ON BUILDING AN INFRASTRUCTURE FOR MOBILE AND WIRELESS SYSTEMS

Report on the NSF Workshop on an Infrastructure for Mobile and Wireless Systems, Oct. 15, 2001

Birgitta König-Ries [¨], Kia Makki[†], Sam Makki [‼], Charles Perkins [•], Niki Pissinou [†], Peter Reiher [┼], Peter Scheuermann [‖], Jari Veijalainen[‡], Ouri Wolfson[±]

I. Introduction

Over the last few years, there have been at least two dramatic changes in the way computers are used. The first has its origin in the fact that computers have become more and more connected to each other. The second was triggered by the increasing miniaturization and affordability of hardware components and power supplies, together with the development of wireless communication paths. These two trends combined have allowed the development of powerful, yet comparatively low-priced, portable computers. In spite of these changes, little attention has been given to reaching a common consensus and to the development of a strong infrastructure in this area.

The recent NSF Workshop on Infrastructure for Mobile and Wireless Systems held on Oct.15, 2001 in Scottsdale, AZ had the goal of defining and establishing a common infrastructure for the discipline of mobile and wireless systems.

This report is a summary of a consensus-based paper written after the workshop that will appear together with the papers of the workshop in a forthcoming Springer-Verlag volume.

The workshop participants were from many different wireless communities, including communications, operating systems, core networking, mobility, databases, and middleware.

E-mail addresses in alphabetical order:
koenig@ira.uka.de,kia@eng.fiu.edu,
s.makki@qut.edu.au,charliep@IPRG.nokia.com,
niki@eng.fiu.edu,reiher@cs.ucla.edu,
peters@ece.nwu.edu,veijalai@cs.jyu.fi,
wolfson@cs.uic.ecu
[†]Florida International University, [‼]Queensland Univ. of Technology-Australia, [•]Nokia,
[¨]Universität Karlsruhe- Germany, [┼]Univ. of California at Los Angeles, [‖]Northwestern Univ., ‡Univ.of Jyvaskyla- Finland,
[±]Univ. of Illinois at Chicago.

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 definitive definition of the infrastructure and its components, but 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. By flashing out the infrastructure requirements that all participants agreed upon and by casting light on the areas where no agreement was reached the report also serves the role of a guide to future research topics in the area of mobile wireless infrastructure. We hope that relevant funding agencies and companies interested in research in this area will consider these unanswered questions when they define new programs and projects in the mobile wireless area.

One issue of discussion in the workshop was the scope of the infrastructure. There was general agreement that the infrastructure should handle wireless cellular networks that rely primarily on single hop communications to a fixed base station that is connected to a wired network as well as 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 Bluetooth and ad hoc IP-based systems currently under development, but ideally it should be flexible enough to handle many other similar 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 that was discussed related to a functional definition of what constitutes "infrastructure." The workshop participants wrestled with different definitions. Here are the two most popular candidates:

- **System-based definition**: 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.
- **Application-based definition:** Infrastructure defines a set of assumptions that application developers can make about the components and behaviors of a wireless mobile network.

Clearly, either definition suggests that there is a common base of hardware, software, and protocols widely deployed for the purpose of servicing common needs of many applications. However, the purpose of the workshop was not to address hardware issues and these will not be discussed further in this report. Rather, the report touches upon protocols and software at the network and transport layers.

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. If particular functionality is only beneficial to a small number of applications, the functionality should be provided in those applications, not in the infrastructure.

There are certain fundamental differences in fixed network environment and wireless environment. Wireless terminals exhibit communication autonomy towards the network components and other terminals, meaning that they are normally detached from the network from time to 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 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 the infrastructure is the support for 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 and use other services connected to the Internet anywhere in the world, as well as communicating directly with other terminals. Mobility support also requires that the terminal is 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.

A security problem inherent in all wireless communication environments is that third parties can capture the radio signals while in the air. This problem cannot be avoided, because the signals must propagate to 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 address also other security issues, because mobile terminals are more vulnerable to loss or theft than fixed terminals. The Wireless Public Key Infrastructure suggested by the WAP Forum is one possible solution to many security issues in wireless cellular networks. However, security for MANETs remains largely an open issue.

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." Thus the infrastructure must be highly scalable.

Although it seems that future telecom networks and terminals that work with then will be most prevalent, there are other wireless environments that are emerging. These include MANETs and sensor networks. Wearable or ubiquitous computing and personal area networks (PANs) can also be included into the above categories.

Bluetooth, a typical MANET technology, has reached already the marketplace, embedded in hand-held terminals, PDAs, and also fixed devices, like printers and cash registers. Typical of these networks is that they reconfigure themselves whenever necessary, without the help of base stations or other central components. The infrastructure must 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, etc).

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. Given these technical trends we have identified the following properties of an infrastructure for a wireless mobile environment:

- The infrastructure must be complete. While the Internet is largely a success, mistakes were made in the definition of its infrastructure, *e.g.,* the lack of security. Also, 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 features that are lacking from the Internet.
- The infrastructure should be minimal. Making the infrastructure smaller increases the chances that its implementations will be correct. Furthermore, if infrastructure features are only added if absolutely necessary, there is less a chance that infrastructure providers will have to pay a high cost to support largely unused features.
- The infrastructure should be secure. A secure mobile wireless infrastructure will never be achieved solely by including some features such as cryptography and authentication. Rather, all infrastructure components must be designed with security in mind. Further, the security of their interactions must be considered.
- The infrastructure must be cost-conscious, since any service to be provided will require resources.

The rest of the report is divided into three sections describing infrastructure services in horizontal slices: network layer services, transport layer services, and middleware layer services.

II. Network Layer Infrastructure

A. Alterations to IP

Because mobile computers using a wireless infrastructure will need 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.

Mobile IP 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 are intercepted by the home agent. The home agent then resends these packets using encapsulation to the care-of-address of the mobile node.

The wireless infrastructure must provide IP support. The main question is whether it needs to support Mobile IPv6 only or whether it should also support Mobile IPv4. Also, a major issue to be addressed is how to deal with the security of care-of addresses.

B. Routing protocols for 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. Mobile IP

assumes that routers well suited to providing routing support are available in 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. Thus, the development of new routing protocols has become necessary. The IETF MANET working group [http://www.ietrf.org/html.charters/manet-charter.html] covers most of the ongoing 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 into two main groups: table-based routing protocols and demand-driven routing protocols.

Table-based ad hoc routing protocols like DSDV and OLSR are adaptations of classical routing protocols. Each node stores a routing table whose entries contain the interface used to reach each destination node or subnetwork and some measure of the distance to the destination via that link. Reachability and distance changes more frequently in ad hoc networks, so these protocols include mechanisms to cope with those differences.

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 discovered. A number of protocols, like DSR and AODV, reduce the message overhead incurred by selecting appropriate nodes to which messages are forwarded and by caching information about known routes.

No consensus has been reached on which ad hoc routing protocols are best. An interesting research issue is whether mobile gateway nodes between an ad hoc network and the Internet should have authorization or responsibility for changing the routing for all nodes within the ad hoc network.

C. *Multicast protocols for mobile and wireless infrastructure*

It is expected that future mobile and wireless networks will support group-based communication such as teleconferencing, multimedia, collaborative work, real-time workgroup, and distributed database access. 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 environment by allowing for node mobility and low-bandwidth, unreliable wireless links.

A multicast routing scheme should reduce the hand-off latency and optimize the multicast tree for stable regions that do not experience frequent group dynamics. It should handle frequent *join* and *leave* requests efficiently and without disturbing the ongoing multicast connections.

D. Investigation of whether other network services should be altered

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 protocols and routing protocols that could require alteration. Transport layer protocols are discussed in Section III. One other key Internet service that should be considered, though, is the Domain Name Service (DNS). DNS is a key component that allows translations of server names, familiar users, to IP addresses managed by routers. If the wireless infrastructure uses a routing solution akin to Mobile IP, mobility should not change the relationship between a name and address. Nor should the use of a wireless network. Intermittent connectivity of a mobile device does not mean that the mapping between its name and IP address should change.

DNS operates well at Internet scale because of its hierarchical nature and because of intermediate results caching. In this respect, some adjustments for the special circumstances of wireless networks and mobility may be useful.

E. Network management issues

A weakness of the Internet is that it lacks good 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 wired mobile environment will be more dynamic and difficult than the wired Internet, and even simple single-hop models of wireless network add further complexity, so that we require more management functionality then any wired environment provides to handle

handoffs and other features. Thus, proper features for network management are a vital part of the infrastructure services.

To the extent that ad hoc networks, personal area networks, and ubiquitous computing networks are included as being part of the mobile wireless world, these networks will require new management solutions. Examples of services to be included are dynamic addition/removal of nodes, self-initialization, fault-tolerance, diagnostics and configuration tools, etc.

Determining the proper set of network management services to include in the infrastructure is an open research question. Choices made for other infrastructure components will impact the choice of network management services.

F. 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 a mobile device's battery may be low, requiring special treatment. Various adaptations need to 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.

Proxies are single nodes, designed to provide services to mobile clients with limited capabilities. Other simple versions of an adaptive service include protocols with adaptive capabilities, single link services or gateway services.

Two communicating nodes could each use wireless links to reach the wired network, requiring the infrastructure to support troublesome links on both ends, even assuming the wired network is trouble-free. Proper handling of such circumstances requires some cooperation between the adaptive services near the endpoints. 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

III. Transport Layer Infrastructure

A transport service in the OSI sense offers a reliable end-to-end connection-oriented transfer of data between endpoints. 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 unable 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 different transport protocols, such as Transmission Control Protocol (TCP). However, TCP is not perfect for wireless networks since it makes assumptions about the behavior of the underlying packet network that are not true for wireless networks.

Modifications to TCP of relevance here include the Selective ACK (SACK) option and Congestion Control. These proposals have relevance for mobile terminals relying on wireless links, because they help distinguish between congestion and corruption, as well as saving bandwidth on the wireless links. However, if a connection exhibits simultaneously both problems, namely corruption and congestion, both of these proposals do not work appropriately.

This problem has led to a suggestion 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 Performance Enhancing Proxy (PEP). One connection is over the error-prone wireless link and the other one over the wireline links(s). The former is aware of the fact that the link used is wireless and the latter can assume wireline network to be used with its typical behavior. The two connections are exchanging data at a (transparent) gateway. 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 criticisms to this proposal. One central point is the end-to-end argument. Closely corresponding to the idea of breaking the connection into two, there could be a wireless TCP profile optimized for wireless links. WAP 2.0 specification addresses this possibility. 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.

IV. Middleware Layer Infrastructure

A. 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. A well-known and frequently used example is that of a user needing to print out something in an unfamiliar environment, e.g., a hotel. . Another example is ordering a taxi in a foreign city relying on Location-based services.

Thus, what is needed is an infrastructure component that enables computers to find services in an unfamiliar network. In the latter example, the geographic location (coordinates) of the terminal must be determined as well. Typically, the proposed architectures for service discovery 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), Jini, HaVI, Web Service Description Language, and UDDI.

The research issues in this area include approaches for the integration of different service discovery mechanism and the development of service discovery methods for ad-hoc networks, as well as for the roaming terminals

B. General Authorization Service

Many security services in networks require some form of access control to allow some users to perform certain tasks while prohibiting others. Because network requests for these services originates remotely, and because today's networks offer no certainty that the traffic is from the claimed source, end systems must perform some kind of authentication on the request.

Because it is so easy to forge network packets, the only practical solution currently available is to use cryptographic techniques. Scaling issues and the need for arbitrary users to authenticate themselves to a wide range of services favor public-key based methods, rather than symmetric cryptography. The simplest form of public key authentication is to assume that the entity trying to verify the creator of a message knows the public key of that creator. The public key can be used to verify that the message is legitimate if the sender signed the message with his private key. Assuming that the cryptography has not been broken and that the private key has not been compromised, only the owner of the public key could have produced the signature, so he must have sent the message.

However, this simple alternative fails in a large-scale world where a node cannot possibly have securely preloaded all public keys for all other entities in the network, so it may receive messages he cannot authenticate.

The standard approach to solving these problems is to use certificates. A certificate is a cryptographic package that validates that a public key contained in the package belongs to a specific entity named in the package. The validity of this binding is guaranteed because the private key of a well known trusted authority signs the certificate.

Some of the open problems here relate to whether any party can be sufficiently trusted by everyone to act as the certificate server, and whether the size of certificates makes them suitable for all kinds of communications

This entire approach is not suitable for sensor networks, where nodes often do not have personal identities. Further, the cryptographic operations required to sign messages and check their signatures are typically computationally expensive, which translates to using a significant amount of battery power. Alternate solutions for providing security in these kinds of wireless mobile networks are required.

C. Location Management

Location-based services (LBS) are perhaps the most important future application of mobile and wireless systems. Location-dependent services use the actual physical location of the terminal to deliver contents or for tracking purposes. The location can be determined by the GPS-enabled terminal itself, by the network infrastructure, or by a combination of these (Assisted GPS). Wireless telecom networks keep track of the location of the terminal in order to be able to set up an incoming call. The finest resolution it keeps is at a cell level.

Location management, the management of transient location information, is an enabling technology for location-aware content delivery and also a fundamental component of other

technologies such as fly-through visualization (the visualized terrain changes continuously with the location of the user), 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 in which cell a user is (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 record, as well as how to ensure their privacy.

The main limitations of current works are that the only relevant operations are point queries and updates that pertain to the current time, and they are only concerned with cell-resolution locations. For broader wireless mobility, queries are often set oriented or they may be temporal, and triggers are often more important than queries**.**

In order to address these problems (and others) a Location Management System (LMS) needs to be part of the infrastructure of mobile and wireless systems. The capabilities required of an LMS include support for modeling of location information, uncertainty management, spatio-temporal data access languages, indexing and scalability issues, data mining (including traffic and location prediction), location dissemination in a distributed/mobile environment, privacy and security, fusion and synchronization of location information obtained from multiple sensors.

V. Conclusion

This report reflects the workshop's participants broad consensus with respect to defining and establishing a common Infrastructure for the discipline of Mobile and Wireless networking. The full report will appear in a Springer-Verlag volume together with the workshop papers . This volume will include one keynote paper, three invited papers and 11 regular papers.

The report identifies three broad areas of research priorities, namely: (a) Network layer Infrastructure, (b) Transport layer infrastructure and (c) Middleware layer infrastructure. The full report discusses a number of additional topics falling within these three areas on which the workshop participants reached no consensus. The report also stresses the fact that the various infrastructure components must work all together and that these components must be tested in a high scale environment before being adopted.

Acknowledgements

We would like to thank the National Science Foundation and Drs. Mukesh Singhal and Randy Chow for their support. The work reported in this paper was supported by the U.S National Science Foundation under Grant CCR-9986080. Any opinions, findings, and recommendations expressed in this paper are those of the authors. We would also like to thank all the participants of the workshop for interesting and engaging discussions**.**