15th Bled Electronic Commerce Conference

e-Reality: Constructing the e-Economy

Bled, Slovenia, June 17 - 19, 2002

Security Issues in Mobile Commerce using WAP

Niels Christian Juul

Roskilde University, Department of Computer Science, Box 260, DK 4000 Roskilde, Denmark
Phone: +45 4674 3860 Fax: +45 4674 3072
ncjuul@acm.org

Niels Jørgensen

Roskilde University, Department of Computer Science, Box 260, DK 4000 Roskilde, Denmark
Phone: +45 4674 3702 Fax: +45 4674 3072
nielsj@ruc.dk

Abstract

The Wireless Application Protocol (WAP) has been proposed as a way to get Internet (or a sort of Internet) to the small wireless and mobile devices, e.g. mobile phones, while accommodating for the special characteristics of such devices.

Originally, WAP was designed with a gateway in the middle, acting as the interpreter between the Internet protocol stack and the Wireless Application Protocol stack. The WAP gateway forwards web content to the mobile phone in a way intended to accommodate the limited bandwidth of the mobile network and the mobile phone's limited processing capability. However, the gateway introduces a security hole, which renders WAP unsuitable for any security-sensitive services.

Through a set of standard releases, primarily version 1.2.1 (June 2000) and version 2.0 (July 2001), security issues have been addressed. We discuss the security hole and the gateway-based design that has led to it, including the business and architectural considerations underlying the design. A number of ways to correct the situation are discussed, including application level security, which still hasn't been fixed in the WAP 2.0 standard of the July 2001 release. Finally we observe, that although version 2.0 allows skipping the gateway thereby tightening security, the added cost is not negligible.

1. Introduction

This investigation of security vulnerabilities in mobile commerce focuses on WAP. The study is based on both a thorough analysis of the technical standards including comparisons with Internet standards and a business-oriented analysis of the stakeholders deploying WAP. Both are focussing on achieving secure mobile transactions using WAP.

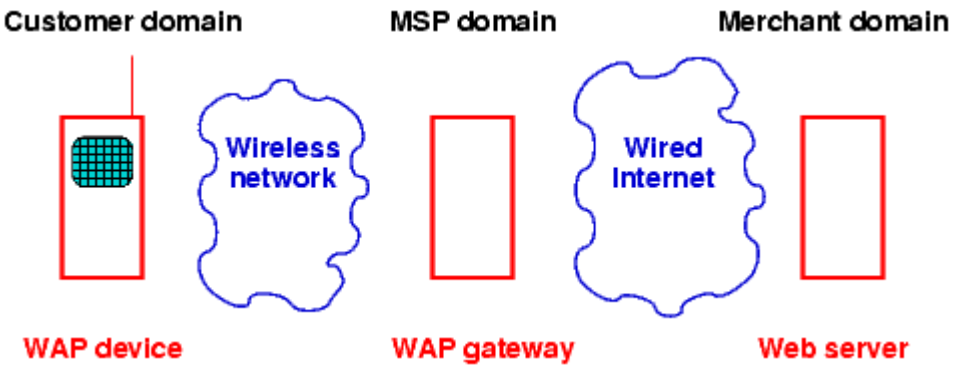

Figure 1 The three parties involved in a generic m-commerce transansaction using WAP: In the middle: The mobile telephone company (referred to as the mobile service provider, MSP) uses a WAP gateway to connect between the wired and wireless Internet. To the left: The customer uses a WAP-enabled mobile phone. To the right: The e-merchant's web site is connected to the fixed Internet. The networks illustrated are simplifications; the first is a mostly wireless, mobile access network, whereas the second is a fixed, mostly wired Internet.

A simple mobile transaction, which needs security, could be the following. A customer places an order with an e-merchant using a mobile device. In doing so, sensitive information is exchanged with the merchant, typically including credit card number, delivery address, etc. If there is a risk that the privacy of this data will be violated anywhere in between the parties, the customer is not likely to engage in this form of e-commerce, and as a consequence the e-merchant is not likely to invest in the technology either. By similar arguments, *authentication*, *message integrity*, and *non-repudiation* must also be supplied. For simplicity, this paper focuses on maintaining *privacy*.

Figure 1 illustrates the WAP solution to m-commerce. When the customer uses a mobile phone and the *Wireless Application Protocol (WAP)* [1,2], the privacy of the data is in fact not guaranteed. Even when encryption is used in accordance with WAP's security protocols [3,4,5] of WAP version 1.2.1, as illustrated in Figure 2, the WAP gateway constitutes a security hole since, inside the gateway, the data is transmitted in its original, un-encrypted form. What WAP fails to provide is end-to-end security, which we define as a secure communication channel between the two parties on top of a potentially insecure network. In WAP, there is a point (the gateway) between the two end points (the customer and the merchant) where the data may be compromised. This critique pertains to version 1.2.1 of the WAP standard, from June 2000.

The WAP gateway is software. Typically, it runs on a computer in a building under the control of the mobile service provider, MSP. Figure 1 illustrates the role of the mobile service provider as an intermediary in a WAP-based e-commerce transaction. Specifically, the security weakness of WAP discussed in this paper means that all data exchanged may be available to people with privileged access to the WAP gateway, e.g. a system administrator of the machine that the WAP gateway is running on. Thus, the privacy of the data depends on such things as the internal security policy of the mobile service provider, the methods used to grant computer access to technical staff, etc.

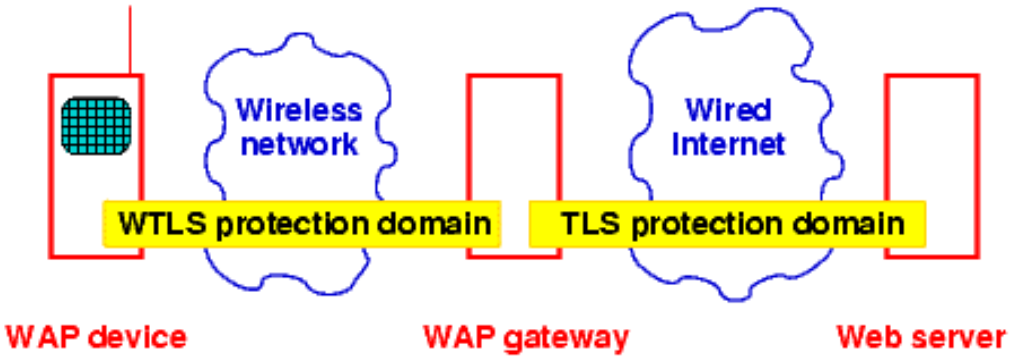

Figure 2 Security zones using standard security services (WTLS and TLS)

The existence of the security hole sketched above is commonly recognized (see e.g. [6, 7]), although companies behind the WAP standard have previously argued that it is negligible [8]. We have analyzed various strategies to avoid the security hole, taking into account the technical and business rationales that led to the gateway-based design and the associated security weakness, in particular the technical rationale, which is based on the limitations of the wireless network and the mobile phones [9].

Version 2.0 of the WAP standard from July 2001 suggests a way around the vulnerability at the gateway. It actually suggests not to use a gateway, but to utilize Internet protocols end-to-end. Besides the loss of the benefits from having a gateway, this also introduces further problems as discussed later.

In short, there are three major remedies to the breach of end-to-end security in WAP:

1. Putting the gateway inside the "vault"
2. Application level security on top of WAP
3. Enabling Internet on the mobile device

The remainder of the paper is organized as follows:

An overview of WAP is given in Section 2, followed by an explanation of the gateway security hole in Section 3. The security solution based on putting the WAP gateway inside the vault of the e-merchant is discussed in Section 4, while applying end-to-end security at the application level is discussed in Section 5. The WAP 2.0 solution of short-cutting the gateway and enabling Internet on the mobile device is discussed in Section 6. Section 7 compares WAP to iMode, the successful mobile Internet service provided by NTT DoCoMo in Japan. Section 8 concludes and discusses the standardization process in semi-opened forums like the WAP Forum.

2. The gateway-based design of WAP

The Wireless Application Protocol (WAP) [10,1,2][1] is a suite of evolving[2] standards for browsing the web with a thin client browser, e.g. a mobile phone. The standard describes a full suite, sometimes referred to as a "stack" of protocols, basically in compliance with the ISO-OSI model for network protocols [11]. The protocol stack of WAP is compared with the Internet protocol stack in Figure 3. Bear in mind that, although illustrated side by side, each protocol layer does not communicate across the two stacks. The obligation of the gateway is to utilize the two stacks and converts between the two protocols HTTP and WSP at the top.

This section introduces the WAP standard with particular emphasis on the WAP gateway. The objective is to identify the technical and business rationale behind the WAP gateway.

2.1 The technical rationale for the WAP gateway

In addition to mobile phones - our prototype example of a WAP device - the WAP standard aims at other hand-held digital devices such as pagers, two-way radios, smart-phones, etc. These devices are characterized by their limited capacities of: processing power, storage capacity, input/output (key-pad and display), and power (battery capacity).

WAP was designed to work not only with GSM but most other digital wireless telephone networks. Some bearers are illustrated on Figure 3. Compared to the well-known Internet, mobile wireless networks are characterized by: limited communication capacity (bandwidth), higher latencies, higher variation in packet-loss (jitter), and variation in long-term connectivity/availability (on/off).

[1] The reader is referred to the web site of the WAP Forum (www.wapforum.org) for the full list of standard documents, both approved and proposed standards.

[2] The standards discussed here are the two latest approved sets of WAP standards, version 1.2.1 of June 2000 and 2.0 of July 2001.

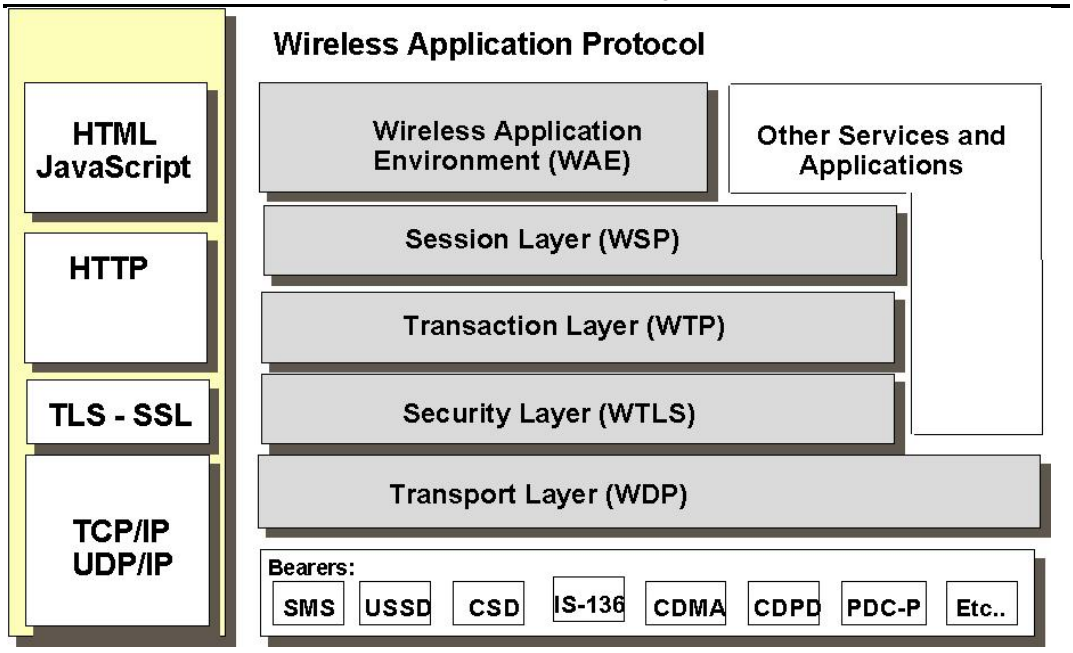

Figure 3 The WAP vs. the Internet Protocol Stack (Source: WAP Forum)

WAP's only requirement pertaining to a document to be delivered from a web server to a mobile phone is that the document is written in WML (Wireless Markup Language). WML is WAP's replacement of HTML and designed specifically for the small and diverging displays of the mobile phone. A WML document is ordered and sent in a request-reply cycle in which the roles of the WML browser, the gateway, and the web server are as follows (the reader may wish to refer again to Figure 1):

The request: The user of the mobile phone orders information by clicking on a link, which is shown on the display by the WML browser. This action spawns a so-called WSP (Wireless Session Protocol) request send from the phone to the WAP gateway. The WSP request parallels the HTTP request send when an ordinary browser requests a document from a web server. In the gateway, the WSP request is converted to a standard HTTP request which is send to the web server.

The reply: The web server's response to the request is the action of sending the WML document to the WAP gateway. After some processing in the gateway (see below), the modified document is send to the mobile phone.

In order for the WAP gateway to help dealing with the low-capacity wireless **network connection**, it has the following functions:

1. **Switching between transport protocols**

   The gateway in its communication with the mobile phone uses a transport protocol (WDP, for Wireless Datagram Protocol[3]), which is of a so-called connection-less and unreliable nature. In contrast, the transport protocol used by HTTP on the ordinary Internet is the connection-oriented and reliable protocol, TCP. In TCP, there is additional communication to ensure that all data actually gets through, and if not, it is re-transmitted. Omitting this in the wireless part of the connection incurs a risk of losing data but saves a significantly on latency and bandwidth.

2. **Compression**

   The gateway compresses the WML document before sending it to the mobile phone. The gateway applies so-called loss-less data compression, which means that the same information can be sent in a format that occupies fewer bites. The compression relies on the prede-

[3] For the purposes of this discussion WDP is identical to the Internet's UDP transport protocol

fined translation between the vocabulary of HTTP and WSP. In comparison, compression of text documents is not used by default on the ordinary Internet.

In addition, to accommodate the limited capacity of the **mobile phone**, the gateway also has the following functions:

3. **Compilation**

   If a WML document contains embedded source code, the gateway compiles such code into a so-called bytecode format, something that relieves the mobile phone of the task of parsing the code. WML documents may contain code written in the scripting language WML-script, which is similar to the JavaScript language that may be embedded inside HTML documents. Such code is executed inside the browser. For example, it may be used for validation, say, checking that a credit card number contains a certain number of digits and allowing the user to correct it before sending the information to the web server.

4. **Decompression**

   The gateway reads and interprets the original WSP request, which is in a compact form that cannot be understood by the web server. The web server understands lengthy HTTP requests in which various commands are given as ordinary text rather than by numerical codes. Also, the gateway translates a symbolic Internet address into an IP number. The gateway does this in the usual way by communicating with a DNS (Domain Name Service) server. Relieving the mobile phone from the use of lengthy HTTP-style commands and from doing DNS look-ups also reduces the amount of data that must be sent from the mobile phone.

Also, local validation of input inside the mobile phone (3) and avoiding DNS look-ups from the mobile phone (4) serve to eliminate request/reply message cycles between the mobile phone and the Internet. Thus, these features of the gateway also reduce consumption of the scarce network bandwidth of the wireless connection. Most importantly, they reduce the total latency between the user presses the button and the result is shown on the mobile phone.

While the gateway is essential for most of the techniques employed by WAP to accommodate the limitations of the mobile phone and the wireless network, other means such as various features of WML are independent of the gateway-based design. This includes the splitting of a WML document into "cards" send as one document. The user browses from one card to the next by clicking on links, thus navigating through all the cards within the same document (or "deck") without needing further communication with the server. This eliminates latencies originating from the time it takes to retrieve the next card in a request/reply cycle across the high-latency wireless network

2.2 The business rationale for the WAP gateway

In addition to the technical rationale, there is also a business rationale underlying WAP's gateway-based design. In particular, the gateway may open new business opportunities for the mobile service provider. The gateway's impact on business, and usability in a wider sense, may be summarized as follows from the point of view of the mobile service provider, the mobile customer, and the e-merchant.

A. The Mobile service provider (MSP) may use the gateway to tie its mobile customers to a portal-like access point to the Internet. Such portals are well-known on the fixed Internet, where they are provided to private/home-based users of the Internet by Internet Service Providers such as America On-line and Tiscali/World On-line. There are, however, two important aspects where a mobile portal may differ from an ordinary Internet-portal: First, WAP services may be combined with the MSP's basic phone services. (For the integration into WAP of basic phone services there is a dedicated protocol, Wireless Telephony Application Interface (WTAI) [12].) Second, the MSP of any given customer has a special, privileged position in the competition with other service providers available to the customer, because the MSP has access to privileged *location* information about the mobile customer, i.e. the exact cell in which the mobile phone is located. The customer's own MSP may utilize this information as part of location dependent WAP services, e.g. ordering take-away meals from nearby restaurants.

B. The mobile WAP user has a dual interest: On the one hand, she would presumably like to have access to those of her own MSP's services that integrate WAP and ordinary mobile telephone service, as made possible by accessing the Internet via the MSP's gateway. On the other hand, she is clearly interested in the cheapest possible, universal access to all WAP-based services on the Internet.

C. The typical e-merchant has one obvious interest in relation to WAP which we want to emphasize: the interest of minimizing the cost - in terms of hardware, software, maintenance etc. - of extending e-commerce to WAP. For most e-merchants, WAP-based e-commerce is just an additional channel to be added to e-commerce via the web. The chief advantage of the gateway is that the e-merchant can open this channel merely by hosting a new type of documents - WML documents - because the adaptation to the special requirements of the wireless network and devices are taken care of by the gateway.

As an aside, we note that the development of the WAP standard has been initiated, not by representatives of any of the above parties, but by producers of hardware and software for mobile devices and networks: Nokia, Motorola, Ericsson, Phone.com (previously Unwired Planted), and others. The main interests of these producers may be identified as selling hardware and software to the mobile service providers (and the companies that own the mobile networks), and mobile phones and other devices to the users. In turn, the prospects for this is of course highly depending on a successful integration with the Internet, including e-commerce.

3. The security hole at the WAP gateway

In general, the mobile customer and the e-merchant involved in an m-commerce transaction want (or should want) to ensure:

Confidentiality: that messages are kept secret.

Authentication: that each party knows whom the other party is.

Message integrity: that messages are passed unaltered from sender to receiver.

Replay attack prevention: that any unauthorized re-sending of messages is detected and rejected.

Non-repudiation: that neither party can later reject that the exchange took place.

All these issues must be addressed in a secure system. Indeed, it is the ambition of the WAP Forum to develop WAP into a standard that covers all relevant aspects of security, and some steps have been taking already with version 1.2.1 while version 2.0 goes a little bit further.

For security, WAP provides a secure protocol for data transport: WTLS, Wireless Transport Layer Security [3]. WTLS also contains features for authentication of both parties, as well as for non-repudiation using message digests and digital signatures. Authentication of the user of a GSM phone may utilize the phone's SIM card [5]. Finally, the definition of WML-Script (introduced above in Subsection 2.1, item "*3. Compilation*") includes a specification of a function library called a "crypto package" [4]. This library contains a signing function, which can be used by a user to digitally sign a message, in the conventional manner in a Public Key Infrastructure. Thus WAP supports non-repudiation of messages (such as a placement of an order) sent by a mobile WAP-user. (The cryptographic library and its possible usefulness in implementing so-called application-level security is discussed below in Section 5).

In the sequel, we merely discuss to what extent WAP achieves message confidentiality.

3.1 WAP's two-stage security model

Consider a WAP-based m-commerce transaction in which a document, say `www.weSell-It.com/orderNow.wml`, must be transferred to the mobile phone in a secure manner.

The basic security feature of WAP provides secrecy in the two half parts of the path that connects the WAP client and the web server: the (presumed) wireless path between the WAP client and WAP gateway, and the (presumed) wired path between the gateway and the web server (Figure 2).

In the middle point constituted by the gateway, incoming data is decrypted; then while the data is in its original, un-encrypted form, it is subjected to some processing; and finally it is encrypted again before it is sent off along the other path. The processing done on the un-encrypted data corresponds to the protocol layers above the security layers on Figure 3 in the gateway.

For encryption over the wired path, WAP simply relies on the Transport Layer Security (TLS) protocol [13], a widely used Internet standard. TLS is standardized by the Internet Engineering Task Force, and is also known under the name Secure Socket Layer (SSL) given to it by Netscape, the company that developed the standard originally.

For encryption over the wireless path, WAP uses WTLS [3], which is in essence an adoption for wireless communication of the TLS protocol. The changes to TLS embodied in WTLS do not weaken security.

WAP's topmost protocol, the Wireless Session Protocol (WSP) [14], initiates and links the two secure halves of the connection. The major steps in this are as follows: The user of the WAP client selects the URL `https://www.weSellIt.com/orderNow.wml`. This tells the WSP layer at the WAP client to initiate the setting up of a WTLS connection to the WAP gateway, and then pass a WSP request (the equivalent of an HTTP request) over the connection for the particular file. Thus prompted by the WAP client, the WSP layer at the gateway will initiate a TLS connection to the web server, in a way completely similar to setting up a connection between an ordinary web client and the server.

This combination of WTLS and TLS provides secrecy (indeed, also integrity) over both halves of the WAP client / web server connection. The crucial weakness is, of course, that all data transferred between the WAP client and the web server is decrypted at the WAP gateway, i.e., all data such as credit card numbers, etc. exists as free text in the memory of the gateway. Technical solutions, such as various programming techniques applied to the software that implements the WAP gateway, can make it somewhat difficult to get access to the data, but not impossible. Organizational solutions, such as tightening the security policy of the organization that hosts the WAP gateway, may limit access to the gateway and it's data; but from the point of view of the two "end users" it is unsatisfactory that the privacy of their data is not under their own direct control.

3.2 Encryption vs. the functions of the gateway

The breach of end-to-end security at the WAP gateway is not an accidental error. Rather it is merely a disadvantage of the gateway-based design which the designers felt was outweighed by its advantages. To argue this, before we discuss (in Sections 4, 5, and 6) how the gateway-based design can be circumvented, let us consider again the technical rationale underlying that design. The question is, which of the functions (listed in Section 2 as 1, 2, 3, and 4) could still be fulfilled by the gateway had the WAP standard prescribed a different approach to confidentiality, one that attained end-to-end security as provided by TLS on the ordinary web?

We believe the answer is that it would probably be possible to use the gateway as a point of switching between the two transport protocols TCP and WDP (cf. function 1), but that it would not be possible for the gateway to employ compression of WML or compilation of WML-Script, nor for the mobile phone to use the bandwidth-saving WSP requests (cf. functions 2, 3, and 4).

To introduce end-to-end encryption while retaining the gateway's function as the point of switching between the transport protocols would be possible by letting the gateway act as a proxy for the WAP client at the level of TCP, similarly to the way it already acts as a proxy at the level of HTTP. A deeper discussion of this strategy is beyond the scope of this paper.

The remaining functions of the gateway cannot be achieved if end-to-end encryption is employed. Clearly, if code written in WML-Script is encrypted, it cannot be compiled, unless the gateway can first decrypt the script, thus breaking the end-to-end encryption. For the same reason, it would not be possible for the gateway to understand (decompress) the encrypted WSP requests, so the mobile phone would have to use the much higher latency-incurring and more bandwidth-consuming HTTP requests. Finally, the gateway's compression of ordinary WML documents is

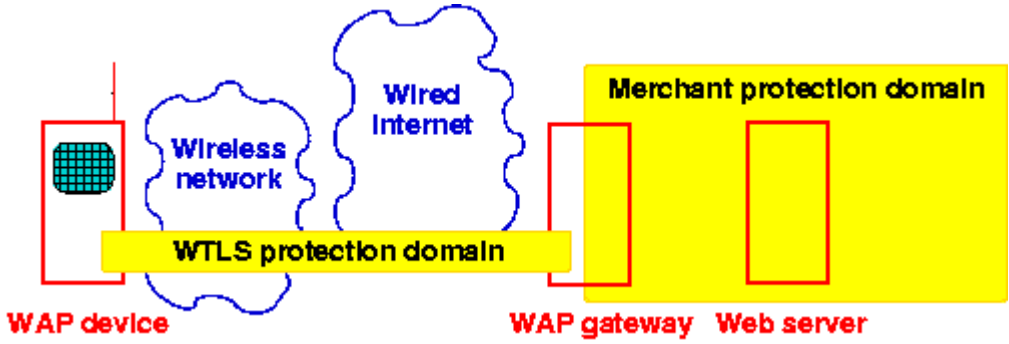

Figure 4 Security zones by replacement of the WAP gateway. The customer may utilizes a dial-up point at the MSP and continue across the Internet, or utilizes the Merchants own dial-up point bypassing the Internet. The secure channel is protected by WTLS in both cases.

ruled out: If the gateway cannot recognize, for example the WML tags `card` and `/card`, it cannot compress them (by replacing them with numbers that consume fewer bytes). In general, the characters of (well) encrypted text are randomly distributed, i.e. with no apparent patterns, and therefore such data cannot be compressed. (For a discussion on encryption versus compression, see [15])

Thus modifying the WAP standard to provide full end-to-end encryption as in TLS conflicts with the compilation, decompression, and compression functions of the gateway that serve to limit the amount of data that has to be transmitted over the wireless network, although, on the other hand, it does not prohibit the use of the fast WDP protocol on that part of the connection.

4. Solution 1: Putting the gateway inside the "vault"

The WAP gateway can be placed at the web server end of the connection, that is, inside the same security zone. When residing inside the local network of the merchant the gateway is protected against the outside world in a similar fashion to the way the web server is protected (Figure 5).

A closer look at the business potential for the three stakeholders reveals that this solution does introduce further problems as well. Whether these will be solved is hard to say at the moment.

4.1 The Mobile Service Provider

For the Mobile Service Provider the main problem is a possible loss of business opportunities, e.g. "locking in" the customer to the provider's m-portal. There are a number of possible WAP-services that the MSP can offer only if it hosts the gateway, this being the MSP's only way of ensuring that all data used in those services stays within the mobile network that the MSP is in control of. For instance, location dependent services utilize the MSP's access to information from its own network about the exact location (cell) of the mobile phone. There may be strong business and security reasons for the MSP to not want such data to be available outside of the network. In the case of location information, the data is clearly sensitive, at the same time, the data may have a high business value, exactly because it is a prerequisite for certain services. Moreover, this solution challenges the MSP to open its network to outsiders like the merchants running their own gateway. Each limitation on the MSP's full control over the usage of its networks may introduce further *traffic management problems* [8].

4.2 The User

The user of the WAP client may witness degraded performance: The WAP protocols, which are tailored for the characteristics of wireless networks, are now used for the entire transport of the

web content to the WAP client, instead of merely for the wireless part as intended. This may incur increased delays if there is congestion in the wired network.

The end-user interface becomes less friendly, because the user of the WAP client will be forced to swap between gateways during Internet browsing. For example, the user that wants to buy from two distinct web sites needs to use the WAP gateways of those sites (given that both transactions are secured by the method discussed in this section).

Swapping gateways raises two problems when changing the gateway profile of the phone. The gateway profile includes a dial-up phone number, the IP number of the gateway, etc. First, in many current implementations of WAP, the user of a WAP device has to go through a cumbersome procedure that involves clicking through several menus, the typing of an Internet address in the form of an IP number, etc. Future WAP implementations may provide features for easy switch of gateway, however it seems that the user should, at least, be asked to confirm a gateway switch. Second, different gateways may have slightly different properties themselves. Although standardized by the WAP Forum, current gateways perform differently on the same task. Furthermore, users facing problems, e.g., quality of service, may have a hard time resolving whether the problem is with the MSP, the merchant, or the mobile phone manufacturer [8].

The option of the mobile service provider running a kind of default gateway - one that the user may switch back to, after using a particular e-merchant's gateway - is not as simple as it may appear at first. A default gateway either requires that the client must explicitly "log out", or alternatively, a "time out" mechanism must be included in the WAP device, letting it switch back automatically.

4.3 The Merchant

The e-merchant is burdened with a complex piece of additional software (the gateway) that must be acquired and maintained. For a WAP gateway, maintenance work includes, for example, ensuring that the security-related software in the gateway is up-to-date, and updating the gateway to new versions of the WAP protocols. Furthermore, location based services, micropayment, etc. may not be available to the m-commerce application neither from the MSP-detached gateway, nor from the short-cut MSP directly.

It is, however, not clear, whether the merchant will be further *"burdened with handset provisioning and activation issues"* as suggested in [8]. Our experience shows that those issues must be addressed by the web server servicing the WML anyhow due to the huge variation in how a WML document is interpreted in the various WML browsers and mobile phones.

4.4 Solution 1 is not perfect

We argue that this solution to the security problem conflicts with one of the fundamental purposes of the WAP gateway, namely to convert between two distinct protocol suites, one for the wireless network (WAP) and one for the traditional wired network (the Internet protocol suite, including HTTP and TCP). One crucial point of difference between these two stacks is the protocols used for transport. Assuming that the wired and wireless networks are sufficiently different to justify the use of the different transport protocols, the gateway should be placed at the intersection of the two networks, so that transportation over both networks uses the appropriate protocols.

In [7], the WAP Forum proposes a variety of gateways and proxies including navigation information to enable the mobile phone to switch between gateways. It remains questionable whether this added complexity will be implemented everywhere following WAP 2.0. While it enables end-to-end security between the transport layer in the handset and the merchant gateway utilizing WTLS [3,16,17], an additional burden is placed on the merchant to also maintain the entire WAP stack of protocols.

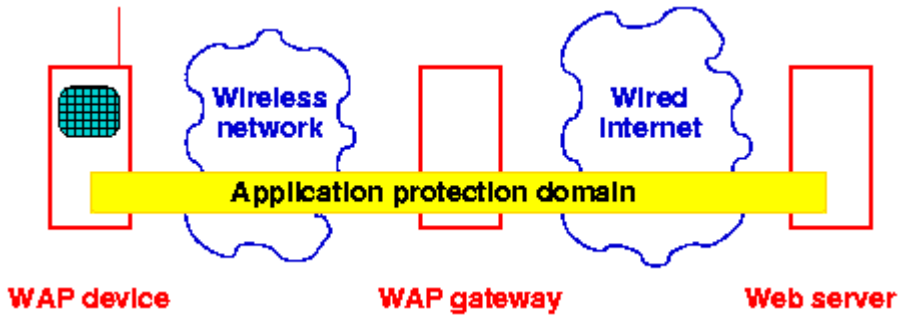

Figure 5 Security zones using application level encryption

5. Solution 2: Application level security on top of WAP

This amounts to introducing security at a software layer above WAP, and considering WAP merely as a potentially insecure communication means (Figure 5). Instead of using WAP's protocol for secure transport (WTLS), security is taken care of by means of dedicated software running at the two "ends", the mobile phone and the e-merchant's web server. Such software could perform encryption in a way that eliminated the security hole at the gateway. In general, the approach would be in line with the conjecture about protocol design made in [18]: that an end-to-end function must be placed at a level where end-to-end control is available, i.e., at the application level.

While technically possible, this solution would add a burden of extra complexity in the mobile phone. Indeed, the implementation of application-level security on top of WAP requires that sophisticated cryptographic functionality is made available to applications on the mobile WAP phone, either in the form of future enhancements of the WML Script Crypto Library [4,19] or in other ways. It should be noted that the single function for digitally signing a message, which is presently the only function contained in the library (as of WAP 2.0), is insufficient. Specifically, the Crypto Library's signing function is not useful for encryption of data, because in principle, everyone in possession of the WAP user's public key can read a message that she has signed digitally.

Moreover, this approach would partly neutralize most of the optimization provided by the WAP gateway: there will be a loss of the benefits of the data conversion and compression taking place in the gateway to accommodate the limited bandwidth of the wireless network (since encrypted data cannot be compiled or compressed/decompressed, as mentioned in Subsection 3.2). To minimize this loss, a balanced approach may be to encrypt only small fragments of the data, not the data as a whole. This would allow the WAP gateway to perform compression, decompression, and compilation by providing access to the WML tags and WSP/HTTP commands in their original, unencrypted form as discussed in Section 2.

Finally, the application-level security approach requires infrastructural remedies to convince the consumer about the security level available. At least something as convincing as the *signed* or *certified applets* on the ordinary web is needed. To enable end-to-end security between the mobile phone and the application-server [20] proposes further infrastructure. Current solutions to the end-to-end security problem on the application layer still lack standardization and support from the WAP Forum.

6. Solution 3: Enabling Internet on the mobile device

The third and last approach is to re-design the WAP protocol to not use a gateway, and employ the existing Internet standards, including the transport protocol (TCP), for the entire wired and wire-

less part of a connection. By definition, this solves the security problem introduced by the gateway.

This change of design has been proposed by the WAP Forum for version 2.0 of the WAP protocol [2]. The standard now allows a range of different gateways, corresponding to having the conversion between the two protocol stacks anywhere from the top to the bottom of the stack. At the top layers, the gateway works like the traditional WAP gateway taking care of the functions mentioned in Section 2.1, and at the bottom layer it works like a traditional bridge/router. Whereas a TCP-level gateway allows for two version of TCP, one for the wired and another for the wireless network, on top of which a TLS channel can be established all the way from the mobile device to the server.

The move away from top-level conversion constitutes a fundamental change of design, which does away not only with the security problem, but also the optimization for the wireless network, made possible by the gateway.

Discarding the WAP gateway makes it possible to attain the same high level of security for an m-commerce transaction as that of an e-commerce transaction on the ordinary web using full end-to-end encryption. Indeed, for WAP to discard the WAP gateway would turn the (fully) WAP-enabled mobile phone into an ordinary Internet device.

Besides the cost of deploying a full Internet protocol stack in the mobile phone, this solution enforces additional messages to be send between the device and the Internet. A simple request for a WML document must - in contrast to the gateway-based WAP solution - wait for a DNS-lookup to derive the appropriate IP-number from the symbolic name in the original request. The gateway-based solution only has this cost over the wired Internet, when the gateway converts the WAP-request to an HTTP-request. A minor optimization would be to utilize a caching DNS server at the border between the wired and the wireless network.

7. Mobile Internet in Japan: a comparison with iMode

A number of wireless Internet solutions that differ from WAP have been deployed in Japan. In terms of subscription numbers, the most successful is iMode, launched by NTT DoCoMo in february 1999. In December 2001, the number of iMode subscriptions exceeded 29.5 millions, by far outnumbering the total number of WAP users in the world.

Besides the service's success, there are two striking differences between iMode and WAP:

The first is that iMode is based more closely on existing Internet and Web standards, in particular TCP for the data transport and HTML as the mark up language for documents. iMode uses a version of TCP termed as "wireless profiled" TCP. (A similar protocol been incorporated into WAP as of WAP version 2.0, where it is one of a number of alternative transport protocols). The mark up language is called iHTML. In contrast to XML-based WML, iHTML shares a common subset with HTML, so that (simple) Web-content can be created which is readable by HTML-browsers on PCs as well as by iMode-compatible mobile phones.

The second is that iMode is a proprietary standard controlled by NTT DoCoMo. Users must purchase an iMode mobile phone, and connect to DoCoMo's iMode service. Via DoCoMo's iMode service, which resembles a portal, the user can access two kinds of sites run by other companies: so-called official and unofficial sites. For a site to become official requires approval by DoCoMo and entails the privilege to charge for the site's services via DoCoMo's billing system (in return for a 9% fee to DoCoMo). The protocols and other standards that iMode is based on have not been fully publicized nor subjected to as much public analysis as those of WAP. The account in this section is based mainly on [21,22].

In emphasizing iMode's conventional approach and proprietary nature we are not asserting that they are causally related. On the contrary, it would appear that in controlling the standard, NTT DoCoMo would have been in a better position to enforce an unconventional and more sophisticated approach such as WAP. Also, the success of iMode may derive from other factors, one of which may be the cheap rates charged for data transfer. In turn these are due to the underlying

mobile network being packet-switched already from the outset in 1999, something that is only currently being introduced in the GSM network (in the form of GPRS, i.e. the Generation "2.5").

With regard to security, NTT DoCoMo in March 2001 enhanced iMode with TLS/SSL as a means of providing end-to-end security between iMode-sites and iMode users. This corresponds to the approach taken most commonly on the wired Internet, and the move was facilitated by the decision to base iMode on the conventional Internet-protocol TCP. However, the success of iMode does not seem related to the absence of a security hole as that of the WAP gateway, since security sensitive services such a e-commerce have not been among the most popular in iMode.

8. Conclusion

Discarding the WAP gateway implies, of course, a loss of the optimization it provides. Even when more powerful mobile phones are developed, and higher bandwidths are attained in future wireless networks, high latency will remain, and so does the relevance, in particular, of the idea underlying WAP's WSP protocol that all data should be obtained by the mobile phone in a single request/reply cycle over the wireless part of the network connection. The fact that the WAP Forum proposes this strategy (at least as one alternative) seems to confirm our critique of the WAP standard [9]. There are inherent difficulties associated with the two other approaches: to place the WAP gateway at the web server end of the connection, or to use application level security on top of WAP.

The WAP Forum has not released the considerations that actually led to the re-design that does away with the WAP gateway. Similarly, although the security hole associated with the gateway is commonly recognized, the WAP Forum did not release its previous deliberations as to why it felt the gateway's advantages would outweigh its disadvantages.

The introduction and subsequent de-route of the WAP gateway - whether preconceived or not - has increased the complexity of the WAP standard, comprising variations both with and without the gateway. The WAP standard has become more difficult to standardize, and difficult to implement for providers of software for mobile phones, e-merchants and other content providers.

We foresee a future where all the communication oriented WAP standards are replaced by Internet standards, leaving only the application level protocols in WAE as WAP-specifications.

It is also interesting to observe how the old and open Internet technology is able to outperform a vendor provided network solution like WAP. Meanwhile the proprietary iMode seems to succeed in part by embracing Internet standards. Though WAP Forum has engaged in a semi-open standardization process, the decision process has been reserved for venders paying a high entrance fee. One may hope that the significance of a fully open discussion and standardization process will be learned also by the vendors behind the semi-opened WAP Forum.

Acknowledgment

An equipment grant from Siemens and a WAP/GSM-air-time grant from the National Danish Telecommunication Company, TDC are greatly acknowledged. We are in debt to Torben Lauritzen of Siemens for many fruitful discussions on WAP. A private e-mail correspondence with CEO Scott Goldman of WAP Forum and the discussion at the SSGRR conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet in L'Aquila, Italy in August 2001 further helped shaping the paper. The constructive comments from the referees of the Bled program committee have further improved the presentation.

This research was also funded by The Research Center for Electronic Commerce at the Copenhagen Business School and partly also by the IT-University of Copenhagen.

References

[1] WAP Forum, "WAP Architecture: Wireless Application Protocol Architecture Specification," Approved Specification (Part of the June 2000 (WAP 1.2.1) conformance release): WAP 100, WAP Forum Ltd., URL: `http://www.wapforum.org/`, Apr. 1998.

[2] WAP Forum, "WAP Architecture: Wireless Application Protocol Architecture Specification," Approved Specification (Part of the July 2001 (WAP 2.0) conformance release): WAP 210, WAP Forum Ltd., URL: `http://www.wapforum.org/`, July 2001.

[3] WAP Forum, "WAP WTLS: Wireless Application Protocol Wireless Transport Layer Security Specification," Approved Specification (Part of the June 2000 (WAP 1.2.1) conformance release): WAP 199, WAP Forum Ltd., URL: `http://www.wapforum.org/`, Feb. 2000.

[4] WAP Forum, "WMLScript Crypto Library: Wireless Application Protocol WMLScript Crypto Library Specification," Approved Specification (Part of the June 2000 (WAP 1.2.1) conformance release): WAP 161, WAP Forum Ltd., URL: `http://www.wapforum.org/`, Nov. 1999.

[5] WAP Forum, "Wireless Application Protocol Identity Module Specification (WIM) part: Security," Approved Specification (Part of the June 2000 (WAP 1.2.1) conformance release): WAP 198, WAP Forum Ltd., URL: `http://www.wapforum.org/`, Feb. 2000.

[6] Sandeep Singhal, Thomas Bridgman, Lalitha Suryanarayana, Daniel Mauney, Jari Alvinen, David Bevis, Jim Chan, and Stefan Hild, *WAP - the Wireless Application Protocol Writing Applications for the Mobile Internet*, ACM Press, 2001.

[7] WAP Forum, "WAP Transport Layer End-to-end Security: Wireless Application Protocol," Approved Specification (Part of the July 2001 (WAP 2.0) conformance release): WAP 187, WAP Forum Ltd., URL: `http://www.wapforum.org/`, June 2001.

[8] Phone.Com, "Understading Security on the Wireless Internet. How WAP security is enabling wireless e-commerce applications for today and tomorrow," White Paper WPSE-R4-001, Phone.com, Jan. 2000.

[9] Niels Christian Juul and Niels Jørgensen, "WAP may stumble over the gateway (security in wap-based mobile commerce)," in *Proceedings of the SSGRR 2001, Internation Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, Veljko Milutinovic, Ed., L'Aquila, Italy, Aug. 2001, Scuola Superiore G. Reiss Romoli (SSGRR), CD-ROM.

[10] Wireless Application Protocol Forum Ltd., Ed., *Official Wireless Application Protocol: The Complete Standard with Searchable CD-ROM*, Wiley Computer Publishing, 2000.

[11] Andrew S. Tanenbaum, *Computer Networks*, Prentice Hall, third edition, 1996.

[12] WAP Forum, "WAP WTAI : Wireless Application Protocol Wireless Telephony Application Interface Specification," Approved Specification (Part of the June 2000 (WAP 1.2.1) conformance release): WAP 170, WAP Forum Ltd., URL: `http://www.wapforum.org/`, July 2000.

[13] T. Dierks and C. Allen, "The TLS protocol version 1.0," RFC 2246, Internet, URL: `http://www.rfc-editor.org/rfc/rfc2246.txt`, Jan. 1999.

[14] WAP Forum, "WAP WSP: Wireless Application Protocol Wireless Session Protocol Specification," Approved Specification (Part of the June 2000 (WAP 1.2.1) conformance release): WAP 203, WAP Forum Ltd., URL: `http://www.wapforum.org/`, May 2000.

[15] Shmuel T. Klein, Abraham Bookstein, and Scott Deerwester, "Storing text retrieval systems on CD-ROM: compression and encryption considerations," *ACM Transactions on Information Systems*, vol. 7, no. 3, pp. 230–245, July 1989.

[16] WAP Forum, "WAP WTLS: Wireless Application Protocol Wireless Transport Layer Security Specification," Approved Specification (Part of the July 2001 (WAP 2.0) conformance release): WAP 261, WAP Forum Ltd., URL: `http://www.wapforum.org/`, June 2001.

[17] WAP Forum, "WAP TLS Profile and Tunneling WAP-219-TLS: Wireless Application Protocol TLS Profile and Tunneling Specification," Approved Specification (Part of the July 2001 (WAP 2.0) conformance release): WAP 219, WAP Forum Ltd., URL: `http://www.wapforum.org/`, Apr. 2001.

[18] J. H. Saltzer, D. P. Reed, and D. D. Clark, "End-To-End Arguments in System Design," *ACM Transactions on Computer Systems*, vol. 2, no. 4, pp. 277–288, Nov. 1984, Revised version of a paper from the Second International Conference on Distributed Computing Systems, Paris, France, April 8-10, 1981, pp. 509-512.

[19] WAP Forum, "WMLScript Crypto Library: Wireless Application Protocol WMLScript Crypto Library Specification," Approved Specification (Part of the July 2001 (WAP 2.0) conformance release): WAP 161, WAP Forum Ltd., URL: `http://www.wapforum.org/`, June 2001.

[20] Diego Ponce and Miguel Soriano, "Secure Intelligent Broker for m-Commerce," in *Proceedings of the SSGRR 2001, Internation Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, Veljko Milutinovic, Ed., L'Aquila, Italy, Aug. 2001, Scuola Superiore G. Reiss Romoli (SSGRR), CD-ROM.

[21] Peter Ashley, Heather Hinton, and Mark Vandenwauver, "Wired versus Wireless Security: The internet, WAP and iMode for E-Commerce," in *Proceedings of the 17th Annual Computer Security Applications Conference (ACSAC)*, Dec. 2001.

[22] Veronika Megler, "iMode. From bandwidth problem into Internet phenomenon," Feb. 2002, URL: `http://www-106.ibm.com/developerworks/library/wi-imode/?dwzone=wireless`.