. Katz, "Improving TCP/IP Performance Over Wireless Networks," *Proceedings of the 1st ACM International Conference on Mobile Computing and Networking (MobiCom '95)*, November 1995.
6. A. Bhattacharya, S. K. Das, "Lezi-Update: An Information-Theoretic Approach to Track Mobile Users in PCS Networks," Proceedings of the Fifth ACM/IEEE International Conference on Mobile Computing and Networking (Mobicom99), Seattle, WA, August 1999.
7. Bluetooth Consortium home page, available at www.bluetooth.com.
8. J. Border, M. Kojo,J. Griner, G. Montenegro,Z. Shelby, Performance Enhancing Proxies Intended to Mitigate Link-Related Degradations. June 2001. RFC3135. http://www.ietf.org/rfc/rfc3135.txt.
9. Antonio Carzaniga, David S. Rosenblum, Alexander L Wolf, Achieving Scalability and Expressiveness in an Internet-Scale Event Notification Service. Proceedings of the Nineteenth ACM Symposium on Principles of Distributed Computing (PODC 2000), Portland, Oregon, July 2000.
10. Antonio Carzaniga, David S. Rosenblum, Alexander L Wolf, Design and Evaluation of a Wide-Area Event Notification Service. ACM Transactions on Computer Systems 19(3), August 2001.
11. Antonio Carzaniga, Alexander L Wolf. Content-based Networking: A New Communication Infrastructure. NSF Workshop on an Infrastructure for Mobile and Wireless Systems, Scottsdale, Arizona, October 2001.

12. Th. Clausen, P.Jacquet, A. Laouiti, P. Minet, P. Muhlethaler,A.Qayyum, L. Viennot Optimized Link State Routing Protocol (OLSR). Internet Draft. Sept. 2001. http://www.ietf.org/internet-drafts/draft-ietf-manet-olsr-06.txt
13. M. Scott Corson and A. Ephremides, "A Distributed Routing Algorithm for Mobile Wireless Networks," ACM Journal on Wireless Networks, Vol. 1, No. 1, 1995.
14. DAML Services Coalition (alphabetically A. Ankolekar, M. Burstein, J.Hobbs, O. Lassila, D. Martin, S. McIlraith, S. Narayanan, M. Paolucci, T. Payne, K. Sycara, H. Zeng). ``DAML-S: Semantic Markup forWeb Services", *Proceedings of the International Semantic Web Working Symposium (SWWS)*. August 1, 2001.
15. DARPA Active Networks Composable Services Working Group, "Composable Services for Active Networks", working document at http://www.cc.gatech.edu/projects/canes/papers/cs-draft0-3.ps.gz.
16. DARPA Active Networks Security Working Group, "Security Architecture for Active Nets", working document, available at http://www.pgp.com/research/nailabs/network-security/secure-active.asp, 2001.
17. S. Dawkins, G. Montenegro, M. Kojo, V. Magret, N. Vaidya, End-to-end Performance Implications of Links with Errors. Aug. 2001. RFC3155. http://www.ietf.org/rfc/rfc3155.txt
18. S. Deering, R. Hinden. Internet Protocol, Version 6 (IPv6), RFC 2460. http://www.ietf.org/rfc/rfc2460.txt
19. M. Dunham, A. Helal, S. Balakrishnan, "A Mobile Transaction Model That Captures Both the Data and the Behavior," *Mobile Networks and Applications*, Vol. 2, No. 2, 1997.
20. W. Keith Edwards. Core JINI. Prentice Hall, 1999.
21. Todd Ekenstam, Peter Reiher, Charles Matheny, and Gerald Popek, "The Bengal Database Replication System," the Journal of Distributed and Parallel Databases, Vol. 9, No. 3, May 2001.
22. D. C. Feldmeier, A. J. McAuley, J. M. Smith, D. S. Bakin, W. S. Marcus, and T. M. Raleigh, "Protocol Boosters," *IEEE Journal on Selected Areas in Communications*, April 1999, 16(3): 437-444.
23. R. Finlayson, A More Loss-Tolerant RTP Payload Format for MP3 Audio, June 2001. RFC3119. http://www.ietf.org/rfc/rfc3119.txt.
24. S. Floyd, J. Mahdavi, M. Mathis, M. Podolsky, An Extension to the Selective Acknowledgement (SACK) Option for TCP. July 2000. RFC2883. http://www.ietf.org/rfc/rfc2883.txt
25. A. Fox, S. Gribble, E. Brewer, E. Amir, "Adapting to Network and Client Variability via On-Demand Dynamic Distillation," *Proceedings of the Seventh International Conference on Architectural Support for Programming Languages and Operating Systems*, October 1996.
26. A. Fox, S. Gribble, Y. Chawathe, E. Brewer, and P. Gauthier, "Cluster-Based Scaleable Network Services," *Proceedings of the 16th ACM Symposium on Operating System Principles (SOSP '97)*, October 1997.
27. E. Guttman, C. Perkins, J. Veizades, M. Day, Service Location Protocol, Version 2. RFC2608, June 1999. http://www.ietf.org/rfc/rfc2608.txt
28. D. Harkins, C. Carrel, The Internet Key Exchange (IKE), November 1998. RFC2409 http://www.ietf.org/rfc/rfc2409.txt
29. D. Hoffman, G. Fernando, V. Goyal, M. Civanlar. RTP Payload Format for MPEG1/MPEG2 Video, Jan. 1998. RFC2250.http://www.ietf.org/rfc/rfc2250.txt
30. HAVi – Home Audio/Video Interoperability. http://www.havi.org
31. IETF. Multi-Protocol Label Switching (MPLS). Overview page to Internet Drafts and other documentation at http://www.ietf.org/ids.by.wg/mpls.html.
32. David B. Johnson and David A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," Mobile Computing, edited by Tomasz Imielinski and Hank Korth, 1996.
33. D.Johnson, D.Maltz, Y-C.Hu, J.Jetcheva, The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR). Internet Draft. Feb. 2002. http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-07.txt

34. D.B. Johnson, C. Perkins, Mobility Support in IPv6. Internet Draft. March 2002. http://www.ietf.org/internet-drafts/draft-ietf-mobileip-ipv6-18.txt
35. S. Kent and R. Atkinson, Security Architecture for the Internet Protocol, November 1998. RFC2401 http://www.ietf.org/rfc/rfc2401.txt
36. Y. Kikuchi,T. Nomura,S. Fukunaga,Y. Matsui,H. Kimata, RTP Payload Format for MPEG-4 Audio/Visual Streams, Nov. 2000. RFC3016.http://www.ietf.org/rfc/rfc3016.txt
37. D. A. Levine, I. F. Akylidiz, and M. Naghshineh, " A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks using the Shadow Cluster Concept," IEEE/ACM Transactions on Networking, Vol. 5, pp. 1-12, February 1997.
38. L. Li, B. König-Ries, N. Pissinou, K. Makki, "Strategies for Semantic Caching," In Proceedings of the 12th International Conference on Database and Expert Systems Applications (DEXA), Munich, Germany, September 2001.
39. Mika Liljebert, Heikki Helin, Markku Kojo, and Kimmo Raatikainen, "Mowgli WWW Software: Improved Usability of WWW in Mobile WAN Environments," *Proceedings of IEEE Global Internet 1996*, London, England, November 1996.
40. Location Interoperability Forum. http://www.locationforum.org/
41. P. Luthi, RTP Payload Format for ITU-T Recommendation G.722.1, Jan. 2001. RFC3047, http://www.ietf.org/rfc/rfc3047.txt.
42. K. Makki, N. Pissinou, and O. Frieder, "Efficient Solutions to Multicast Routing in Communication Networks," Mobile Networks and Applications, Vol. 1, No. 1, pp. 221-232, 1996.
43. Mobile Ad-hoc Networks (manet). Home page at http://www.ietf.org/html.charters/manet-charter.html
44. M. Mathis, J.Mahdavi, S. Floyd, A. Romanov, TCP Selective Acknowledgement Options, Oct. 1996. RFC2018. http://www.ietf.org/rfc/rfc2018.txt
45. K. McKay, RTP Payload Format for PureVoice(tm) Audio, August 1999. RFC2658. http://www.ietf.org/rfc/rfc2658.txt
46. D. Maughan, M. Schertler, M. Schneider, J. Turner, Internet Security Key Management and Association Protocol (ISAKMP), November 1998. RFC2408 http://www.ietf.org/rfc/rfc2408.txt
47. B. Clifford Neuman and Theodore Ts'o, "Kerberos: An Authentication Service for Computer Networks," IEEE Communications Magazine, September 1994.
48. Nokia. http://www.nokia.com/press/nps_white_papers.html.
49. NTT DoCoMo home page, April 25th Press Release on V-Service at http://foma.nttdocomo.co.jp/english/
50. Open Mobile Alliance. http://www.openmobilealliance.org/
51. S. Paul, "Reliable Multicast Transport Protocol (RMTP)," IEEE Journal on Selected Areas in Communications, Vol. 15, No. 3, pp. 407-421, April 1997.
52. V. Paxson (ed). Known TCP Implementation Problems. RFC2525, March 1999. http://www.ietf.org/rfc/rfc2525.txt.
53. J. Postel, Transmission Control Protocol. Sept. 1981. RFC793.http://www.ietf.org/rfc/rfc793.txt
54. Filip Perich, Sasikanth Avancha, Anupam Joshi, Yelena Yesha, and Karuna Joshi, "Query Routing and Processing in Mobile Ad-hoc Environments", Technical Report, UMBC, November, 2001.
55. C. Perkins, IP Mobility Support for IPv4. Jan. 2002. RFC3220.http://www.ietf.org/rfc/rfc3220.txt.
56. C.Perkins, E.Belding-Royer, S.Das, Ad hoc On-Demand Distance Vector (AODV) Routing. Internet Draft. Jan. 2002. http://www.ietf.org/internet-drafts/draft-ietf-manet-aodv-10.txt.
57. C.E. Perkins, Ad-Hoc Networking, Addison-Wesley, 2002
58. N. Pissinou, B. Bhaghavati, B. König-Ries, and K. Makki, "On Fault Wireless Network Management, Annual Review of Communications, June 2001.

59. N. Pissinou, B. Bharghavati, K. Makki, “Mobile Agents to Automate Fault Management in Wireless and Mobile Networks,” In Proceedings of the IEEE International Parallel and Distributed Processing Systems Workshops 2000, pp. 1296-1300, 2000.
60. E. Pitoura, G. Samaras ”Locating Objects in Mobile Computing,” IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No. 4, July/August 2001.
61. G. Polyzos, and G. Xylemenos, “ Enhancing Wireless Internet Links for Multimedia Services,” Proceedings of the Fifth International Workshop on Mobile Multimedia Communications, Berlin, October 1998.
62. David Ratner, Peter Reiher, Gerald Popek, and Geoffrey Kuenning, ”Replication Requirements for Mobile Computing,” Mobile Networks and Applications, Vol. 6, No. 6, pp. 525-534, November 2001.
63. P. Reiher, R. Guy, K. Eustice, V. Ferreria, and M. Yarvis, “Cooperative Adaptation Between End Points,” Workshop on Active Middleware Services, August 2000.
64. Q. Ren and M. H. Dunham, “ Using Semantic Caching to Manage Location Dependent data in Mobile Computing,” MobiCom, August 2000.
65. RSA http://www.rsasecurity.com/rsalabs/index.html.
66. Simonas Saltenis, Christian S. Jensen, Scott T. Leutenegger, and Mario A. Lopez, "Indexing the Positions of Continuously Moving Objects", Proc. Of the SIGMOD Conference", 2000, pp. 331-342.
67. M. Satyanarayanan, J. Kistler, P. Kumar, M. Okasaki, E. Siegel, and D. Steere, “Code: A Highly Available File System for a Distributed Workstation Environment,” IEEE Transactions on Computers, Vol. 39, No. 4, April 1990.
68. H. Schulzrinne, S. Casner, R. Frederick,V. Jacobson, RTP: A Transport Protocol for Real-Time Applications. January 1996. RFC1889. http://www.ietf.org/rfc/rfc1889.txt
69. H. Schulzrinne, A. Rao, R. Lanphier, Real Time Streaming Protocol (RTSP), April 1998. RFC2326. http://www.ietf.org/rfc/rfc2326.txt
70. H. Schulzrinne, S. Petrack, RTP Payload for DTMF Digits, Telephony Tones and Telephony Signals. May 2000. RFC2833. http://www.ietf.org/rfc/rfc2833.txt
71. SemanticWeb.org: The Semantic Web Community Portal http://www.semanticweb.org
72. R. Steele (Ed.), “Mobile Radio Communications”, Prentech Press Publishers (London) & IEEE Press (New York), USA 1995, ISBN 0-7803-1102-7. 779 p
73. Tayeb, O. Ulusoy, O. Wolfson, A Quadtree Based Dynamic Attribute Indexing Method, Computer Journal, Vol. 41(3), 1998.
74. D. Tennenhouse and D. Wetherall, “Towards an Active Network Architecture,” *Computer Communications Review*, April 1996.
75. G. Trajcevski, O. Wolfson, S. Chamberlain, F. Zhang, "The Geometry of Uncertainty in Moving Objects Databases", Proceedings of the International Conference on Extending Database Technology (EDBT02)}, Prague, Czech Republic, March 2002. Springer LNCS 2287, pp. 233-250.
76. Universal Plug-and-Play (UPnP) Forum. Microsoft Corporation. http://www.upnp.org
77. M. Vazirgiannis, O. Wolfson, "A Spatiotemporal Query Language for Moving Objects", Springer Verlag Lecture Notes in Computer Science, number 2121, Proceedings of the 7th International Symposium on Spatial and Temporal Databases, Los Angeles, CA, July 12-15, 2001, pp. 20-35.
78. J. Veijalainen, Transactions in Mobile Electronic Commerce. In: Gunter Saake, Kerstin Schwarz, Can Türker (eds.), Transactions and Database Dynamics. Lecture Notes in Computer Science Nr. 1773, Springer Verlag, Berlin, December 1999, pp. 208- 229
79. J. Veijalainen, A. Tsalgatidou, Electronic Commerce Transactions in a Mobile Computing Environment. In Q.Jin, J. Li, N. Zhang, J. Cheng, C. Yu, S. Noguchi (eds) Enabling Society with Information Technology. Springer Verlag (Tokyo 1/2002), ISBN 4-431-70327-6, pp. 131-140.

80. J. Veijalainen, M. Weske, Modeling Static Aspects of Mobile Electronic Commerce Environments. A chapter in "Advances in Mobile Commerce Technologies", Lim Ee Peng, Keng Siau (eds.). Kluwer (forthcoming).
81. An-I Wang, Peter Reiher, Rajive Bagrodia, and Gerald Popek, "A Simulation Evaluation of Optimistic Replicated Filing in a Mobile Environment," 18th IEEE International Performance, Computing, and Communications Conference, February 1999.
82. B. Wang and J.C. Hou, "Multicast Routing and its QoS extension: Problems, Algorithm, and protocols," IEEE Networks, Vol. 14, January 2000.
83. Wapforum. www.wapforum.org
84. Web Service Description Language (WSDL). http://www.w3.org/TR/wsdl
85. O. Wolfson, H. Cao, H. Lin, G. Trajcevski, F. Zhang, N. Rishe, "Management of Dynamic Location Information in DOMINO", Proceedings of the International Conference on Extending Database Technology (EDBT02), Prague, Czech Republic, March 2002. Springer LNCS 2287, pp. 769-771.
86. Mark Yarvis, Peter Reiher, and Gerald J. Popek, "A Reliability Model for Distributed Adaptation," *Proceedings of the Third IEEE Conference on Open Architectures and Network Programming (OPENARCH '00)*, Tel-Aviv, Isreal, March 2000.

## Appendix A: List of Papers

### A.1 Invited Papers

- Keynote Speech: Charles Perkins (Nokia Research Center, Mountain View, California): Recent Developments with Mobile IPv6.
- Radek Vingralek (STAR Lab, Intertrust Technologies Corporation): Supporting E Commerce in Wireless Networks.
- Ouri Wolfson (University of Illinois at Chicago), Sam Chamberlain (Army Research Lab), Kostas Kalpakis (University of Maryland), Yelena Yesha (University of Maryland): Modeling Moving Objects for Location Based Services.
- P. R. Kumar (University of Illinois): Ad hoc wireless networks: Analysis, protocols, architecture, and convergence.

### A.2 Accepted Papers

- Content-Based Networking: A New Communication Infrastructure (Antonio Carzaniga and Alexander L. Wolf, University of Colorado)
- Flexible Integrated Cache for Efficient Information Access in Mobile Computing Environments (Mohan Kumar and Sajal Das, The University of Texas at Arlington)
- Design Considerations for Mobile Database Applications (Wai Gen Yee, Shamkant B. Navathe, Georgia Tech)
- Smart Environments: Middleware Building Blocks for Pervasive Network Computing (A Position Paper) (Jon Weissman and Zhi-Li Zhang, University of Minnesota)

- Agents, Mobility, and M-Services: Creating the next generation applications and infrastructure on mobile ad-hoc networks (Anupam Joshi, Timothy Finin, and Yelena Yesha, University of Maryland)
- ANTARCTICA: A Multiagent System for Internet Data Services in a Wireless Computing Framework (A. Goni, A. Illarramendi, E. Mena, Y. Villate, J. Rodriguez, Univ. of Zaragoza, Spain and Univ. of the Basque Country, Spain)
- A Collaborative Infrastructure for Mobile and Wireless Systems (L. Wegner, Morad Ahmad, S. Fröhlich, and Ch. Schmidt Universität GH Kassel, Germany)
- The Rational for Infrastructure Support for Adaptive and Context-Aware Applications: A Position Paper (Nigel Davies, Keith Cheverst, Christos Efstratiou and Adrian Friday, University of Lancaster and University of Arizona)
- Ambient Wireless Interfaces - in Search for the Limits (A position paper) Jari Veijalainen and Tom Gross GMD-FIT, Germany and University of Jyvaskyla, Finland)
- Multilayer Secure Real-time Video Transmission over CDPD Networks (Farid Hatefi, Gamze Seckin, Forouzan Golshani, Arizona State University)
- Middleware for Location Content Services in Mobile Environment (Y. Lee, N. Prabhu, E. K. Park, University of Missouri-Kansas City)

## Appendix B: Workshop Participants

- Lutz Wegner, Universität GHKassel
- Guohong Cao, Penn State University
- Eduardo Mena, University of Zaragoza, Spain
- Yugyung Lee, University of Missouri Kansas City
- Jari Veijalainen, University of Jyvaskyla, Finland
- Antonio Carzaniga, University of Colorado
- Alexander Wolf, University of Colorado
- Farid Hatefi, Arizona State University
- Wai Gen Yee, Georgia Tech
- Nigel Davies, University of Arizona
- S. R. Subramanya, University of Missouri-Rolla
- Sham Navathe, Georgia Tech
- S.A.M. Makki ***, Queensland Univ. of Technology, Australia
- Jon Weissman, University of Minnesota
- Birgitta Koenig-Ries **, Universität Karlsruhe, Germany
- Peter Scheuermann **, Northwestern University
- Charles Perkins, Nokia Research Center, Mountain View, California
- M. Satyanarayanan, Carnegie Mellon University
- Alan Blatecky, National Science Foundation
- Radek Vingralek, Intertrust Tech
- P. R. Kumar, University of Illinois
- Kia Makki *, Florida International University
- Niki Pissinou *, Florida International University

- Peter Reiher, University of California at Los Angeles,
- Mohan Kumar, University of Texas at Arlington
- Ouri Wolfson, University of Illinois at Chicago

* Workshop Co-General Chairs
** Workshop Co-Chairs
*** Workshop Program Vice Chair