Mobile Access to the Internet: Improving Performance with Movement Predictions

Claude Castelluccia, Vincent Roca

June 24, 1997

INRIA Rhône-Alpes - projet SIRAC
655, av. de l'Europe; 38330 Montbonnot St Martin; FRANCE

e-mail: (Claude.Castelluccia — Vincent.Roca)@inrialpes.fr

1 Introduction

The fast development of wireless telephony is a sign of a new erea. Many experts predict that the next century will be the century of mobile services. In the near future, it will be possible to access information anywhere and at anytime. The rapidly expanding technology of cellular, wireless LAN, and satellite will make it possible for an user, equipped with a laptop, to connect to the Internet in order to book a plane ticket, read electronic mails, or surf the web, while roaming the world.

These connections will, most probably, not rely on an homogeneous network but on an overlay mobile network composed of several different sub-networks (GSM, Wireless LAN, satellite networks, etc.). Because these sub-networks use different technologies, the technical characteristics (bandwidth, transmission delay, loss rate, etc.) of this virtual mobile network will largely vary. This heterogeneity creates new kinds of problems to the communication layers and applications.

The goal of this paper is to introduce these problems and to sketch a few possible directions to solve them. The rest of this paper is structured as follows. Section 2 presents the future wireless overlay network, providing continuous connectivity to the Internet. It presents its main characteristics and its impacts on the communication layers. Section 3 describes our approach and solutions to the problems of mobile access to the Internet. Section 4 presents our conclusions.

2 Mobile access to the Internet

2.1 Description of the future wireless overlay network

The figure 1 presents the mobile access environment that will someday be available [8]. This environment is composed of several wired and wireless technologies, providing a continuous access to the Internet.

With this environment, a mobile could move across the world and still remain connected to the Internet. To illustrate this futurist scenario, let us consider a typical working day of a business man, John. John has to visit one of its client in city B today. He takes the train at his home town, A, but since he is a very busy man, he can not enjoy the landscape during the 3 hours trip to city B, but has to do some work. Therefore, he needs to get connected to his company by means of the Internet. He runs on his laptop several applications such as a multiuser audio-conference application, an electronic mail application, a remote login and a file transfer tool. During the trip, John accesses the Internet via several types of links:

- From A to t_5, John uses a bidirectional 9.6 Kbps/s GSM connection.
- From t_1 to t_2 he has also access to a Wireless LAN (WLAN), e.g. an AT&T Wavelan), at 2 Mbps and covering a 150 meter area. This WLAN has an incoming 12 Mbps connection to the Internet via a satellite link, and an outgoing connection via the GSM network.
- From t_2 to t_3, John can only access the Internet via a bidirectional GSM link.
- From t_3 to t_4, he has again access to a WLAN network.
- From t_4 to t_5, John can only access the Internet via a bidirectional GSM link.
- From t_5 to t_7, he looses connectivity (the GSM network is not connected to the Internet) .
- From t_7 to t_8 he can enjoy a Xircon Netware WLAN at 1 Mbps, covering a 240 meter area, and connected to the Internet through a 2 Mbps ATM network.
- Between t_8 and t_9, he looses once again connectivity to the Internet.
- And finally, when arrived at B, he can use a classic local Ethernet at 10 Mbps.

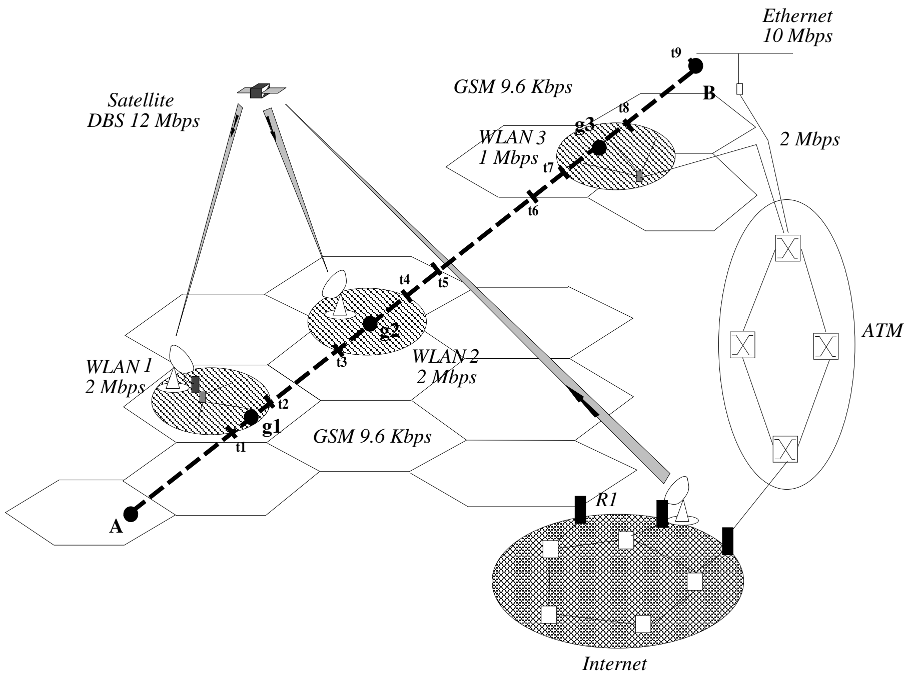

Figure 1: *An example of mobile-capable environment for an Internet access*

This scenario shows that:

- heterogeneity of Internet connectivity is a key point when considering mobile hosts.
- connectivity can be far from perfect with the possibility of unreachability areas.
- several communication infrastructures can be available which raises the problem of choosing the most appropriate interface.

Besides this example includes two different dimensions:

- wireless communications (GSM, WLAN, satellite, etc.), and
- mobility support (enable the mobile host to "transparently" access the hosts on the Internet).

Note that mobility can also be a feature within classic wired networks. For instance, an user having several offices may want to keep its laptop configuration when moving from one place to another so that other hosts can still access him.

Our work addresses essentially the second point, namely the mobility support in a heterogeneous and non perfect environment.

The following section analyzes more in details the characteristics of mobile connections to the Internet.

2.2 The characteristics of mobile environments

The mobile environments differ from fixed environments by essentially three aspects:

- *high variability of the wireless connection characteristics*

  The characteristics of wireless networks are much different from that of wired networks. Globally, wireless communications have lower qualities: lower bandwidth, higher and widely variable error rates and frequent disconnections due, for example, to cell transitions. Unlike wired networks, wireless networks are subject to temporary high user concentrations in a given cell, therefore reducing the bandwidth available to each of them. All these reasons make the foreseeing of a given wireless network performance highly speculative.

- *heterogeneousness of the sub-networks*

  In addition to the intrinsic variability of the features of a given technology, the mobility adds another dimension. It enables the successive choice of different technologies according to the communication infrastructure available at the moment, and to the criteria of an user [1].

- *asymmetry of connections*

  In mobile environments, due to technical constraints, the incoming and outgoing flows can use two different technologies. This is the case for some satellite networks [12]. In addition, different data flows coming from different applications and having different requirements can make use of different networks. Both of these characteristics are unusual in classic configurations.

2.3 Impacts of the mobility on protocol layers and applications

Mobility has many impacts on the communication stack [10, 7]. We can identify three concerned levels: the network layer, the transport layer, and the applications.

2.3.1 Network layer: impacts on routing

- *Mobile IP; a point to point routing protocol to support mobility*

  IPv4 and IPv6 extensions to support mobility are currently being analyzed by the IETF [4, 9]. This proposition, called mobile IP, aims at hidding a host mobility by making it still accessible by other hosts at its home address. Several propositions have been done. Most of them introduce the concept of "foreign" and "home" agents (respectively FA and HA). When a mobile host (or MH) arrives in a remote network, it subscribes to the local FA. This latter informs the HA of the new MH address. All the packets destined to the MH are thus intercepted by the HA which redirects them to the FA which hands them out to the MH. This concept of agent, even if non optimal from a routing point of view (a cut-through routing optimization is currently defined), hides mobility while keeping current transport protocols and applications. While mobile IP provides a solution to mobility, it has several drawbacks:

  - it can lead to connection losses when moving to another foreign network, and
  - it does not take into account the possibility of network interface choice.

- *Multicast routing*

  Multicast routing as defined in [6] relies on the Reverse Path Forwarding (RPF) algorithm. RPF creates the spanning tree for each source of multicast stream. If it works fine in a static environment, introducing mobile hosts creates efficiency problems. If the source of a multicast stream changes of location, then a new spanning tree must be created, and until the pruning of unused branches takes place, the whole multicast backbone is flooded by this multicast flow. In case of a mobile receiver, this latter needs to unsubscribe and then subscribe to the multicast group each time it moves, therefore creating unreachable periods and, with certain applications, the need to go through a state recovery procedure. Eventually, the RPF algorithm is inherently incompatible with asymmetric connections required by certain wireless technologies.

2.3.2 Transport layer: inadequacy of classic protocols

The mobile IP solution makes the host mobility transparent to the transport protocols and applications. But this solution is quite inefficient because point-to-point, multicast transport protocols and existing applications are not really adapted to wireless and mobile environments.

For instance using TCP in a wireless environment leads to many inefficiencies [2, 5]:

- the packets losses, caused by interferences, are usually misinterpreted as being caused by a router congestion. Consequently TCP enters a new slow-start cycle to reduce its outgoing throughput. Unfortunately it does not help improve the situation.

- Also, TCP has currently no selective retransmission feature which is penalizing in a highly disrupted environment.

It is important that there are more interactions between the network and transport layers, so that the latter can be informed of cell transitions, or of the underlying communication technology and adapt properly.

2.3.3 Applications

For simplicity and philosophical reasons, most of applications are designed independently of the underlying communication infrastructure. This layering is the source of inefficiencies.

This is all the more true in a mobile environment where the network features can quickly vary in large proportions [11]. A mobile application should:

- take into account the heterogeneity of the crossed networks, e.g. by changing the coding algorithm when moving to a low-bandwidth GSM network during an audio-conference,
- provide a certain level of informations to the user, e.g. warning him when approaching a non-covered area, enabling him to choose among several communication technologies according to criteria like pricing, raw performances, etc.

3 Our approach to mobile access to the Internet

3.1 Our vision of the future mobile Internet access environments

We believe that in the near future three types of mobile access environments will be available, namely: (1) local area access which will be available within a building or a room, (2) metropolitan access which will cover a campus, a city and (3) wide area access which will cover a region, a country, a continent and eventually the planet.

Those three kinds of environments have different characteristics which generate problems of different orders.

- In local area environments, the connectivity is important, the available bandwidth is plentiful (few Mbps), delays are small, the mobility is usually limited and the networking technology is homogeneous. Those characteristics are in fact very similar to the those of wired links. Little efforts are required to adapt existing protocols to this type of environments.
- In metropolitan environments, the connectivity is (or will be) also important, although less important than in a local access environment, the available bandwidth is large (few hundreds of Kbps), the delays are small, the mobility is quite limited (pedestrians, tramways, cars,...) and the networking technology is heterogeneous. The impact of mobility, in this type of environments, is therefore also limited.
- In the wide area environments, the connectivity is weak and variable (disconnections would be usual, especially when the mobile will be roaming in sparsely populated area), the available bandwidth is small (few Kbps), the delays are important, the mobility is important and the networking technology heterogeneous. The impact of mobility on the communications layers is very important in this type of environment.

  However, we believe that unlike the two previously described types of environments, host movements are highly predictable in wide area environments. In fact, on a long trip, a mobile can usually predict its route using a road-map, a GPS system or by looking at the train stops though its destination. This movement predictions could be profitably used by the mobile to determine in advance the characteristics of the access environments available to him during its trip.

  The figure 2 shows the bandwidths and delays of the access environments that our business man of example 2.1 will encounter during its train trip. Those predictions could be loaded by the mobile in A and used to optimize the communications.

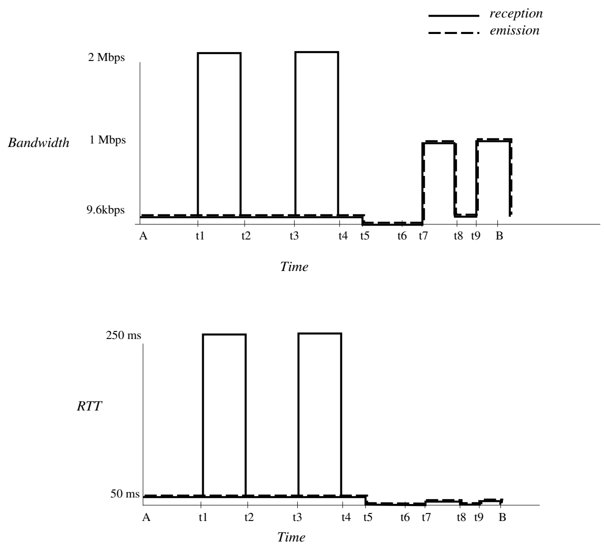

Figure 2: *Bandwidth and delay predictions of our mobile environment*

In the following session, we show how this information can be profitably used to solve most of the problems generated by mobility on each of the communication layers.

3.2 Using movement predictions in mobile environments

3.2.1 Basic Principles

Our approach consists in utilizing the movement predictions of the mobile to estimate the characteristics of access environments that it will encounter during its trip and use this information to optimize communications.

These estimations are optimistic; they inform the mobile of the best characteristics it can pretend to at each connection point and not the real characteristics it will get. In other words, those estimations won't tell the mobile what it will get, but instead what it will not get. For example, when moving from a WLAN to a GSM network, the mobile could anticipate that its bandwidth will drop drastically from few Mbps to few Kbps. This information, even inaccurate, can in fact be very useful to optimize protocol mechanisms, as shown in the rest of this section.

3.2.2 Network layer

- With mobile IP, the mobile must register to a new foreign agent when it moves to a new subnet. This registration phase takes some time and consequently can lead to unreachability periods of the mobile (the mobile is between the previous and new subnets). Those unreachability periods generate losses which degrade the communication performances. By anticipating its movement, a mobile can predict the transitions from one subnet to another one and optimize them. For example, by predicting its transition to a new subnet, the mobile can register in advance to the new foreign agent. The packets for the mobile would then be multicasted to the new and old foreign agents. The mobile could then unsubscribe from its previous foreign agent as soon as it starts receiving packets from its new foreign agent. As a result of this mechanism, the unreachability periods and resulting losses would be minimized.

- Predictions can also be used to solve the problems, exposed in section 2.3, generated by mobility on the multicast routing protocols [6]. Predictions could be used by the IGMP (Internet Group Multicast Protocol) to pre-establish the spanning trees when the source is moving. This would avoid the flooding of the whole Mbone (Multicast backbone) each time the source changes subnets. Protocol efficiency would thus be greatly improved.

  As explained in section 2.3, when the receiver is moving, losses can be generated by unreachability periods of the mobile due to its transition from one subnet to another. Solutions similar to the one proposed previously for the unicast case can be used to solve this particular problem.

3.2.3 Transport layer

The estimation of the access point characteristics that a mobile visits during its movement can be of great help to improve the performance of transport protocols. As explained in section 2.3.2, transport protocols suffer considerably, in mobile environments, from the variations of network characteristics and from the numerous losses.

If the mobile knows in advance the characteristics of its access environment it can adapt its transport mechanisms accordingly and use the correct version of the protocol. For example, if the mobile knows that it accesses the Internet though a satellite link, it can use the long-delay paths extension of TCP [3]. Its performance would be greatly improved. Similarly, when the mobile anticipates a transition from a WLAN to a GSM network, it can reduce its transmission window accordingly. Those examples demonstrates that estimations, even inaccurate, can be very beneficial to protocols. They also show that protocols need to be highly adaptive.

Also, as explained in the previous section, losses at subnet crossing can be reduced by the use of predictions. The transport protocols would benefit from this reduction.

3.2.4 Applications

Predictions can also be of great interest to the applications. They can use them to adapt their coding algorithm and manage their resources properly. For example, when a mobile anticipates that it will soon cross an area that does not provide him any access to the Internet, it can inform the user of an incoming disconnection period by a blinking icon. This signaling would significantly reduce user's frustrations that could then work locally until a new access point becomes available.

When an user runs several applications on a mobile, predictions could be used to schedule the different flows or applications in order to optimize the network resource utilization. It could, for example, use satellite links, which provide high bandwidth but large delays, to receive asynchronous bulk data, such as emails or files, and GSM links, which provide low bandwidth but low delays, for low-bandwidth interactive applications, such as the telnet application. This scheduling can also be done dynamically. For example, in the scenario presented in section 2.1, if the user requests the transfer of a large file between t_4 and t_5 (it accesses then the Internet via A GSM link), it would be reasonable to postpone this request until the mobile accesses the Internet via a high bandwidth link, such as between t_7 and t_8.

4 Conclusions

In this paper, we propose a solution to improve the performance of protocols in wide area mobile environments. A mobile, that roams relatively long distance, can usually anticipate the path that it will take toward its destination. This information can then be used to estimate the characteristics of the Internet access environments that it will encounter during its trip. We show that those predictions can be used to solve most of the problems generated by mobility on communication protocols and to optimize the network resource utilization.

References

[1] Mary G. Baker, Xinhua Zhao, Stuart Cheschire, and Jonathan Stone. Supporting mobility in mosquitonet. In *Proceedings of USENIX Technical Conference*, San Diego, CA, January 1996. USENIX.

[2] Hari Balakrishnan, Venkata N. Padmanabhan, Srinivasan Seshan, and Randy H. Katz. A comparison of mechanisms for improving tcp performance over wireless links. In *Proceedings of the ACM SIGCOMM Conference*, pages 256–269, Stanford University, CA, August 1996. ACM.

[3] D. Borman, R. Braden, and V. Jacobson. Tcp extensions for high performance RFC-1323. In *DDN Protocol Handbook*. U.S. Department of Defense, May 1992.

[4] Editor C. Perkins. IP mobility support RFC-2002. In *DDN Protocol Handbook*. U.S. Department of Defense, October 1996.

[5] Ramon Cacares and Liviu Iftode. Improving the performance of reliable transport protocols in mobile computing environments. *IEEE Journal on Selected Areas in Communications*, 13(5), 1995.

[6] S. Deering, C. Partridge, and D. Waitzman. Distance vector multicast routing protocol RFC-1075. In *DDN Protocol Handbook*. U.S. Department of Defense, January 1988.

[7] George H. Forman and John Zahorjan. The challenges of mobile computing. *IEEE Computer*, pages 39–47, 1994.

[8] Randy H.Katz and Eric A. Brewer. The case for wireless overlay networks. In *Proceedings of the SPIE Multimedia and Networking Conference*, San Jose, CA, January 1996.

[9] Charles E. Perkins and David B. Johnson. Mobility support in IPv6. In *Proceedings of the Second Annual International Conference on Mobile Computing and Networking*, Rye, N.Y., November 1996. ACM.

[10] M. Satyanarayanan. Fundamental challenges in mobile computing. In *Proceedings of the Fifteenth Annual ACM Symposium on Principles of Distributed Computing*, pages 1–7, Philadelphia, PA, May 1996. ACM.

[11] Terri Watson. Application design for wireless computing. In *IEEE Workshop on Mobile Computing Systems and Applications*, Santa Cruz, CA, December 1994.

[12] Yongguang Zhang and Son Dao. Integrating direct broadcast satellite with wireless local access. In *Proceedings of the First International Workshop on Satellite-based Information Services*, Rye, N.Y., November 1996. ACM.