Terminal-oriented solutions for seamless service delivery via heterogeneous radio access networks

S. Murphy, M. Atiquzzaman, J. Fitzpatrick, J. Murphy

***Abstract*—In this paper, options for providing seamless service delivery to mobile terminals via heterogeneous radio access networks are considered. Despite its current popularity, the Mobile IP solution for mobility management of mobile hosts is discounted due to problems associated with introducing the required network infrastructure and poor handover performance. Rather, the solution described here comprises of a number of key technologies working together. More specifically, the paper argues that the most appropriate solution comprises of the following: adaptive applications, an abstraction layer between application and transport layer, an intelligent transport layer with multi-homing and multi-path support and a means to obtain current state information about each of the available access networks. While work has been ongoing on each of these individual components little has been done to address how they can all be integrated and consider how they may perform in the context of different radio access technologies, potentially operated by different entities. This paper attempts to address some of these issues.**

***Index Terms*—Heterogeneous wireless access networks, Handover, Mobile Networking**

I. Introduction

It has been clear for some time now that multiple options for wireless access will exist in the future: indeed, there are already markets where multiple wireless access options exists – e.g. South Korea offers WiBro services as well as 3G cellular services and WLAN based access – and multimode terminals are starting to appear in the marketplace.

Many within the community subscribe to the vision that these multiple access networks will support some interoperability such that services can be accessed seamlessly via the most appropriate access network (for some definition of the most appropriate access network, e.g. low delay). However, it is still unclear how the business relationships will evolve in this context: some are of the opinion that there will be many wireless access networks operated by different entities and devices will be able to seamlessly switch between them. However, others are of the view that individual network operators will offer multiple wireless access options and handover between the different access networks will be limited to a single operator's domain. In any case, it is clear that there is a need for handover between heterogeneous wireless access networks.

An important issue for this scenario is that different Radio Access Networks (RANs) will have different capabilities and associated costs. More specifically, different RANs will be able to support different data rates, with potentially different delay characteristics and may have different usage costs. Indeed, the picture is more complicated than this: the service that even a single RAN may be able to offer can vary significantly, depending on the location of the user. For this reason, then, it is important that applications operating in the context of heterogeneous RANs have some adaptive capabilities and may even need to have some awareness of what access network is being used by the device at any time.

While there have been many contributions to the literature to address different aspects of this problem, it is clear that a rethink of the accepted TCP/IP layered model is required for this context. Here, an alternative solution to the standard model is described which provides more flexibility and may be more appropriate for a world comprising of many different wireless and mobile devices and different RANs.

The document is structured as follows. Section 2 discusses work related to that described here. This is followed in section 3 by a discussion of the requirements of mobile and wireless applications, which is tempered by the capabilities of current and future wireless interfaces. This is then followed by a discussion of a middleware which can encapsulate the different capabilities of different transport protocols, providing an abstraction which can hide details of transport layer protocols from application developers: this is particularly useful in a mobile context. Finally, there is a short conclusion.

II. Related Work

The issues associated with deficiencies in the TCP/IP protocol stack for wireless networks are well known and many enhancements/extensions to the stack have been proposed for the mobile context. Indeed, many ideas have been proposed to address these problems, some of which employ existing protocols in novel ways and some of which involve the

S. Murphy, J. Fitzpatrick and J. Murphy are with the School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland.

M. Atiquzzaman is with the School of Computer Science, University of Oklahoma, Norman, Oklahoma, USA.

Corresponding author: sean.murphy@iname.com

development of entirely new protocols.

The Mobile IP (MIP) [1,2] is probably the most known protocol in this area. It leverages ideas from the cellular telephony world to address the problems of location management and handover in wireless IP networks: it operates at the network layer, which means that many applications should work normally over MIP. This solution, however, has a number of problems and deployment of MIP is taking some time. One of the most critical issues for MIP is that it requires significant changes to network infrastructure: history shows that technologies requiring substantial infrastructure modifications often have limited success – e.g. RSVP/Intserv [11]. Other issues with MIP relate to the fact that it has been designed to be independent of the wireless interface, even though knowledge of the wireless interface has been shown to be useful; also, there have been issues with the handover latency for MIP, although solutions to these problems have been developed, which typically involve adding more complexity to the network.

As MIP rollout is somewhat underwhelming, researchers are starting to consider alternative options. The solutions that are receiving most interest at present are those in which more intelligence is located in the end-terminal and the amount of network intelligence is reduced [12]. Indeed, these types of solutions are consistent with the end-to-end model, which stipulates that the network core should have a minimal amount of intelligence.

Two promising ideas in this space are that of transport layer handover, typically combined with Dynamic DNS for location management and a solution based on the introduction of a new abstraction layer between the transport and network layers. Each of these will be discussed separately.

A number of different proposals have been tabled to provide handover support at the transport layer. TCP-Migrate by Snoeren et al was one of the first contributions. The core idea behind this proposal is that a TCP connection can be recommenced once a device switches IP address by sending an appropriately modified TCP SYN message with an identifier for the TCP connection. This solution, however, suffers from some issues with latency and has no support for protocols other than TCP [3].

Another approach which has been proposed is one based the Stream Control Transmission Protocol (SCTP) [4,5,13,14]. This transport layer protocol enables a so-called association – effectively, a transport layer connection in SCTP parlance – to span multiple IP addresses. Handover can be effected, then, by switching from one IP address on a device to another, the former being associated with one RAN and the latter with another. The advantage of this solution is that it results in much lower handover latency than the Migrate approach; it also has some support for mode of operation which does not have head of line blocking and can be used for applications which have more stringent delay requirements, this extension to SCTP is known as Partial Reliable SCTP (PR-SCTP). The key disadvantage of this solution is that it requires that applications be rewritten to make use of this new transport protocol.

The Datagram Congestion Control Protocol (DCCP) [6] is a third transport layer protocol which is receiving some interest and is currently being standardized within the IETF. A key benefit of this scheme is that it has multiple modes of operation which behave differently, but can work well in the presence of TCP traffic. More specifically, it provides some support for delay sensitive applications which have some ability to adapt: as such, it seems particularly suited to mobile and wireless applications. Multi-homing operation was considered when designing the protocol – partially for handover support – using a mechanism similar to SCTP, but was initially ruled out due to security concerns. However, there is still some interest in seeing whether these can be addressed such that this new protocol can provide good handover support.

Another solution which is receiving much attention right now is that of the so-called SHIM layer [7]. This is an extra layer of abstraction that is introduced between the transport layer and the network layer. It has the purpose of separating the use of IP addresses as both transport layer connection identifiers and interface identifiers. Essentially, it operates by tracking transport layer connections and providing mapping functionality such that when IP addresses change, the transport layer is isolated from this change and the IP address that was used to initiate the connection is passed to the transport layer in place of the currently active address. It also has signaling mechanisms similar to those employed by the mobile SCTP variants which enable the hosts to maintain a list of valid IP addresses for each other.

The key advantages of the SHIM concept are that no modifications are required to existing applications and it can be realized in the end terminals, thus requiring minimal network support. However, as the current solution leverages some of the increased flexibility offered by IPv6 – the ability to add arbitrary headers, essentially – it can only operate with IPv6 connections. While there is increasing use of IPv6 today, especially in the orient, it will still be some time before large amounts of content is available via IPv6 and hence IPv4 solutions to these problems are also required.

The IETF is not the only standards organization which has an interest in solving this problem. The bodies which ratify standards for cellular networks, 3GPP and 3GPP2 are developing their own solutions, although they are discounted here as their solutions are typically complex and are tightly integrated with network architecture. The IEEE is also developing a solution, within the context of the 802.21 activity, which appears to have more promise.

The 802.21 initiative has two important aspects [8]: firstly, it provides a standardised call-back based mechanism by which different functions within a device can be aware of the state of the different link layer attachments and secondly, it provides a means by which a terminal and a network can communicate regarding the different handover options in a heterogeneous networking context. The first mechanism enables, say, the network layer and, if necessary, the transport layer to be aware of different events at the link layer, e.g. link going down, link active etc. while the second mechanism can be used to enable a node to know what different RANs are available in a given area and, for example, what frequency/channel they are operating on. It is important to note that the 802.21 idea comprises of both terminal and network functions.

A final initiative which is of note is the so-called Unified Link Layer being developed by EU-funded GOLLUM project [9]. This initiative has developed a single abstraction for different types of link layers which can be used by OS programmers to simplify access to multiple RANs. The abstraction is an enhancement over the current approach in which OS developers typically have to work with many different device drivers to realize this capability. Indeed, the consortium has shown how their abstraction can be used to easily develop a rudimentary 802.21 implementation. [10]. However, they limit their focus to the link layer and do not give much consideration to the higher layers of the protocol stack.

It is clear from the above, then, that there is much interest in developing appropriate solutions which provide good application support in the context of mobility within heterogeneous wireless access networks. However, it is also clear that the community has not yet reached any consensus on the most appropriate solution, or indeed, if a single solution can cater for all of the different use cases.

III. Communications Requirements in Future Heterogeneous Mobile/Wireless Networks

The constant cost reduction of computing and communications components is already resulting in powerful mobile devices, capable of running quite sophisticated applications. However, it is clear that the era of powerful mobile/wireless devices running complex, potentially distributed applications is still in its infancy.

Many believe that the future will see increasing diversity in mobile devices, made possible by ever-decreasing component costs: different devices will have different form factors, different interfaces, different wireless access options, etc. There will be more mobile and wireless devices in vehicles, about the person, in public spaces, etc. These heterogeneous devices running different applications will have a broad mix of requirements of future networks: they will have differing data rates, delay sensitivity, adaptive characteristics etc.

Advances in network technology continue unabated. While new radio technologies are providing for faster wireless connectivity, often with increased range and mobility, there are still large parts of the world for which slow and costly satellite access is the only option. Mesh networking technology is a very promising technology which is set of have a profound impact on the way wireless networks will evolve in the future. In their infancy at present, mesh networks will provide for cheap, low cost coverage, more resilient networks and reduced management costs. However, mesh networks will likely deliver quite varied performance – there can be large variations in network performance arising from the multi-hop nature of the systems, distances to wired gateways and variation in distance between transmitters and receivers. In general, this can pose problems for applications which may not be able to realistically adapt to arbitrary network conditions.

As is clear from experience with the wired Internet, adaptive applications are necessary to operate well under varying network conditions. Non-delay sensitive applications, typically operating over TCP, adapt their maximum bitrate to the available resources, VoIP applications adapt to prevailing network conditions by adjusting their interpacket spacing and video delivery applications support some adaptation in terms of frame rate and/or spatial resolution.

As conditions on wireless networks can be even more variable, fundamentally due to the variable channel characteristics exhibited by wireless networks, adaptive applications are even more critical. Wireless networks differ from their wired counterparts in other ways. One key point is that multiple wireless access networks can be available simultaneously – indeed, they can be accessed simultaneously. This offers one possibility for providing increased network capacity if a single wireless access option is deficient.

From the above discussion, then, some requirements of future mobile and wireless networks are clear:

- Applications will need to be able to adapt to variations in end-to-end network performance
- Some support for existing applications is necessary;
- Devices will be potentially connected to multiple RANs using multiple IP addresses simultaneously: this can create an opportunity for multi-homed wireless protocols;
- Devices could leverage multiple RANs simultaneously in order to achieve improved performance: protocols should provide support for this;
- Devices will need to be able to operate with existing IPv4 infrastructure as well as the emerging IPv6 infrastructure.

It is clear that the current TCP/IP model has some difficulty meeting the above requirements: some alternative thinking is required.

IV. Middleware to Support Multiple Transport Layers

Many of the proposals discussed in Sec II address parts of the requirements described in Sec III. However, none provides a very compelling view of how the future of wireless networking will look.

The requirements identified above are quite diverse and hence a number of different techniques should be combined to realize an appropriate solution. As deployment problems have been highlighted with solutions requiring substantial network modifications, the emphasis here is strongly on terminal oriented solutions.

Here, some thoughts on alternative ways to provide communications support to application developers which have the potential to provide greater flexibility, albeit at the cost of greater complexity.

The essence of the approach is to provide an alternative abstraction for communications which shields the application developer from the specifics of the transport layer protocol in use. A middleware layer then performs a mapping from a generic communications identifier into specific transport layer protocols.

While many variants of this idea have been proposed in different contexts heretofore, eg [15], there are two related points which differentiate the ideas proposed here. Firstly, an essential point of the work proposed here is that mobility support is critical: the mechanism must provide good support for seamless (subject to constraints) delivery of services via multiple, potentially heterogeneous wireless access networks. Secondly, the mechanism must provide support for communications between devices with different capabilities. More specifically, as there are an increasing number of transport layer protocols available, it is realistic to assume that not all devices will support these hence some capability exchange between devices within the network is necessary.

The middleware, then, will have the function of determining the most suitable means for communication between remote hosts. This decision will be taken based on both the capabilities of the remote hosts and the requirements specified by the application developer. This can involve selecting the appropriate transport layer protocol(s) and wireless interface(s) for the communication; the middleware could also intelligently adapt either wireless interface or transport protocols if necessary.

An attractive aspect of this approach is that it can be introduced in an incremental fashion. The default operation of the middleware could be to use TCP or UDP connections – depending on application developer requirements – which are almost universally available. The middleware could also query the remote host via some known TCP port to obtain information on the capabilities of the remote host: if the remote host does not provide such services, then an aggressive timeout mechanism can be invoked and the communications can be initiated using the default mechanisms. Indeed, some basic learning mechanisms could be added, such that an end terminal could know which remote hosts support the middleware.

Another important characteristic of this middleware is that it will need to provide accurate information to applications which reflect the diverse operating conditions: thus the application can determine how to adapt. This can be done via specific middleware query mechanisms as well as some event based mechanisms.

As is usually the case, security introduces a significant number of issues and further research is necessary to determine how well these can be resolved. The proposed middleware can operate in the context of existing security mechanisms, although providing support for connection transfer from one secured TCP connection to another – because a new IP address works better, for example – would prove difficult under the envisioned model.

TLS or DTLS would appear to be a natural solution to the security problems arising in this context. However, the important issue of dealing with changing IP addresses is not sufficiently addressed. This remains an open issue which is under study in a number of contexts.

V. CONCLUSION

The current TCP/IP stack is clearly exhibiting significant limitations in the context of large amounts of mobile devices having frequently varying internet attachment points. An important challenge for the future internet is to devise a means to provide good mobility support and good support for legacy applications. In this paper, an argument for a middleware which encapsulates multiple transport layer mechanisms with a view to providing seamless mobility support in the context of changing IP addresses has been proposed. The key advantage of this middleware is that it can separate the requirements of application developers from the details associated with realizing novel transport protocols. Many issues remain to be resolved with this concept, most of which have been identified throughout the discussion: these will be the subject of future research.

ACKNOWLEDGMENTS

S. Murphy wishes to acknowledge the support of the Informatics Research Initiative of Enterprise Ireland. The work of M. Atiquzzaman was supported by NASA grant no. NNX06AE44G. J. Fitzpatrick wishes to acknowledge the support of the Irish Research Council for Science, Engineering and Technology.

REFERENCES

[1] C. Perkins, "IP Mobility Support", http://www.ietf.org/rfc/rfc2002.txt, Oct, 1996.
[2] C E Perkins & A Myles "Mobile IP" IEEE Communications Magazine, May 1997, Volume 35.
[3] A. Snoeren and H. Balakrishnan, "An End-to-End Approach to Host Mobility", Proc. 6th ACM MobiCom, August 2000.
[4] R. Stewart, Q. Xie, K. Morneault , C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, V. Paxson."Stream Control Transmission Protocol", RFC2960, October 2000.
[5] R. Stewart, C. Metz, "SCTP: new transport protocol for TCP/IP", IEEE Internet Computing, Dec 2001.
[6] E. Kohler, M. Handley, S. Floyd , "Designing DCCP: Congestion Control Without Reliability," Proc. ACM SIGCOMM 2006.
[7] E. Nordmark and M. Bagnulo, "Level 3 multihoming shim protocol", Internet Draft (work in progress), IETF, May 2006. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf shim6-proto-05.txt.
[8] V. Gupta and D. Johnston, "A Generalized Model for Link Layer Triggers", IEEE 802.21 proposal.
[9] M. Bandholz, A. Gefflaut, J. Riihijärvi, M. Wellens and P. Mähönen, "Unified Link-Layer API Enabling Wireless-Aware Applications," Proceedings of PIMRC'06, Helsinki, Finland, September 2006.

[10] M. Wellens, J. Riihijärvi, K. Rerkrai, M. Bandholz, P. Mähönen, "Enabling Seamless Vertical Handovers Using Unified Link-Layer API," Proceedings of IEE Mobility Conference, Bangkok, Thailand, October 2006.

[11] M. Handley, "Why The Internet Only Just Works," BT Technology Journal, Vol 24, No 3, July 2006.

[12] W. M. Eddy, "At What Layer Does Mobility Belong", IEEE Communications Magazine, Oct 2004.

[13] S. Fu, L. Ma, M. Atiquzzaman and Y. Lee, "Architecture and Performance of SIGMA: A Seamless Handover Scheme for Data Networks", IEEE International Conference on Communications (ICC 2005).

[14] J. Fitzpatrick, S. Murphy, J. Murphy, "SCTP based Handover Mechanism for VoIP over IEEE 802.11b Wireless LAN with Heterogeneous Transmission Rates", Proc. of IEEE International Conference on Communications (ICC 2006), June 2006.

[15] J. Salz, A. C. Snoeren, and H. Balakrishnan, "TESLA: A Transparent, Extensible Session-Layer Architecture for End-to-End Network Services," 4th Usenix Symposium on Internet Technologies and Systems, Seattle, Washington, March 2003.