On Architectures for Broadband Wireless Systems

Shengming Jiang, research associate
Danny H.K. Tsang, assistant professor
Department of Electrical & Electronic Engineering
Hong Kong University of Science & Technology
Clear Water Bay, Kowloon, Hong Kong
email: {eejiang, eetsang}@ee.ust.hk

Sanjay Gupta, senior staff engineer
Motorola, GSM Products Division
1501 West Shure Drive (IL27-3205)
Arlington Heights, Illinois 60004
email: guptasn@cig.mot.com

Abstract

The growing popularity of portable and mobile computing and communication devices along with the introduction of wireline broadband networks is fueling demand for wireless broadband networks. Since ATM has been accepted as the standard for broadband integrated service networks we consider the problem of providing "ATM-like services" to mobile devices. In this paper, we discuss the fundamental issues that have to be tackled in order to provide broadband services that are currently available (or will very shortly be) to wireline hosts. We compare and contrast the architectures that have been proposed for extending the broadband wireline network infrastructure to the wireless environment. We argue that transporting small ATM cells over the air leads to inefficient utilization of the already scarce transmission capacity. Subsequently, we propose an architecture for integrated wireless and wireline broadband networks. The proposed architecture allows us to design the wireline and the wireless parts of the network independently, therefore allowing each to be optimized for the environment they operate in. Yet, the proposed architecture also permits a simple and efficient interconnection between the wireline and wireless infrastructures.

Keywords: ATM networks, wireless systems, mobility, multimedia

1 Introduction

In recent years there has been an increasing trend towards personal computers and workstations becoming "portable" and "mobile". This ever growing group of mobile users has been demanding access to network services similar to their "tethered" counterparts. The desire to provide universal connectivity for these portable and mobile computers and communication devices is fueling a growing interest in wireless packet networks. At the same time, wireline communication networks have been undergoing a revolutionary change themselves with the introduction of *Asynchronous Transfer Mode* (ATM) based *Broadband Integrated Services Digital Network* (B-ISDN). Given these rapid advancements, tomorrow's communication networks are expected to employ wireless media in the local area and utilize wireline physical media in the metropolitan and wide area environments. Wireless systems and networks will provide communication capability not only between mobile terminals but also permit mobile devices to have access to "wireline" networks. This will allow mobile computing and communication access to network resources that will enable them to run applications (browse the World Wide WEB (WWW) on the Internet, send and receive email, and broadband and multimedia applications) just like their wireline desktop counterparts.

To achieve the above goals, it is necessary to construct a wireless networking infrastructure that can support Quality of Service (QoS) guarantees essential to providing broadband services. Since ATM is the standard for wireline broadband networks, it is often assumed that broadband services are best provided to wireless users by extending ATM to wireless users. The above thought is reflected in most existing proposals for providing broadband services to mobile users that transport ATM cells over the "air". The approach realizes a seamless ATM-based communication system that provides end-to-end QoS guarantees and definitely achieves the objective of operating an ATM network over a wireless channel. However, we believe that the objective should be to provide ATM-like services to mobile hosts and an architecture that is best suited for the wireless environments should be employed for this purpose. Transporting ATM cells over wireless channels should definitely be considered as a possible approach but this approach should not preclude us from considering others that are potentially better suited to the wireless environment.

ATM based B-ISDN networks are best described as a public switched network infrastructure capable of supporting both narrowband and broadband services on a single flexible network platform and providing customer access over a single interface. The basic units of data transfer in ATM networks are 53-byte packets that are referred to as *cells*. ATM networks are designed to meet user specified QoS guarantees that are typically defined in terms of bounds on delay distributions and loss probabilities given the characteristics of the traffic generated by a source. ATM is made for high speed and very reliable communication channels. For satisfactory working of ATM protocols, it is typically required that the bit error rates be below 10^{-9}.

Wireless channels are nothing like the communication channels assumed by ATM. Compared to wireline channels, wireless channels are typically low bandwidth, unreliable, and non-stationary (i.e., their characteristics change with time). For instance, the bandwidth available in the ISM band for radio waves (or infrared optical channels) is typically less than a few tens of Mbps whereas the bandwidth of over 100 Mbps can be obtained in a routine manner over wireline channels. Additionally, while the bit error rate of fiber optic cables is typically less than 10^{-9}, the bit error rate for wireless channels, even under optimistic conditions, is seldom better than 10^{-6} and is frequently worse than 10^{-1}.

Since the characteristics of the wireless communication channels are significantly different from those of wireline channels and no procedures currently exist in ATM to support user mobility, solutions that are designed for wireline networks cannot be expected to be appropriate for wireless environments. A crucial part of our study is to demonstrate that ATM over wireless channels is a very poor choice indeed for providing broadband services to mobile users. We argue that transporting "small" ATM cells over wireless networks leads to inefficient utilization of the wireless channel capacity. On the other hand, even though we suggest that ATM should not be used over wireless channels, a key objective is to develop an architecture that is as close as possible to that of ATM networks to facilitate interconnection between wireless and wireline networking infrastructures. Therefore, we propose that the wireless part of the network should be isolated from the wireline ATM network. This enables the construction of the wireless and wireline network infrastructures that are best suited for the environment they operate in.

The paper is organized as follows. In Section 2 we discuss the organization of integrated wireless and wireline networks and summarize the salient features of systems designed for providing broadband services to mobile hosts. In Section 3 we give brief introductions to some existing proposals on wireless ATM. In Section 4 we analyze the "cellification" approach, and find that transporting ATM cells over the "air" leads to inefficient resource utilization. Then in Section 5, we propose a new architecture and discuss the implementation issues for some non-ATM air interface standards. Finally, we conclude the paper in Section 6.

2 Integrated Wireless and Wireline Networks

Future communication networks that integrate wireless and wireline networks will employ wireless media in the local environment in order to satisfy the demand for mobility and tetherless access to network services and will also employ wireline networks to interconnect wireless networks in order to provide a high speed backbone. An example of such an integrated network is shown in Figure 1.

As in typical wireless systems, in Figure 1, the area to be covered is divided into "cells" (different from ATM cells). In each cell is a *base station* that is connected with the wireline network and the mobiles as shown in Figure 1. At any given time, a mobile belongs to exactly one cell and is said to be associated with the base station in that cell. The base stations communicate with the mobiles on the one hand and with other base stations through the wireline networking infrastructure on the other. Therefore, the base station provides mobiles with access to wireline hosts and mobiles located in other cells. Throughout this paper we consider only the case where the wireline networking infrastructure is an ATM network. Before proceeding further, we outline the necessary details of ATM networks and pertinent factors relating to the wireless networking infrastructure.

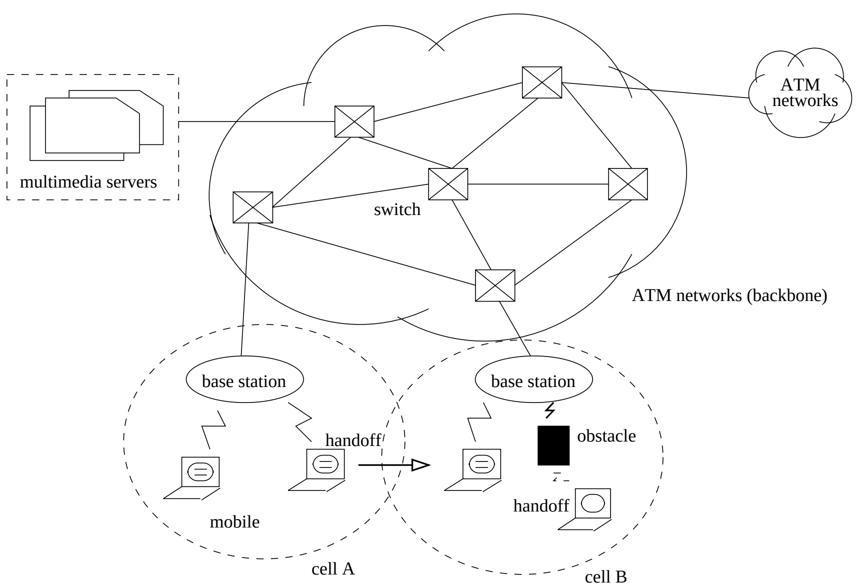

Figure 1: An example of broadband wireless systems

2.1 ATM Background

ATM is a high-speed, virtual circuit oriented, packet-switching technique that uses short, fixed-length packets called *cells*. A *Virtual Circuit* in ATM is a contract between the network and the customer to deliver traffic of specified statistical characteristics to the destination with a specified QOS, which is typically defined in terms of limits on cell loss, cell delay, and cell jitter. The architecture of ATM based B-ISDN consists of three layers: the physical layer, the ATM layer, and the ATM adaptation layer (AAL) as shown in Figure 2 [6].

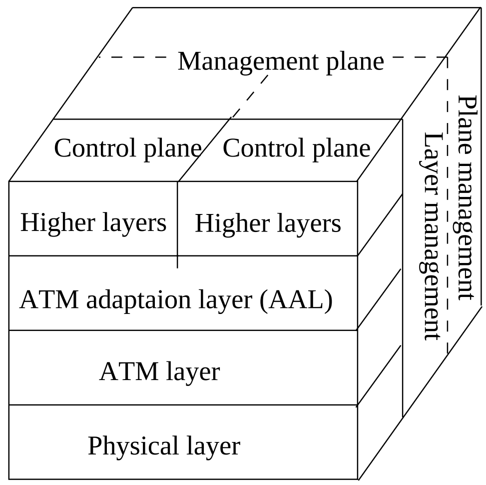

Figure 2: B-ISDN protocol reference model

The Physical Medium (PM) sublayer interfaces with the physical medium and at the transmitter, transmits bits on the physical medium. At the receiver, the PM delivers the recovered bit stream to the Transmission Convergence (TC) sublayer. At the transmitter the TC sublayer inserts ATM cells (received from the ATM layer) into the bit stream to be transmitted. Similarly, at the receiving end, the TC extracts ATM cells from the received bit stream and delivers them to the ATM layer. The ATM layer performs multiplexing, switching, and control functions based on the information contained in the ATM cell header. The fixed 53-byte cell is the basic data unit processed at the ATM layer and consists of 5-byte header and 48-byte payload for the user information. The ATM layer provides services necessary for cell transmission with “agreed” QoS guarantees.

The ATM Adaptation Layer (AAL [5]) provides an interface for higher layers to access the ATM services. The AAL layer is subdivided into the Convergence Sublayer (CS) and Segmentation and Reassembly Sublayer (SAR). The AAL handles Protocol Data Units (PDUs) that could be either variable or fixed in length (different from the length of an ATM cell). Currently, AAL layers for constant bit rate (CBR) (AAL1), variable bit rate (VBR)

(AAL3/4), and lightweight variable bit rate traffic (AAL5) have been standardized. The functions performed by the AAL layers depend on the application being supported. For example, the functions to maintain the timing relationship between the source and the destination exist at the AAL layer for AAL1 to support CBR delay-sensitive applications, which, however, need not be required at the AAL layer for AAL3/4 and AAL5. Additionally, the CS sublayer for AAL3/4 and AAL5 is further broken down into the Common Part (CP) and the Service Specific parts (SS), which are null for AAL1. Throughout this paper we will assume the presence of appropriate ATM signaling protocols and connection acceptance procedures.

2.2 The Wireless Infrastructure

As discussed previously, the wireless part of the network consists of "cells" in each of which a base station is present. The base station provides the functions that are needed to support the communication requirements of the mobiles that belong to its cell. Base stations, in cooperation with additional infrastructure, provide services required to support *mobility*, i.e., movement of a mobile from one cell to another. In addition, the base station provides functions needed to support authentication services, power management features and other miscellaneous functions. The above functions are implementation specific and since these are not critical to our discussion we do not discuss them here.

For wireless networks that are designed to guarantee QoS requirements, the base station initiates necessary connection acceptance procedures. Within a single cell, mobiles may communicate with each other with or without the intervention of the base station. However, information destined to/from other stations in the network must necessarily pass through the base station. For example, if each cell is an independent IEEE 802.11 compatible LAN, for intra-LAN communication, intervention of the base station (called an Access Point) is not necessary. However, the access point will be required during the connection establishment phase if adherence to QoS services is demanded. Since the base station typically understands the protocols used in the wireline and the wireless parts of the integrated network, it performs the necessary protocol translation functions.

The wireless channel capacity provided by wireless media is fundamental in broadband wireless systems because some multimedia applications such as video are bandwidth consuming. Therefore, the physical layer has to be designed to be able to provide sufficient bandwidth to satisfy multimedia applications. The wireless channels are often divided into forward channel and reverse channel. The forward channel is used by the base station to convey traffic to the mobile users from the base station, and the reverse channel is shared by several mobile users for transmission. The use of the forward channel is controlled by the base station; however, medium access protocol (MAC) is often needed for the reverse channel to coordinate channel access among users. The MAC protocol must be able to guarantee the bandwidth and QoS to multimedia applications while it is also efficient to achieve high utilization of the scarce radio resource.

2.3 Issues in Providing Broadband Services

In this subsection we discuss the challenges involved in providing broadband services to mobile users. Clearly, the most important bottlenecks are (i) the limited transmission capacity available and (ii) the poor quality of the wireless channels. Given the significant difference between the wireline and wireless communication channels, the most critical objective is to design the wireless and the wireline networking infrastructures that are specially made for their respective environments and that facilitate simple interconnection techniques to be employed such that the interconnection devices can be very simple and would not become the bottleneck.

To circumvent the problem of scarce wireless transmission capacity, increased spatial multiplexing can be performed. Spatial multiplexing refers to the use of the same part of the electro-magnetic spectrum in two different cells that are sufficiently separated from each other. Increased spatial multiplexing leads to increased capacity being available per user while minimizing the power that must be transmitted by the mobiles, thus helping them conserve valuable battery power. The current trend in wireless networking is to use small size cells referred to as micro and pico cells. (The diameter of a micro cell is typically from 100 to 500 meters and that of a pico cell is usually no more than 100 meters.) However, the use of pico cells results in increased hand-offs[1] as the frequency of the hand-off is inversely proportional to cell sizes. Therefore, if one can increase the utilization of radio resources, larger-size cells can be used to decrease the hand-off rate. With this in mind, we focus not only on providing ATM-like services to the mobile user, but also maintaining a high level of radio resource utilization by designing an appropriate interface between the "air" and ATM networks.

Processing of hand-offs requires substantial management and control effort from the network. Clearly, the QoS requirements of a connection should be maintained (if possible) after a hand-off. During the time the hand-off is accomplished, it is inevitable that the performance experienced by the mobile is degraded. Therefore, from the mobile's perspective, a

[1] Hand-off refers to the phenomenon of the mobile associating with a new base station, which is necessitated by the movement of the mobile from one cell to another.

key objective is to complete the hand-off procedures in the smallest time possible. At the same time, inefficient use of network resources has to be avoided. For this purpose, efficient routing of connections has to be maintained even after a hand-off.

Broadband services require that end-to-end QoS be maintained. QoS is typically expressed in terms of bounds on maximum delay, delay variation, and packet loss. Therefore, in order to provide broadband services to mobile users, it is essential that support for QoS be provided by the wireless part of the integrated network. Wireless communication channels are broadcast in nature and, therefore, medium access control (MAC) protocols that can provide QoS guarantees, similar to those provided by the ATM network, need to be developed. For example, in Ethernet LANs, the adopted MAC protocol, Carrier Sense Multiple Access with Collision Detection (CSMA/CD), cannot guarantee bounded delay, which is essential for real-time applications. Recently, MAC protocols that provide support for QoS have been developed for wireless environments. Examples of such protocols include that specified by the IEEE 802.11 working group [15].

Given that wireless channels are unreliable and non-stationary, use of appropriate error control techniques is a must. These error control techniques can impact the QoS experienced by end-user applications. The error control schemes adopted for the air medium need to be efficient to satisfy the requirements of different applications and also to adapt changing wireless environments. Since different applications have widely different QoS, appropriate error control techniques must be developed.

Finally, mobiles are tetherless devices and operate on battery power. It is therefore imperative that the computational and processing burden placed on these devices be minimal so that power can be conserved. In addition, since base stations interconnect high-speed wireline networks to low speed wireless networks and they are likely to be the spot for congestion. Hence, the design of the base station, and how it accomplishes the desired functionality are critical in achieving the design objectives.

3 Overview of Previous Work

In this section, we review some of the existing proposals for providing ATM-like services to the wireless users (see Table 1). A majority of the proposals suggest that the mobiles behave just like wireline ATM hosts, and therefore it is the mobiles that are the termination points for an ATM connection. Examples of systems that follow the above approach include those presented in [2],[4], [7] and [9]. In this scenario, the base station performs the functions of an ATM switch. Table 1 gives a brief summary of the above proposals. Figure 3 shows the

Table 1: Summary of existing proposals for wireless ATM

Models	Cell Size	Air cell	Error control	MAC	Reference
WATMnet[1]	micro	$2+2+\frac{48}{n}+2$	ARQ	TDMA	[2][3][8]
RATM[1]	pico	?+?+48+2	ARQ	Slotted ALOHA	[4]
SWAN[1]	pico	6+5+48+2	FEC, ARQ	Dedicated channels	[7]
BAHAMA[1]	micro	?+5+48+?	adaptive	DQRUMA	[9][10]
PSTN-ATM[2]	not specified by the paper				[1]

[1] ATM connections end at mobile

[2] ATM connections end at base station

Air cell = header+ATM header+payload+trailer

'?' indicates the value is not clarified by the paper

layered protocol structure of various network entities in such a network. In the following, we examine how the various functionalities are accomplished in each of the above proposals.

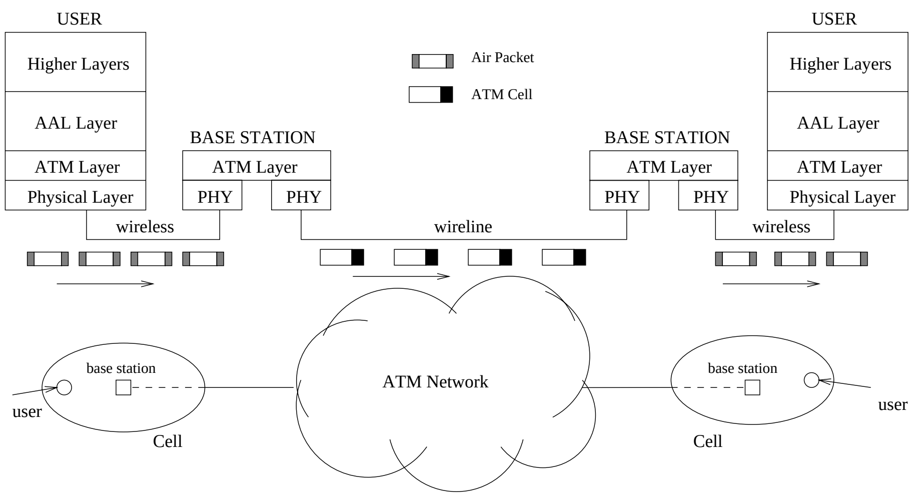

Figure 3: Architecture of existing proposals for broadband wireless networks

Air Packet Format

In each of the above systems the format of an ATM cell used is slightly different from that used in the wireline network. Almost all approaches employ *header compression* to reduce the header size and thereby increase the information transmission rate. For example, in WATMnet, the 5-byte ATM cell header is compressed to 2 bytes with 12 bits for VCI and 4 bits for ATM controls.

Error Control

Since ATM assumes the use of reliable physical media, no error control is performed by the ATM layer. However, given the poor quality of the wireless channels, some form of error control has to be provided. The proposals discussed above differ widely in how error control is accomplished. For instance, in WATMnet and RATM, each ATM cell is subdivided into smaller units and any units received in error are recovered using an automatic repeat request protocol. SWAN employs 24 parity bits for forward error correction to each ATM cell and employs selective repeat to recover cells that are received in error. BAHAMA employs a multi-layer adaptive error control scheme based on FEC and ARQ, that is, the bit-level FEC against random error at the physical layer, byte-level and packet-level FEC at the data link layer. The byte-level FEC tries to correct the packet in error with some FEC codes embedded in the packets. If the embedded codes do not suffice to recover the original packets, the packet-level FEC with an Automatic Repeat Request (ARQ) mechanism requires the source to transmit some additional packets for reconstructing the erroneous packets.

Mobility Management

In WATMnet the end-to-end VC connection is split into static and dynamic segments for mobility support. The static and the dynamic parts meet at special switches referred to as Hand-Off Switches (HOS). When hand-off happens, only the dynamic segments need to be reset by HOS, where a group of new dynamic segments are pre-established to connect the original static segments. A similar approach is used in RATM where the HOS is referred to as the Mobile Switching Point (MSP). In SWAN, once a mobile moves from one cell to another, the path of each active connection (that terminates at the mobile under consideration) is extended by the addition of a hop from the previous base station to the new base station. The "Homing Algorithm" used in BAHAMA is a similar procedure for supporting user mobility.

Medium Access Control

WATMnet employs separate uplink and downlinks. On the uplink, TDMA is employed and broadcast is used on the downlink. RATM suggests that slotted ALOHA be used as the MAC protocol. Since contention-based MAC protocols can not provide upper bounds on delay, RATM fails to provide strict QoS guarantees. SWAN currently uses dedicated channels for each mobile and therefore no procedures are needed to guarantee QoS on the wireless portion of the network. BAHAMA employs Distributed Queuing Request Update Multiple Access (DQRUMA) protocol in [11] as the MAC protocol in each cell.

Another proposal for interconnection between the base station and Public Switched Telephone Networks (referred to as PSTN-ATM) was introduced in [1]. Different from the above proposals, the ATM connection ends at the base station rather than at the mobile in this proposal. Thus, the mobiles need not be aware of the existence of ATM so that the air packet may be periodic with either fixed or variable sizes, or it even may be aperiodic.

4 How Large Should "Air" Packets Be?

The high bit error rates that are typical of wireless channels have often been used to justify the use of small packet sizes in packet switched wireless networks. The above studies argue that as the size of a packet increases, for a given bit error probability, the probability of it being received in error increases as well. The justification put forth for using small packet sizes is sound under the assumption that no *forward error correction* is employed on the wireless channel. FEC is a technique that increases the apparent reliability of the channel by using error correcting codes to combat bit errors (due to channel imperfections) by adding redundancy, henceforth *parity bits*, to information packets before they are transmitted. This redundancy is used by the receiver to *detect* and *correct* errors.

Before we proceed with a discussion of the relevance of FEC, it is necessary to outline preliminary details of the FEC procedure. Assume that the packets generated by the source are k symbol[2] long and these packets are transformed to n symbol *codewords* by an (n, k) *linear block* code with *minimum distance* d_{min} and *redundancy* $r = n - k$ for error detection and correction. A coder/decoder pair correct with minimum distance d can correct up to $t^* \leq \lfloor (d_{\text{min}} - 1)/2 \rfloor$ symbol errors and detect all errors between t^* and $d_{\text{min}} - 1$. If the number of symbol errors exceed t, the decoder will either (i) declare "failure" or (ii) commit an "error", i.e., decode the packet incorrectly [20] [19]. Examples of error correcting block codes include binary *Bose, Chaudhri, and Hocquenghem* (BCH) codes, (henceforth, we will use BCH to refer to *binary* BCH codes), and *Reed Solomon* (RS) codes (the most popular non-binary BCH codes).

Consider a case where an (n, k) block code is being employed to correct up to $t \leq t^*$ symbol errors. Assume that the symbol error probability is p and the events corresponding to successive symbols being received in error are independent. Define $b(t, n, p) = \binom{n}{t} p^t (1-p)^{n-t}$ and let $S_{n,p}$ be a binomial random variable with parameters n and p, i.e., $P\{S_{n,p} = t\} = b(t, n, p)$. Denote by $P_E(p, n, t)$ the probability that any given packet is not decoded (due to decoder failure) or decoded incorrectly. Since either decoder failure or decoder error

[2] Each symbol can take on one of m values and therefore each symbol contains $\log_2 m$ information bits.

result in the number of symbol errors in a received codeword exceed t, we have[3]

$$\begin{aligned} P_E(n,k,t) &= 1 - \sum_{i=0}^{t} \binom{n}{i} p^i (1-p)^{ni}, \\ &= P\{S_{n,p} > t\}. \end{aligned} \tag{1}$$

Note that as t increases, the error probability for a packet decreases. Although the use of Forward Error Correction (FEC) decreases the probability of a packet being received in error, the increased number of parity symbols decrease the potential throughput that can be achieved. Clearly, the relevance of a FEC scheme then depends on the ratio of the length of the packet, k, to the length of the codeword, n, for a given error correction capability. It can be argued that the error correction capability, t, of a (n,k) block code, where $n = 2^m - 1$ exceeds $\frac{n-k}{c}$, where $c > 1$ is a constant that depends on the coding-decoding scheme being employed [18]. For example, for binary BCH codes of length $n = 2^m - 1$, $c = m$ and for RS codes of all lengths $c = 2$.

Let $\overline{X}$ denote the time required to transmit a codeword of length n. The maximum throughput (henceforth, referred to as just throughput) measured as the rate of successfully decoded packets times the number of symbols in each packet per unit time

$$= \frac{k(1 - P_E(p,n,t))}{\overline{X}}. \tag{2}$$

Clearly, the maximum throughput (normalized by channel transmission rate in symbols per unit time) that can be obtained is

$$\Lambda^*(p,n,t) = \frac{k}{n}(1 - P_E). \tag{3}$$

The above equation (3) can be solved to obtain the optimal error correction capability t^* that should be employed given codeword length and channel symbol error probability. In general it is difficult to obtain a closed-form expression for either the optimal error correction capability, t^*, or the maximum throughput, $\Lambda(n,k,t^*)$. Either can however be calculated using numerical methods. In any case, it is interesting to consider the following asymptotic results[4] (as np gets large) [18]

$$t^* \sim np + O(\sqrt{np}) \text{ and} \tag{4}$$

$$\Lambda(p,n,t^*) \sim (1 - cp) + o(1). \tag{5}$$

[3]In general, it is difficult to estimate the exact decoder error and failure probabilities. However, upper bounds are available in the literature for a number of special cases of interest [19]. A useful approximation (widely conjectured to be an upper bound) for the undetected error probability of BCH codes is $2^{-(nk)}$.

[4]It should be emphasized that the above results are valid for channels that are bursty as well in which the error probability of successive symbols are not independent.

Even though we derived the above result under the assumption that the probability of successive symbols being in error are independent, the above results can be shown to hold true for other channels as well (for example, channels where errors happen in bursts).

The result of (4) can be intuitively explained as follows. The parity bits in a codeword are wasted either (i) if the number of symbol errors are greater than the error correction capability of the code being used or (ii) if the symbol errors are much fewer than the error correction capability (since in this case fewer parity symbols would have sufficed). As the codeword size increases, the coefficient of variance[5] of the random number of symbol errors decreases. Therefore, as n increases, fewer parity symbols are wasted. Further, since the expected number of symbol errors in a codeword of size n are np, we should therefore provide for a FEC scheme that can intuitively correct np symbol errors. The asymptotic results of (4) and (5) neglected the overhead due to the fixed-size header that has to be appended to each packet. Since the header can be assumed to be of a fixed size, the overhead per bit decreases as the packet size decreases, further emphasizing the need to employ large packet size in order to increase efficiency.

FEC alone can not guarantee error free delivery of packets and typically ARQ schemes have to be employed to recover packets received in error if necessary. Since ARQ schemes rely on retransmission of packets received in error, their use results in increased delays. However, if the round trip delay is small (as is the case within a single cell), the effect is minimal if the number of retransmission attempts is kept reasonably small. It is demonstrated in [17] and [18] that (i) the choice of the ARQ protocol has little effect on the optimal error correction capability and (ii) the asymptotic results presented above hold independent of the ARQ protocol being used. Therefore, even when an ARQ protocol is being employed, the largest possible codeword should be employed.

The above discussion assumed that maximizing throughput is the primary objective. However, frequently throughput has to be optimized subject to constraints on maximum delay. The most common constraint is one that ensures that no more than ϵ fraction of the packets suffers delay exceeding d.[6] The above constraint places an upper bound on the number of retransmissions that are allowed and therefore an upper bound on the loss probability that can be tolerated for each attempt. Once again, larger codeword sizes are recommended in this case since they offer the advantage of requiring the smallest overhead (that includes both the parity bits due to FEC and packet header) given a target codeword error probability.

[5]Coefficient of variance of a random variable is defined as the ratio of the variance to the square of the mean.

[6]Note that $\epsilon = 0$ implies unbounded d since no error recovery protocol can guarantee finite delays.

From the above arguments, it can be concluded that in wireless networks, the largest possible codeword size for an application should be used. The exact size of the codeword to be used would depend on the application under consideration. For instance, if the application is packet voice, the packetization delay increases with an increase in codeword size. Since for packet voice maximum end-to-end delay has to be bounded, the maximum size of the codeword is limited. For example, for packet voice with a peak rate of 32 Kbps, and maximum packetization delay of 10 msec the maximum packet size (with the FEC bits) possible is 40 bytes. However, if the application is video with a peak rate of 1 Mbps, the maximum packet size is 1250 bytes for a maximum allowable packetization delay of 10 *msec*. Choosing a large codeword size increases the coder-decoder complexity and issues related to their implementation are likely to limit the maximum size of the codeword used.

5 The Proposed Architecture

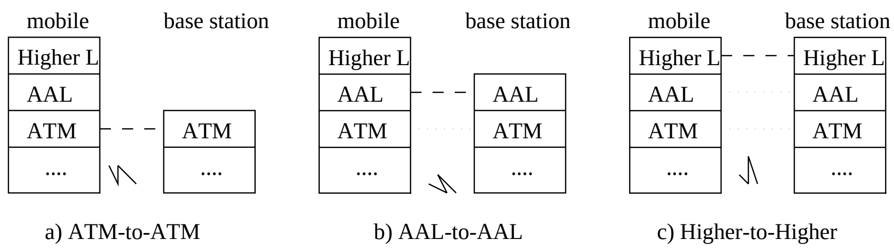

Figure 4: Peer-to-peer connections between mobile and base station

Before we proceed with a discussion of the proposed architecture, it is instructive to consider the choices available for accomplishing the required functions. With the peer-to-peer connection approach, which requires the same protocol to be operated by the corresponding peer layers in connection at the sender and receiver, there are three types of connections between the mobile and the base station according to the connected-layer as shown in Figure 4:

1. *ATM-to-ATM*: The ATM-to-ATM layer connection is equivalent to the approach of transporting ATM cells over the air medium, which has been analyzed in Section 4.

2. *AAL-to-AAL*: In this approach users access the ATM services through the AAL interface for multimedia applications and the air packet can be longer than the ATM cell to achieve high resource utilization. However, with this mode, a long CS-PDU needs to be segmented at the mobile for transmission over the air medium. At the base station, these segments have to be reassembled into the CS-PDU which needs to be segmented again into SAR-PDUs at the SAR sublayer as shown in Figure 5.a.

3. *Higher-to-Higher layers*: The Higher-to-Higher layer connection (e.g., TCP) makes the ATM layer completely transparent to the users so that the ATM capabilities such as QoS guarantees cannot be exploited by higher-layer applications.

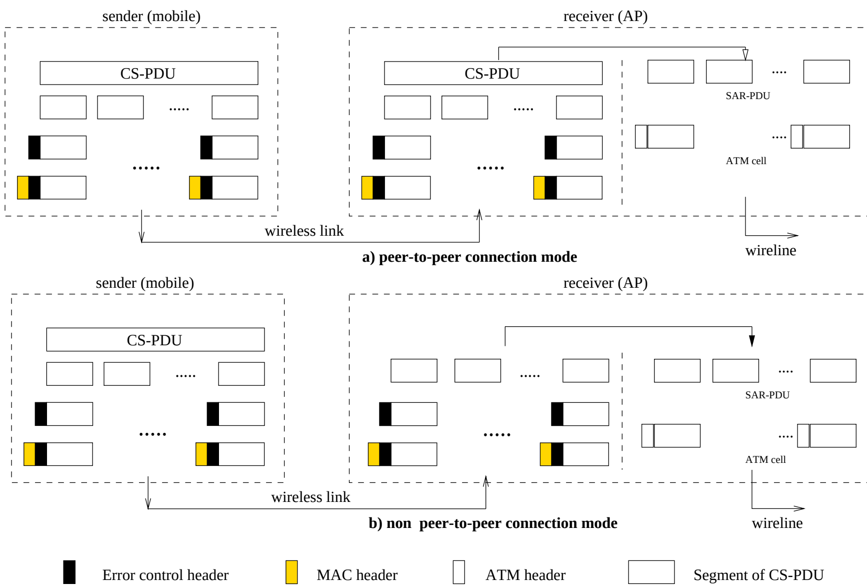

Figure 5: Data conversion for AAL connection between mobile and base station

Given that all peer-to-peer approaches suffer from severe problems and are unable to utilize the wireless network better, we propose that a non peer-to-peer approach shown in Figure 5.b be adopted. The proposed architecture for interconnecting wireless LANs and ATM distributes the SAR sublayer of the AAL layer between the mobile and the access point (AP) such as the base station[7] as shown in Figure 6. The portions of the SAR sublayer at the mobile and the AP are referred to as the SAR+ and SAR-, respectively. Further, an intermediate sublayer, MAC$^+$ sublayer, is inserted between the SAR+ and the SAR- sublayers. Observe that the integration of the protocol stacks at the mobile unit and the AP (linked through the wireless channel) forms the complete protocol stack of a conventional ATM terminal. The MAC$^+$ sublayer, as shown in Figure 6, consists of a MAC sublayer to arbitrate access to the wireless channel and an application specific error control sublayer. The error control sublayer may be omitted in cases where the MAC$^+$ sublayer provides the desired level of reliability. The MAC$^+$ sublayer is a key component of the proposed architecture and is designed such that the splitting of the SAR is transparent to the mobile and the base station. This requires the MAC$^+$ sublayer to be able to provide negotiable and bounded QoS to satisfy different applications.

[7]Henceforth, we use Access Point (AP) instead of base station

The SAR+ sublayer (at the mobile) accepts CS-PDUs from the CS sublayer and fragments them into MAC-PDUs (the MAC-PDUs in IEEE 802.11 have a maximum length of 2312 bytes). Further, it also reassembles the received MAC-PDUs into CS-PDUs and passes them to the CS sublayer. Similarly, the SAR- sublayer (at the AP) fragments the received MAC-PDUs into ATM cells and forwards them to the wireline network. Further, the SAR- sublayer reassembles ATM cells it receives (to be forwarded to a mobile host) into MAC-PDUs and transmits them on the wireless channel. In order to simplify the fragmentation and the reassembly of long MAC-PDUs (i.e., $\geq$ the ATM cell payload) into ATM cells, only MAC-PDUs with payloads necessary to fill up an integer number of ATM cells are allowed.

The error control sublayer provides application specific error control procedures. The MAC sublayer arbitrates access to a (broadcast) wireless channel. These sublayers are discussed in detail in the following sections.

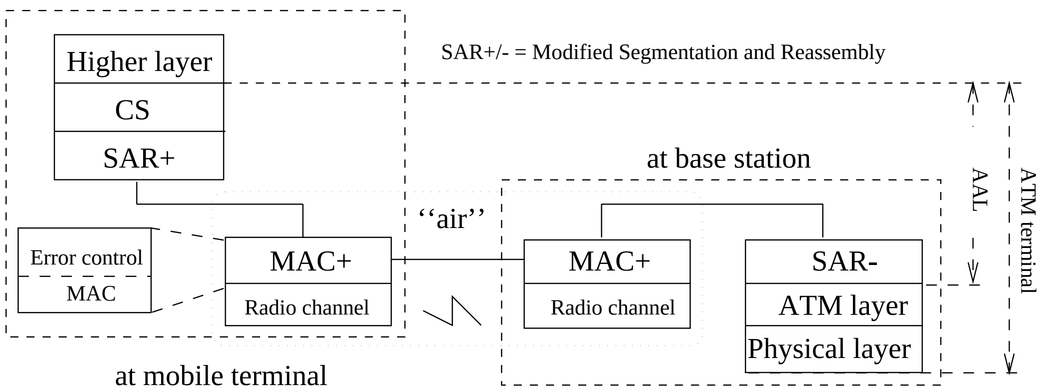

Figure 6: Protocol stack for the proposed architecture

A key feature of the proposed architecture is that the mobile is isolated from the ATM network. This removes the constraints imposed by "cellification" on the air packet. Note that in the proposed architecture the functions of the SAR sublayer are split between the SAR+ and the SAR- sublayers. This permits MAC-PDUs with both small sized payloads (e.g., $48/n$ bytes, $n = 1, 2, ...$) and large sized payloads (which could be as large as 2312 bytes in 802.11, for example) to be exchanged over the wireless medium; but only MAC-PDUs smaller than or equal to the ATM payload are allowable in the cellification architecture. Therefore, the proposed architecture can flexibly determine the MAC-PDU size according to the application requirement and the adopted error control scheme.

Both "cellification"-based and the proposed architectures can provide ATM-like services to the mobile through the AAL interface. However, no ATM layer is needed at the mobile with the proposed architecture which allows a simpler mobile structure than that based on the cellification architecture. This feature of the proposed architecture is helpful to reduce the computational and processing burden operated by the mobile to save the battery power. With the proposed architecture, the functions for hand-off support can be provided by AP and no changes to existing ATM protocols are required because the ATM connection ends at the base station. This provides flexibility in designing hand-off support schemes and permits the proposed architecture to co-exist with the existing ATM products, which are prototyped according to the current ATM standards. Further, each mobile in the proposed architecture does not have to be assigned an ATM address since the mobile is no longer an ATM terminal. The ATM address of an AP can be shared by the mobiles supported by it, saving valuable ATM network address space.

We now present how the proposed architecture can be implemented on existing standards for the air interface and MAC layer protocols such as the IEEE 802.11 (a draft for WLAN MAC and physical layer standard) and IS-95 (a U.S. digital cellular standard).

5.1 IEEE 802.11

IEEE 802.11 is a standard for wireless LAN being developed by the IEEE 802.11 working group. The basic access method in the IEEE 802.11 MAC protocol is the Distributed Coordination Function (DCF), which is best described as the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). This method can be used to support asynchronous data transmission and in an ATM environment suitable for use by adaptation layers AAL3/4 and AAL5.

In addition to the DCF, the IEEE 802.11 also incorporates an alternative access method called Point Coordination Function (PCF) - an access method that is similar to "polling" and uses a point coordinator to determine which station has the right to transmit. This service is designed to support bounded delays and delay variance, and is therefore suitable for use by AAL1. (Readers should refer to [21] for details of how the PCF can be used to guarantee QoS.)

The 802.11 MAC protocol can be used as the MAC layer protocol in the proposed architecture over a variety of 802.11-specific physical layers as shown in Figure 7. The physical layers can be 2.4 GHz Frequency Hopping Spread Spectrum (FHSS), 2.4 GHz Direct Sequence Spread Spectrum (DSSS) and $850 - 950$ nanometer Infrared (IR). The maximum channel data rate provided by 802.11 is 2.0 Mbits/s. In this implementation, the AP functions as the Point Coordinator for the PCF service, and the error control at the top of the MAC layer proposed by the architecture can be omitted. This is because the 802.11 MAC layer uses immediate positive acknowledgment where retransmission is scheduled by the

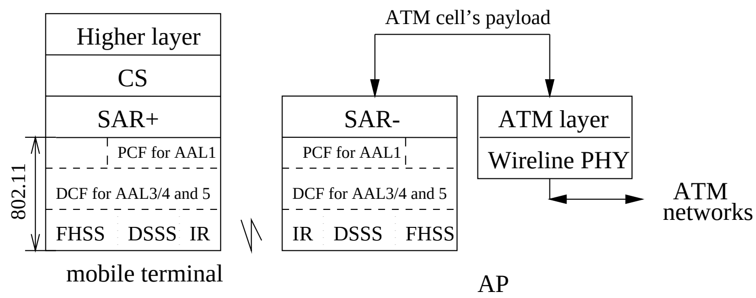

Figure 7: Implementation of 802.11 in the proposed architecture

sender if no acknowledgment is received for the DCF service, and the PCF service operates on the base of DCF.

As shown in Figure 8, when the mobile is sending data, the data enter the CS sublayer[8] via the AAL service access point (AAL-SAP) and are then encapsulated into CS-PDUs[9]. At the SAR^+ sublayer, CS-PDUs are segmented into SAR^+ PDUs by adding SAR^+ headers. The SAR^+ header is used by the peer entity at the AP to learn the sequence of the SAR^+-PDUs and the end of a CS-PDU. Note that the size of the SAR^+-PDU cannot be longer than the maximum payload of the 802.11 MAC PDU, i.e., 2312 bytes. The SAR^+-PDUs go down to the 802.11 MAC layer and are encapsulated into 802.11 MAC PDUs, which are then transmitted to the peer entity at the the AP by the MAC layer. Which service, DCF or PCF, is used for transmission depends on the QoS requirement of applications.

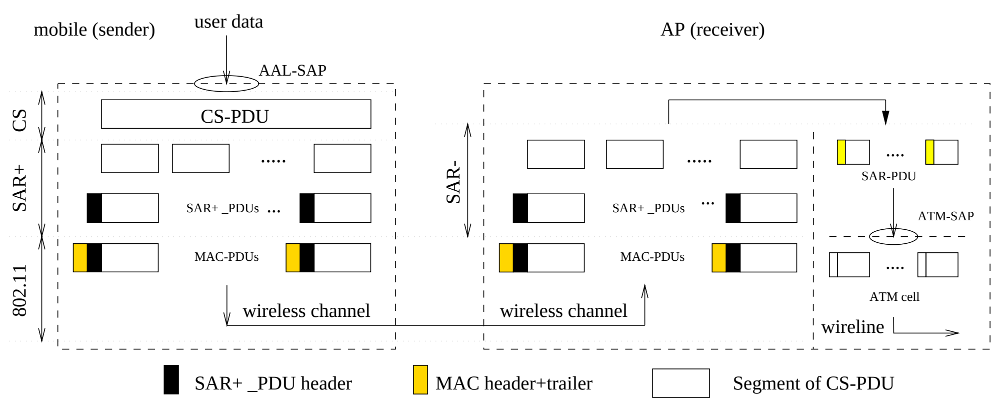

Figure 8: Data conversion between mobile and AP

At the AP (i.e., receiver) as shown in Figure 8, the MAC layer retrieves the MAC payloads from the correctly received MAC-PDUs and passes them (i.e., SAR^+-PDUs) up to the SAR^-

[8]The CS sublayer for AAL3/4 and AAL5 is further divided into Common Part CS and Service Specific CS which, however, are ignored here for simplicity.

[9]The maximum length of CS-PDUs for AAL3/4 and AAL5 is 65,535 bytes [5].

sublayer. Note that the SAR$^+$-PDUs are not reassembled into CS-PDUs at the SAR$^-$ sublayer to avoid the overhead introduced by segmentation-reassembly-segmentation upon the same CS-PDU. Instead, the SAR$^-$ sublayer directly forms SAR-PDUs with the payload of the received SAR$^+$-PDU if the payload is not smaller than 48 bytes for AAL5, for example, or if the received SAR$^+$-PDU indicates the end of the CS-PDU; otherwise, the SAR$^-$ sublayer needs to collect enough SAR$^+$-PDUs to fill a SAR-PDU payload. The SAR-PDUs go down to the ATM layer via the ATM SAP and then are put into ATM cells, which are then transmitted through wireline ATM networks to the destination.

Similarly, when AP is the sender, the SAR$^-$ sublayer forms SAR$^+$-PDUs with the payloads of the ATM cells received from the wireline network, and then passes the SAR$^+$-PDUs to the MAC layer, which transmits them to the peer entity at the mobile through the wireless channel. The SAR$^+$ sublayer at the mobile reassembles the SAR$^+$-PDUs received from the MAC layer into CS-PDUs, and the CS-PDUs will be passed up to higher layers. To simplify the conversion between the payloads of the SAR$^+$-PDU and SAR-PDU, only SAR$^+$-PDUs with payloads necessary to fill up an integer number of SAR-PDUs are allowed. For example, the payload of SAR$^+$-PDU has to be $m \times 48$ bytes for AAL5, where $m = 1/4, 1/3, 1/2, 1, 2, 3, \ldots$.

In the case where two mobile terminals communicate with each other without the AP's coordination, the data conversion is illustrated in Figure 9. Observe that no ATM layer operation is involved in this case.

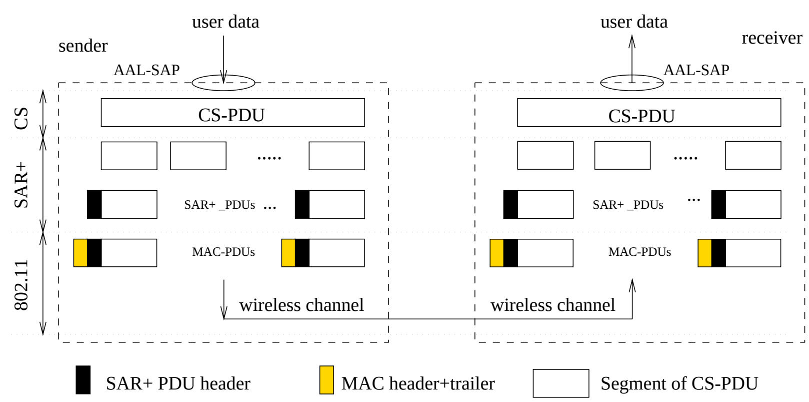

Figure 9: Data conversion between two *ad-hoc mobile terminals*

5.2 IS-95

IS-95 [23] is a digital cellular standard based on Code Division Multiple Access (CDMA). This standard was designed by the U.S. Telecommunications Industry Association (TIA) to be compatible with the existing U.S. analog cellular system (AMPS) frequency band. IS-95 uses 824 - 849 MHz band for the reverse link and 869 - 894 MHz for the forward link, and each channel occupies 1.25 MHz. It can provide maximum user data rates of 9.6 kb/s and 14.4 kb/s with different code rates. The user data rate can be changed in real-time, depending on the voice activity and requirement in the network. The reverse CDMA channel consists of access channels and reverse traffic channels. The access channel is used by the mobile unit to initiate communication with the AP and to respond to the page channel messages [22].

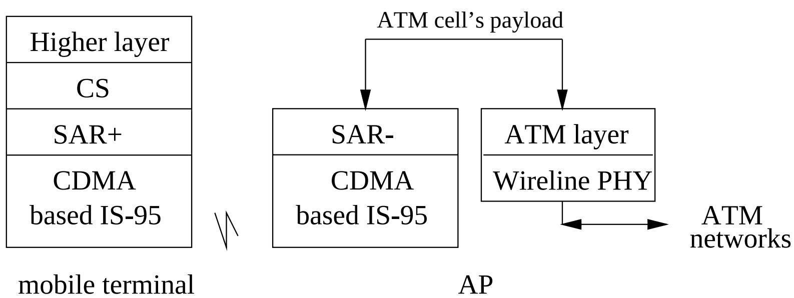

Figure 10: Implementation of IS-95 in the proposed architecture

Figure 10 illustrates how an IS-95-based cellular network can be used to provide ATM-like services to mobile users. The support of data applications at the SAR^+ and SAR^- sublayers with IS-95 is similar to that for 802.11 described above. In the following, we present a brief discussion about the support of constant bit rate (CBR) applications.

IS-95 provides hierarchical user data rates at 9600, 4800, 2400 or 1200 bps and 14,400, 7200, 3600 and 1800 bps with different code rates. This can support CBR applications with AAL1. As shown in Figure 11, when the mobile constantly sends data to the AP, the data enters the CS sublayer via the AAL-SAP at a constant rate, and they are used to form CS-PDUs. The maximum length of the CS-PDU for AAL1 is 47 bytes so that the SAR^+ sublayer at the mobile normally does not segment the CS-PDUs if the channel status is good enough for transmission of such long packets. In this case, the SAR^+-PDU header can be omitted for non-error-sensitive applications such as voice transmission to save bandwidth as shown in Figure 11. The CS-PDU is then passed down to the IS-95 layer to be transmitted to the peer entity at the base station. The SAR^- sublayer at the AP encapsulates the packets received from the IS-95 layer to SAR-PDUs, and then passes the SAR-PDUs to the ATM layer to form ATM cells. The reverse procedure takes place when the AP sends data to the mobile.

One major problem for voice transmission over ATM is the "cell payload assembly delay", which is the amount of time needed to collect enough data to fill a cell. This will affect the

QoS of some real-time applications such as voice transmission. For example, the time to collect 47 bytes (the user information payload of an ATM cell for AAL1) at the user data rate of 14.4 kbps is around 26 *msec*. Although this time can be reduced by sending cells that are only partially full by adding dummy bytes (refer to [24] for detail), transmitting the added dummy bytes will waste radio bandwidth. Therefore, the SAR^+ sublayer at the mobile needs to be able to filter the padding in AAL1 CS-PDUs before passing them to the IS-95 layer for transmission as shown in Figure 11. Meanwhile, the SAR^- sublayer at the AP needs to re-pad the received SAR^+-PDU to 47 bytes before forming SAR-PDUs, and vice versa.

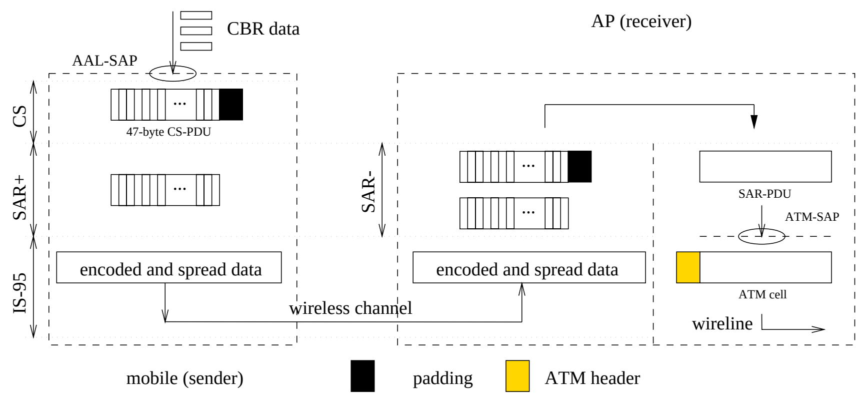

Figure 11: Data conversion between mobile and AP for AAL1

A major advantage of this implementation is hand-off support capability provided by CDMA-based IS-95, which is desirable in PCS. However, the maximum user data rates provided by IS-95 cannot satisfy some applications such as video transmission.

6 Conclusion

This paper first discussed some important issues of providing the ATM-like services to the mobile devices and introduced some existing proposals for wireless ATM, most of which are based on the "cellification over air" approach. Then the paper presented the analysis of "cellification over air" and found that transporting "small" ATM cells over wireless networks leads to inefficient utilization of the wireless channel capacity. Finally, we proposed a new architecture that gives flexibility to the designer of the wireless system to keep resource utilization as high as possible. With this architecture, the ATM layer is transparent to the mobiles, but they can still exploit the ATM capabilities to support multimedia applications. This is accomplished by splitting the SAR sublayer to the mobile and the AP. The implementations for some non-ATM air interface standards have been discussed.

References

[1] B. T. Doshi, A. Sawkar, An ATM Based PCS/Cellular Architecture, WINLAB 1995, pp23-35.

[2] D. Raychaudhuri, N. D. Wilson, ATM-Based Transport Architecture for Multiservices Personal Communication Networks, IEEE JSAC, Vol. 12, No. 8. October 1994, pp1401-1414.

[3] D. Raychaudhuri, L.J. French et al., WATMnet: A Prototype Wireless ATM System For Multimedia Personal Communication, ICC 1996, pp469-477.

[4] J. Porter and A. Hopper, An ATM based protocol for Wireless LAN, Olivetti Research Limited, April 19, 1994, from ftp:cam-orl.co.uk:/pub/docs/ORL/tr.94.2.ps.Z.

[5] ITU-T, Recommendation I.363, B-ISDN ATM Adaptation Layer (AAL) Specification, 03/93.

[6] CCITT, Recommendation I.321, B-ISDN Protocol Reference Model and Its Application, Geneva 1991.

[7] P. Agrawal, E. Hyden et al., SWAN: A Mobile Multimedia Wireless Network, *IEEE Personal Communications*, April 1996, pp18-33.

[8] H. Xie, P. Narasimhan et al., Data Link Control Protocols for Wireless ATM Access Channels, *1995 4th International Conference on Universal Personal Communications*, November 6-10, 1995 Japan, pp753-757.

[9] K.Y. Eng et al., A wireless broadband ad-hoc ATM local-area network, *Wireless Networks*, vol. 1 (1995), pp161- 174.

[10] K.Y. Eng et al., BAHAMA: A Broadband Ad-Hoc Wireless ATM Local-Area Network, ICC 1995, pp1216-1223.

[11] Mark J. Karol et al., An efficient demand-assignment multiple access protocol for wireless packet (ATM) networks, *Wireless Networks*, vol. 1 (1995), pp267-279.

[12] R. Yuan et al., Mobility Support in a Wireless ATM Network, WINLAB 1995, pp335-345

[13] Victor O.K. Li and X. M. Qiu, Personal Communication Systems (PCS), *Proceedings of the IEEE*, vol. 83, No. 9, September 1995, pp1210-1243.

[14] G.A. Halls, The high performance radio local area network standard, *Electronics & Communication Engineering Journal*, December 1994, pp289-296.

[15] IEEE Standards Department, IEEE 802.11 Draft Standard for Wireless LAN, P802.11D5.0, P802.11D3, 1996.

[16] Jost Weinmiller et al., Performance Study of Access Control in Wireless LANs - IEEE 802.11 DFWMAC and ETSI RES 10 HIPERLAN, from http://maui.ee.tu-berlin.de/bibl/ours/index.html.

[17] S. Gupta and M. El Zarki, Design of error control schemes for wireless data networks, *Wireless Computer Networks*, volume 3, pages 771–775, 1994.

[18] S. Gupta and M. El Zarki, Optimal design of error control schemes for packet radio networks, *IEEE International Conference on Personal and Wireless Communications*, 1994.

[19] S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood Cliffs, N. J., 1983.

[20] R. J. McEliece and L. Swanson, On the decoder error probability for reed-solomon codes, *IEEE Transactions on Information Theory*, 32(5):701–703, September 1986.

[21] Coutras and S. Gupta and N. Shroff,Transportation of Real time traffic over IEEE 802.11 WLANs, *In preparation.*

[22] T. S. Rappaport, *Wireless Communications*, Prentice-Hall, 1996, pp519 - 533.

[23] U.S. Telecommunications Industry Association, TIA/EIA/IS-95-A.

[24] The ATM Forum Technical Committee, Circuit emulation services interoperability specification (af-vtoa-0078.000), version 2.0.

Biographies

Shengming Jiang received the B.Sc degree in Computer Science and Applications from Shanghai Maritime Institute, Shanghai, China, in 1988, D.E.A. (MPhil equivalent) and Ph.D degrees in Computer Science in University of Paris VI and University of Versailles, France, in 1992 and 1995, respectively. From 1988 to 1990, he worked as a computer engineer in Nanjing Oil Transportation Company, Nanjing, China. He has been a research associate in Department of Electrical & Electronic Engineering since February 1995 and then in the Computer Science Department at The Hong Kong University of Science & Technology since February 1997. His research interests include medium access control (MAC) protocol design for ATM-compatible high-speed LAN and MAN and HFC systems, ATM networking, ABR service, wireless ATM and PCS.

Danny H.K. Tsang was born in Hong Kong. He received the B.Sc. degree in Mathematics and Physics from the University of Winnipeg, Winnipeg, Canada, in 1979, and the B.Eng. and M.A.Sc. degrees both in Electrical Engineering from the Technical University of Nova Scotia, Halifax, Canada, in 1982 and 1984, respectively. He also received the Ph.D. degree in Electrical Engineering from the University of Pennsylvania, Philadelphia, PA, in 1989.

He joined the Department of Mathematics, Statistics & Computing Science, Dalhousie University, Halifax, Canada, in 1989, where he was an Assistant Professor in the computing science division. He has joined the Department of Electrical & Electronic Engineering at the Hong Kong University of Science and Technology since summer of 1992. His current research interests include statistical modeling of variable-bit-rate video traffic, queuing analysis of ATM multiplexer, congestion controls in B-ISDN/ATM networks, and wireless ATM. He has been a member of the Technical Program Committee for IEEE INFOCOM since 1994. His email address is: eetsang@ee.ust.hk.

Sanjay Gupta received his Ph.D in Systems in 1993 and M.S. degree in Electrical Engineering in 1989, both from the University of Pennsylvania. He received his B.Tech degree in Electrical Engineering in 1988 from Indian Institute of Technology, Kanpur.

Dr. Gupta joined the GSM products division at Motorola in 1997 where he is working on Universal Mobile Telecommunications System (UMTS). Prior to that he was employed in Department of Electrical and Computer Engineering at Illinois Institute of Technology in 1993 as an Assiatant Professor. He has worked as visiting research scholar at Hong Kong University of Science and Technology during the summer of 1995 and 1996. He is currently working on problems in the area of integration of wireless and wireline communication networks. His other research interests include wireless local area networks, ATM based broadband networks, management and control of communication networks, and stochastic modeling and simulation. He edited a special issue of Wireless Networks on error control and has served as a Technical Program Committee member for the IEEE Infocom conference from 1993 to

1997 and the IEEE Vehicular Technology Conference 1995.