In Search of the All-IP Mobile Network

PETER NEWMAN, NETILLION INC.

Abstract

Much of the recent standardization activity from the 3G mobile wireless community has been directed at the "all-IP" network to support multimedia services. This "all-IP" effort defines a separate IP multimedia domain attached to the mobile network. In this article we consider what a mobile network might look like were it based on IP technology within the core of the network — if it were in fact all IP.

Introduction

Two solutions to the problem of wireless mobility have emerged from the communications industry: one based on packet switching and one based on circuit switching. One is an extension of local area networks and the Internet, the other an extension of the digital phone network. In the wired world, much activity has been focused on the integration of the packet- and circuit-based approaches, such as asynchronous transfer mode (ATM) and voice over IP. In this article we explore the integration of the packet- and circuit-based approaches in wireless mobility.

The most frequently cited motivations for the integration of voice and data in the wired world were economics of operation and multimedia applications. The Internet has recently emerged as the natural habitat for multimedia applications. As the market for mobile voice saturates, operators are increasingly looking to mobile multimedia services to maintain their revenue stream [1]. The third-generation (3G) standardization effort is defining an "all-IP" multimedia subsystem as a separate extension to the existing network [2]. It is likely that a more congenial environment for mobile multimedia services might be achieved by the integration of IP technology within the core of the mobile network. We consider what form such an integrated mobile network might take and briefly review some of the technical difficulties.

In the following section we give a brief overview of the solutions to mobility developed by the data networking community. Then we turn to an overview of the mobility solution developed for voice telephony. We consider the direct application of IP technology to the radio access network (RAN) of the cellular mobile network and review the technical difficulties that render this naive approach impractical. The characteristics of an air interface designed specifically for high-speed data are reviewed and contrasted with the proposals from the 3G standards. Finally, we look at some more evolutionary upgrades whereby IP technology is already being deployed within the RAN.

Mobilizing the Internet

The problem of adding mobility to the Internet can be considered in three stages: untethering the terminal, handling terminals that move, and scaling to lots of terminals that move. As is traditional in these networks we will temporarily ignore the problems of security and billing.

Wireless LANs

A wireless LAN permits us to untether the terminal from the wired infrastructure but in itself offers no support for mobility beyond the domain of a single wireless subnet. The IEEE 802.11 family of wireless LAN standards offers a direct shared medium replacement for Ethernet. Thus, IP runs over the wireless LAN in much the same manner as in its natural habitat, wired Ethernet.

The basic concept of the wireless LAN is being extended to the public metropolitan area network (MAN). In IEEE 802.16a, fixed or nomadic service is contemplated, extending to a range of up to 30 mi and offering a level of service similar to digital subscriber line (DSL). Both 802.16e and 802.20 are working on support for vehicular mobility.

While the implementation details differ in each of these layer 2 wireless networks, the service offered to the IP layer remains the same. It is that of a shared medium packet network supporting high-peak-rate data transmission.

Mobile IP

Mobile IP was developed before the widespread adoption of wireless LANs. The original intent was to permit a mobile terminal to communicate using its permanent home IP address while connected to a foreign wired network. Terminals that do not require a permanent IP address can simply borrow a temporary local address using the Dynamic Host Configuration Protocol (DHCP). However, this does not allow them to move between subnets. Since the introduction of wireless LANs, Mobile IP has been extended to enhance the ability of the terminal to move between wireless subnets while maintaining its active connections.

The basic concept is to use a home agent to maintain a binding between the mobile terminal's home IP address and its current location. When a mobile enters a foreign subnet it obtains an IP address, called a care-of address, from that subnet's address space. The mobile registers the new care-of address with its home agent. Subsequently, all packets received for the mobile by the home agent are tunneled across the network using the care-of address.

Recent enhancements allow the terminal to reduce the time taken to detect that it has moved into a new subnet and to update the binding to a new care-of address in the home agent. Protocols are also under development to support context transfer and dormant mode alerting (paging). When handoff occurs between subnets, context transfer allows state such as authentication information, security context, quality of service (QoS) properties, and header compression to be transferred to the new subnet. Dormant mode permits the mobile terminal to conserve power while attached to the network but not actively engaged in communication.

Tunneling is attractive as a solution to the problem of mobility because it only requires changes to the network equipment in a small number of places, the tunnel endpoints. The rest of the network can remain unaware of the tunnels and needs no enhancement to support mobility. A tunnel-based solution has the disadvantage that intermediate routers

0163-6804/04/$20.00 © 2004 IEEE

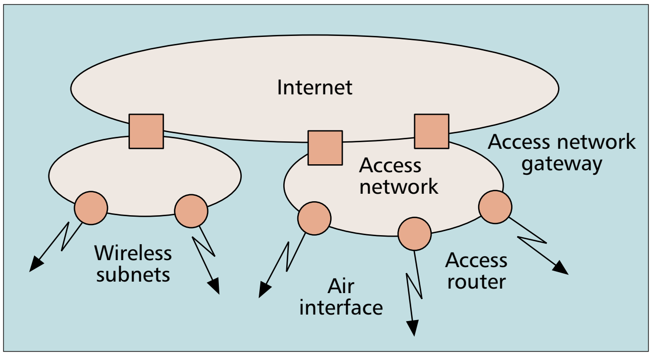

FIGURE 1. *Architecture of the mobile Internet.*

cannot see the packet headers. This makes functions like caching, multicast, and QoS differentiation difficult to implement. It can also affect the operation of existing protocols because it changes the end-to-end semantics of the Internet architecture [1]. For example, one current area of concern is the interaction of Mobile IPv4 tunnels with IPSec tunnels used to support encrypted access to a virtual private network.

Micromobility

It is generally agreed that in order to scale to a large number of mobile terminals, the mobility problem needs to be divided into two complementary parts: mobility over a large area (macromobility) and mobility over a small area (micromobility). Mobile IP is a macromobility solution. Much recent work has addressed the micromobility problem, although it is still considered an area of current research by the Internet Engineering Task Force (IETF) and not yet ready for standardization activity.

The resulting architecture for the mobile Internet is depicted in Fig. 1. Access routers connect directly to wireless LANs. Within the access network a micromobility protocol is used to maintain connectivity to mobile terminals. Mobile IP provides mobility beyond the confines of the access network and supports handoff between separate access networks.

In Mobile IP, when a terminal moves between subnets an update message must be sent to the home agent. If the home agent is located in a distant network, this incurs unacceptable delay. With a large number of mobile terminals it also generates an unacceptable signaling load across the network. Micromobility protocols confine the location information to the access network and thus reduce the handoff latency and reduce the signaling load.

Two approaches to micromobility have been pursued, one based on tunnels, the other on per-host forwarding [1, 3]. In tunneling schemes a proxy mobility agent is introduced into the access network that shields the home agent from all updates to the care-of address that are local to the access network. A local tunnel is maintained from the local mobility agent to the current subnet in which the mobile is located (or to the mobile itself). Per-host forwarding schemes tackle the problem of mobility directly in the routing protocol. Information about the location of each mobile is distributed within the access network so that routers can forward packets directly to mobile terminals without the use of tunnels. The trick is to minimize the amount of location information that needs to be distributed, or rather to minimize the number of routers to which it needs to be distributed while maintaining full connectivity.

The 3G All-IP Architecture

Turning to the mobility solution developed for voice telephony, Figs. 2 and 3 give a simplified view of the CDMA2000 and the Universal Mobile Telecommunications System (UMTS) 3G cellular mobile architectures respectively. (CDMA2000 is the 3G evolution of the cdmaOne network originating in the United States, and UMTS the evolution to 3G of the GSM network in Europe.) The terminology has been simplified to focus on data traffic and reflect the commonality with wired access networks such as DSL and dialup. Table 1 gives a mapping of this simplified terminology onto the 2G GSM system, and the 3G CDMA2000 and UMTS systems.

In both networks a data session is established to carry IP packets between the network access server and the mobile terminal [4]. Both networks use tunnels to support mobility. In CDMA2000 the tunnel is between the network access server and the base station controller, and permits the mobile to move between base stations and between base station controllers without interrupting the data session. Mobile IP is used to allow the mobile to move between areas controlled by different network access servers. In UMTS the tunnel from the network access server is routed through a tunnel switch. The mobile can move between base stations and base station controllers without changing tunnel switches by moving one leg of the tunnel. It can move to an access network controlled by a different tunnel switch by moving both legs of the tunnel without disturbing the data session.

Fundamentally, the architecture of the mobile Internet of Fig. 1 is all IP because the problem of mobility is solved in the IP layer. In contrast, the 3G mobile architecture solves the problem of mobility at layer 2, the 3G mobile network itself. If mobility is solved in the IP layer, routers have access to the packet headers and could implement IP-based QoS algorithms. IP/ATM interworking has already shown us the complexity that results from attempting to translate quality of service requirements from one network to another.

A native IP architecture supports all the functionality of IP throughout the network. Caching can be employed at any point within the network, but particularly within the access network where it can significantly reduce load and latency. Multicast can be used throughout the network to support multimedia services and conserve bandwidth. A tunneling solution for mobility hides the packet headers, and makes caching and IP multicast difficult to implement within the access network of CDMA2000 or anywhere within UMTS.

In a native IP architecture, routing to local resources is efficient. Packets do not have to travel to the core of the network and back again just to reach another mobile within the same cell, as required in UMTS. Integration with wireless LANs is simple because they use exactly the same architecture. New services are easily introduced just as they are on the Internet.

Of course, the above is contingent on the problem of micromobility being solved for IP. Until then we have to make do with Mobile IP tunnels that share some of the same difficulties as the tunnels of the 3G architectures. In the parlance of 3G, "all-IP" mostly refers to the IP multimedia subsystem. It includes the Session Initiation Protocol in the handset and the corresponding call control servers in the IP multimedia subsystem beyond the network access server. This is IP technology applied as an external addendum to the network. We now consider what might be achieved were IP technology integrated into the core of the network.

	GSM	UMTS	CDMA2000
Base station	BTS	Node-B	BTS
Base station controller	BSC	RNC	BSC
Tunnel switch	SGSN	SGSN	n/a
Network access server	GGSN	GGSN	PDSN
Circuit switch	MSC	MSC	MSC

Table 1. *Mobile network terminology.*

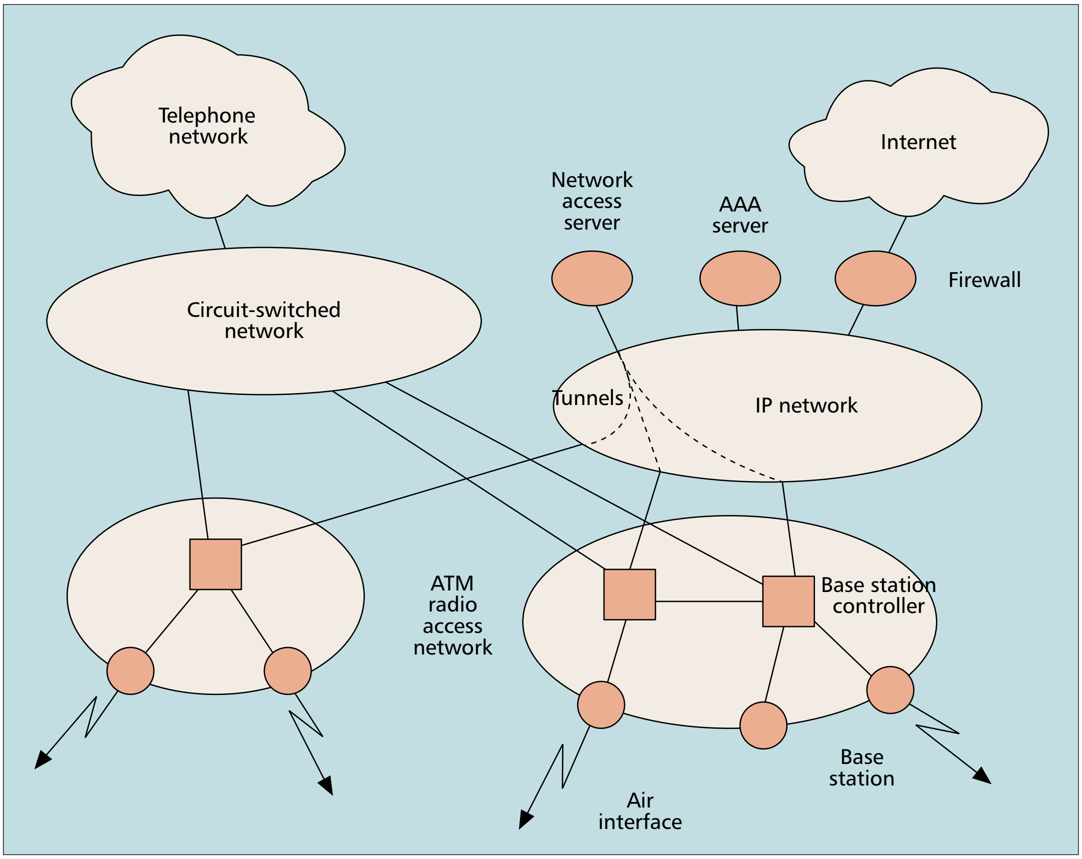

FIGURE 2. *CDMA2000 network architecture.*

IP Technology in the Cellular Mobile Network

IP multimedia applications are to be found in the wild on the Internet, and their natural habitat is packet switching in a distributed architecture with native IP functionality. Several times in the past they were encouraged to migrate to a circuit-switched architecture, first with the integrated services digital network (ISDN) and later with the virtual circuits of ATM. However, market forces prevailed. Is the lure of wide-area mobility sufficient to attract them to a mobile network optimized for circuit-switched voice? Or should we focus on creating a more congenial environment for mobile multimedia services by the integration of IP technology within the core of the mobile network?

Here, we take a look at introducing IP technology into the core of the mobile network, particularly into the RAN. We begin by looking at the introduction of the flat mobilized Internet architecture of Fig. 1 into the hierarchical 3G networks of Figs. 2 and 3.

A Native IP Radio Access Network

First we examine the direct approach, replacing the hierarchical 3G RAN with a distributed native IP network without changing the air interface. This approach does not work, but it is instructive to see why.

Air Interface — The fundamental problem is that the 3G network is designed and optimized around a single service — mobile, circuit-switched voice. The air interface selected for 3G is code-division multiple access (CDMA). CDMA is most efficient when a large number of low-bit-rate streams are spread across the radio channel. Circuit-switched voice offers a large number of long-duration low-bit-rate sources and is thus an ideal candidate for CDMA. Packet-switched data, however, is not. It prefers variable-duration high-bit-rate bursts, particularly in the downlink direction (base station to mobile). At high bit rates CDMA is limited by low spreading gain or intercode interference [5].

Power Control — CDMA is interference limited and is only efficient if tight closed-loop power control is used to minimize the amount of mutual interference between users of the system or between radio cells. Without such power control, the capacity of the system could be reduced almost to that of the first-generation frequency-division multiplexed systems. Closed-loop power control works well for transmissions of long duration with respect to the frequency of feedback information. Voice traffic exhibits this characteristic. For packet traffic, the validity of feedback information for closed-loop power control decreases as the interval between packets increases [6]. Data traffic is typically very bursty, which makes it more difficult to employ closed-loop power control.

Soft Handoff — Soft handoff is the simultaneous communication with two or more radio cells during transfer of a connection between cells. CDMA requires soft handoff for the same reason as closed-loop power control, to minimize interference and thus maximize system capacity. A mobile at the edge of one cell will transmit at high power. If it moves into another cell, it will cause unnecessary interference if it is not very quickly power controlled. While hard handoff could be used to control the mobile, the inevitable delay reduces the capacity of the system, and the high frequency of handoff required would impose too great a load on system control [7].

Soft handoff is intrinsically a centralized process. A central arbiter (the selection and distribution unit) is required to combine the signals arriving from multiple base stations. These signals must arrive at the arbiter within tightly controlled delay bounds (typically within 10 ms of each other) to participate in the combination process.

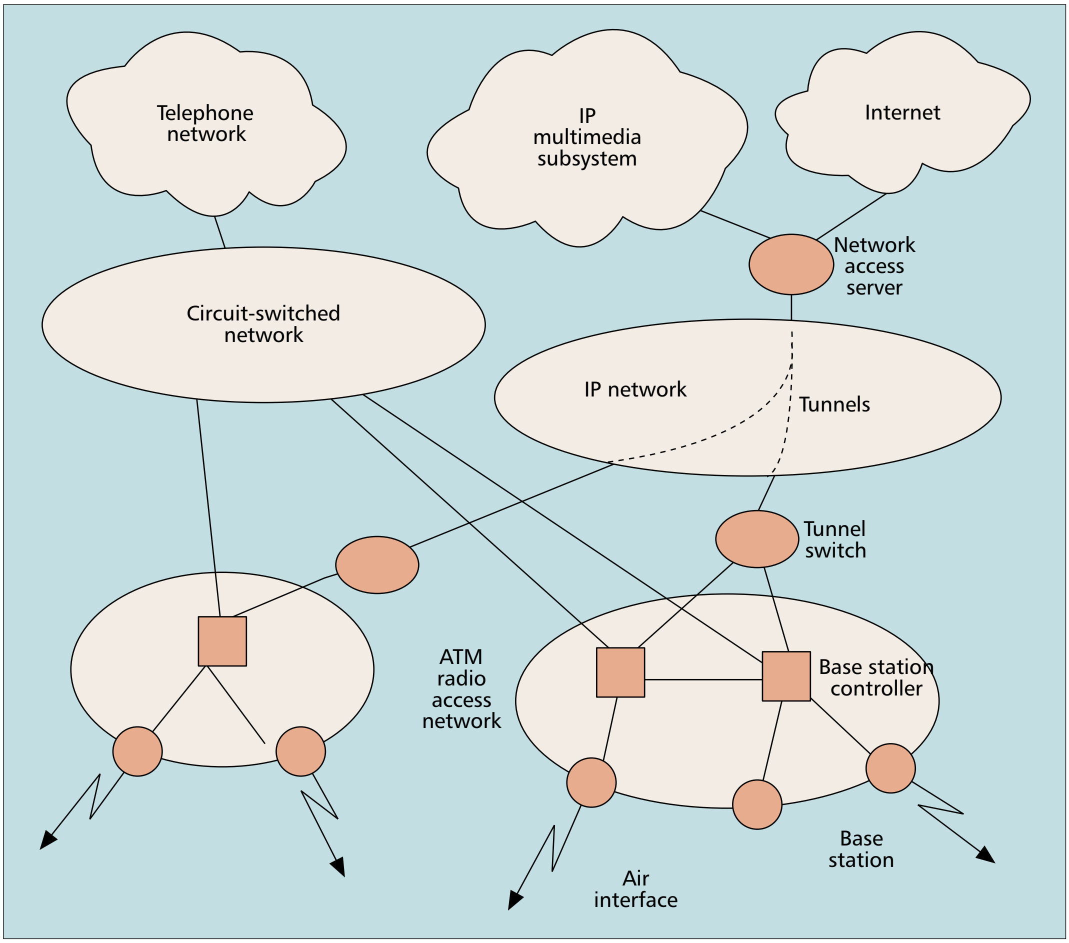

FIGURE 3. *UMTS network architecture.*

The arbiter is most naturally located at the center of a hub and spoke hierarchical architecture. That way, each of the separate signals travels up its backhaul link only once. It is possible to locate the arbiter in a base station and have that base station perform the combination function for its peers in a more distributed manner. However, if the physical topology of the backhaul is hub and spoke, the load on the backhaul links will be almost doubled because signals now have to travel up to the hub and then back out along one spoke before being combined.

High-Speed Packet Services

Given that CDMA is best suited to the characteristics of voice traffic, we now consider the design of an air interface optimized for the characteristics of data traffic.

OFDM — To achieve high bit rate channels, orthogonal frequency-division multiplexing (OFDM) is often selected. OFDM with sufficiently long symbol periods supports high bit rates in time delay spread environments with performance that improves with increasing delay spread up to a point of extreme dispersion [8]. The major high-capacity wireless data networks use a form of OFDM: IEEE 802.11a, 802.11g, 802.16a, Hiperlan II, [9].

CDMA is most efficient for a large number of low-bit-rate streams. In contrast, OFDM divides a single high-bit-rate data stream into a large number of low-bit-rate subchannels that are transmitted in parallel. These narrowband subchannels are sufficiently narrow to minimize the effects of multipath delay spread. OFDM is intrinsically better suited to the transmission of high-peak-rate packet data, although in wide-area cellular systems techniques are required to minimize the interference between adjacent cells [10]. An OFDMA proposal was one of the candidates for the selection of the 3G air interface [11]. It was rejected, in part, on the basis of the large peak-to-average power ratio of the OFDM signal [7]. This leads to high power consumption by the mobile.

HSDPA and 1xEV-DO — Both of the 3G standards organizations have recognized the need for enhanced capabilities for data traffic. Both have come up with similar solutions: high-speed downlink packet access (HSDPA) [7, 12] for UMTS and 1xEV-DO [13] for CDMA2000. Both solutions are essentially alternative air interfaces, optimized for data, in the downlink. They are well integrated with the rest of the system but at the air interface they can be considered the wireless equivalent of asymmetric DSL (ADSL): voice and data carried separately, each in its own piece of spectrum. In HSPDA the channel bandwidth allocation between voice and data can be configured and a single-carrier solution, sharing codes and transmit power, is possible. In 1xEV-DO an entire channel must be allocated to data, but CDMA2000 uses smaller channels. A similar competing standard for CDMA2000, 1xEV-DV, dynamically shares the channel between voice and data [14].

Both approaches have made similar design decisions for the simple reason that both are attempting to support a high-peak-rate shared data channel in the downlink within the context of a CDMA air interface. Conventionally, CDMA uses fast power control to mitigate the effects of rapid fading in the radio channel. In contrast, both high-speed packet proposals transmit at constant power and use adaptive modulation and coding to combat fading because power control is inefficient with bursty data. Both systems also use a scheduling algorithm that instantaneously allocates the shared channel to those mobiles

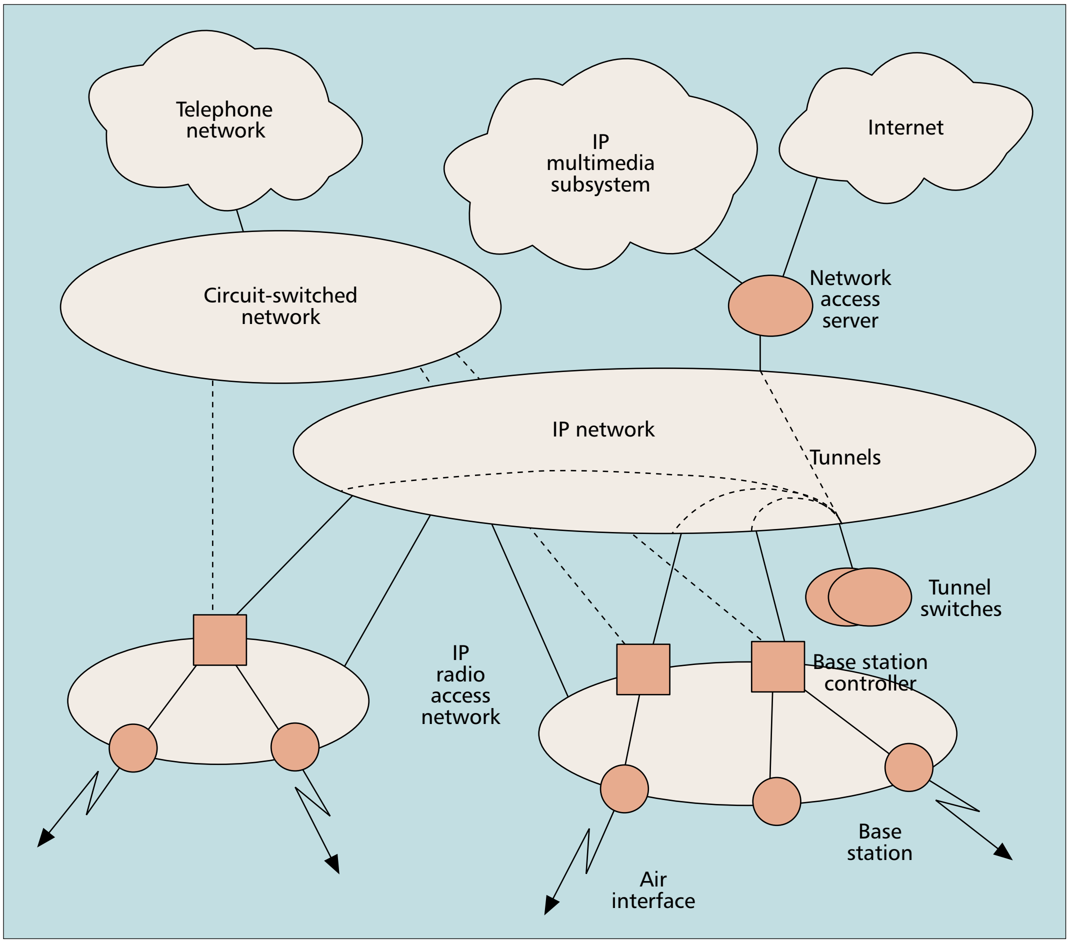

FIGURE 4. *UMTS with compressed voice over IP.*

currently experiencing the best channel characteristics. Use of such a scheduling algorithm prevents soft handoff from being used in the downlink because the scheduling decisions at adjacent base stations must be independent. In HSDPA, the related medium access control (MAC) layer has been moved to the base station (Node-B) to improve the delay characteristics [12]. For both systems, the uplink is not significantly changed, and continues to use power control and soft handoff.

If the air interface designed for high-speed shared packet transmission is fundamentally different from that designed for voice, why retain the hierarchical RAN architecture? Could we couple the high-speed packet air interface directly to a native IP RAN?

The uplink remains a stumbling block in both HSDPA and 1xEV-DO because it retains soft handoff. However, the time-division duplex (TDD) alternative [7] in the UMTS specifications does not use soft handoff. The high-speed packet channel HSDPA is also specified for TDD mode. Additionally, in TDD mode bandwidth can be asymmetrically allocated between uplink and downlink, which may lead to a more spectrum-efficient implementation.

IP Transport in the Radio Access Network

Turning now to consider less dramatic changes to the mobile network, IP technology is already being introduced into the RAN. In this context IP is employed as a transport alternative to the 16 kb/s circuits in GSM, and is proposed as an alternative to ATM adaptation layer 2 (AAL-2) in 3G networks. As a transport mechanism within the RAN, IP is only required to support communication between base stations and base station controllers, not to the mobile terminals themselves. So there is no problem of mobility to solve for IP in this application. As the base stations tend to be connected to their controllers via point-to-point links, improvements in performance are mostly due to statistical multiplexing and more efficient encapsulation.

Cost reduction in the most cost-sensitive part of the network is the main motivation for introducing IP transport. In terms of capital expenditure, the popularity of IP puts price pressure on the cost of IP equipment. In terms of operating expenses, IP offers more autoconfiguration features and is thus easier to manage. For example, it is much simpler to move a base station from one controller to another with IP as it automatically reconfigures itself with DHCP and reconnects itself with a routing protocol. Operation and maintenance is already based on IP using remote telemetry and the Simple Network Management Protocol (SNMP). Running it over IP saves operation and management costs. Also, IP can be used for remote surveillance with low-cost hardware.

IP is layer 2 independent. It will run over almost anything. If metro Ethernet or fiber to the curb becomes available as an alternative backhaul technology, IP will run over it. Also, with an IP network out to the radio tower it may be of interest to offer other IP-based services, such as IEEE 802.16a fixed "DSL-style" last mile Internet access in urban areas.

A simulation study by the Mobile Wireless Internet Forum investigated several possible IP encapsulations and reported a 10 percent improvement over ATM/AAL-2 [15]. Commercial implementations based on Point-to-Point Protocol multiplexing (PPPmux) and compressed Real-Time Transfer Protocol (RTP) claim 25–50 percent improvement over the 16 kb/s circuit-based backhaul of GSM and project 15–25 percent improvement against ATM/AAL-2. For GSM, further gains may be obtained from the inherent statistical multiplexing of packet switching.

Transcoder-Free Operation

Due to advances in compression technology, a number of different codecs are now in use in mobile handsets producing digitized speech mostly in the range 8–13 kb/s. To ensure interworking regardless of the codec in use at the far end, the practice until recently has been to convert all digitized voice streams to 64 kb/s using a transcoder. The transcoder is typically situated at the base station controller or the first switch in the 64 kb/s circuit-switched network. Transcoding is an expensive operation in terms of both the cost of the required equipment, and the delay and distortion introduced into the voice stream. There is considerable interest in avoiding this unnecessary expense wherever possible.

If a mobile is communicating with another mobile on the same network, or possibly even on a remote network, it is highly likely that they share a codec in common. Efforts are underway to define a signaling protocol that allows the mobile to negotiate with the network and remote mobile to determine whether they share a codec in common. If so, communication can proceed using compressed speech without the need for transcoding.

This signaling negotiation occurs with the switches in the circuit-switched network. Another development underway is that these circuit switches are being transformed into soft switches in which the control software is physically separated from the switching matrix. This enables us to consider using the IP network to transport compressed voice under the control of the soft switch controller.

By using the IP network for compressed speech we avoid the need for transcoders, reduce the delay for mobile-to-mobile calls, and possibly support direct connection to voice over IP networks. We can also support new services such as push-to-talk (walkie-talkie style multicast voice to a private group), generally implemented using voice over IP. Most push-to-talk applications occur within a small geographical range, so the best implementation would avoid shipping the voice streams across the network unnecessarily.

Two types of packet voice encapsulation are under discussion. One multiplexes voice samples from a number of conversations into a single IP packet to amortize IP overhead. This resembles an IP implementation of AAL-2. It restricts the entire multiplex to the same two endpoints. A more flexible encapsulation encodes speech from only a single voice stream into each IP packet and uses compression techniques to reduce overhead. It results in less overhead than the current ATM/AAL-2 approach and is directly compatible with voice over IP.

Figure 4 illustrates a UMTS network with support for compressed voice transport over IP. The residual media gateway function of the soft switch is integrated into the base station controller, which connects directly to the IP transport network. Signaling between the soft switch and the base station controller determines whether to use IP or the circuit-switched network for a call. Mobile-to-mobile calls use a common codec, and mobile-to-voice-over-IP calls use IP transport. Mobile calls to the public telephone network, and anything that requires advanced services like conference bridging, use the circuit-switched network. Call control is left as an exercise for the reader. Most of the work on the IP multimedia subsystem relates to call control for voice over an IP network.

Conclusion

Looking at Fig. 4 we see we are approaching the mobilized Internet architecture of Fig. 1. We are moving toward an all-IP architecture but in the core of the network, not merely external to the network. We are being driven there for reasons of capital and operational cost reduction. None of this evolution is being driven by multimedia applications. However, if they do appear, there is an IP network in place ready to handle them.

We have IP transport in the RAN, yet retain the hierarchical architecture of the 3G network. If demand for high-speed data services grows, we can deploy an alternative air interface optimized for data. If the IP community solves the problem of micromobility, we might adopt a native IP access network for data. If not, we continue to use the existing tunnel-based architecture of either 3G or Mobile IP. There are strong operational reasons for moving some voice to the IP network. With voice on the IP network we can connect directly to voice over IP networks and interface to those multimedia applications so eager to descend on us.

It should not come as a surprise. The architecture has evolved in the wired world. For all our efforts at integration over the years, we still handle voice and data separately. On a precious resource, such as the last-mile copper, voice and data share the medium, but separately as in ADSL. As soon as it is convenient we ship them off to their respective networks. The addition of mobility does not change the fundamentals. Market forces will lead to the same result in the mobile network as they did in the wired world.

Acknowledgments

This work was supported by the Mobile Wireless Group of Cisco Systems, Inc. The author would like to thank the numerous Cisco subject matter experts who shared their experience and insights about integrating IP and mobile networks.

References

[1] D. Wisely, P. Eardley, and L. Burness, *IP for 3G*, Wiley 2002.

[2] L. Bos and S. Leroy, "Toward an All-IP-Based UMTS System Architecture," *IEEE Network*, Jan. 2001, pp. 36–45.

[3] P. Reinbold and O. Bonaventure, "IP Micro-Mobility Protocols," *IEEE Commun. Surveys*, vol. 5, no. 1, Q3, 2003.

[4] P. J. McCann and T. Hiller, "An Internet Infrastructure for Cellular CDMA Networks Using Mobile IP," *IEEE Pers. Commun.*, Aug. 2000.

[5] G. J. Pottie, "System Design Choices in Personal Communications," *IEEE Pers. Commun.*, Oct. 1995, pp. 50–67.

[6] D. M. Novakovic and M. L. Dukic, "Evolution of the Power Control Techniques for DS-CDMA Toward 3G Wireless Communication Systems," *IEEE Commun. Surveys*, vol. 3, no. 4, Q4, 2000.

[7] H. Holma and A. Toskala, *WCDMA for UMTS*, 2nd ed., Wiley, 2002.

[8] J. Chuang and N. Sollenberger, "Beyond 3G: Wideband Wireless Data Access Based on OFDM and Dynamic Packet Assignment," *IEEE Commun. Mag.*, July 2000, pp. 78-87.

[9] T. S. Rappaport *et al.*, "Wireless Communications: Past Events and a Future Perspective," *IEEE Commun. Mag.*, May 2002, pp. 148–61.

[10] J. Chuang *et al.*, "High-Speed Wireless Data Access Based on Combining EDGE with Wideband OFDM," *IEEE Commun. Mag.*, Nov. 1999, pp. 92–98.

[11] T. Ojanperä and R. Prasad, "An Overview of Air Interface Multiple Access for IMT-2000/UMTS," *IEEE Commun. Mag.*, Sept. 1998, pp. 82–86.

[12] Kolding *et al.*, "High-speed Downlink Packet Access: WCDMA Evolution," *IEEE Vehic. Tech. Soc. News*, vol. 50, no. 1, 2003, pp. 4–10.

[13] P. Bender *et al.*, "CDMA/HDR: A Bandwidth Efficient High- Speed Wireless Data Service for Nomadic Users," *IEEE Commun. Mag.*, July 2000, pp. 70–77.

[14] A. C. K. Soong *et al.*, "Forward High-Speed Wireless Packet Data Service in IS-2000 — 1xEV-DV," *IEEE Commun. Mag.*, Aug. 2003.

[15] MWIF, "IP in the RAN as a Transport Option in 3rd Generation Mobile Systems," MWIF tech. rep. MTR006v2, May 2001.

Biography

PETER NEWMAN (peter.newman@ieee.org) is a co-founder of Netillion, a Silicon Valley startup. He has been a member of the founding team of three prior startups: Ensim, where he was chief scientist; Ipsilon, inventors of IP switching, a precursor to MPLS; and Adaptive, where he designed the first ATM switch to fail in the enterprise market. In a previous life he was a research fellow at the University of Cambridge where he received a Ph.D. for research in fast packet switch design before anyone called it ATM.