EURASIP Journal on Applied Signal Processing 2005:1, 57–66


# A Proxy Architecture to Enhance the Performance of WAP 2.0 by Data Compression

**Zhanping Yin**
*Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada V6T 1Z4*
*Email: zhanping@ece.ubc.ca*

**Victor C. M. Leung**
*Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada V6T 1Z4*
*Email: vleung@ece.ubc.ca*

*Received 11 June 2004; Revised 17 November 2004; Recommended for Publication by Weihua Zhuang*

This paper presents a novel proxy architecture for wireless application protocol (WAP) 2.0 employing an advanced data compression scheme. Though optional in WAP 2.0, a proxy can isolate the wireless from the wired domain to prevent error propagations and to eliminate wireless session delays (WSD) by enabling long-lived connections between the proxy and wireless terminals. The proposed data compression scheme combines content compression together with robust header compression (ROHC), which minimizes the air-interface traffic data, thus significantly reduces the wireless access time. By using the content compression at the transport layer, it also enables TLS tunneling, which overcomes the end-to-end security problem in WAP 1.x. Performance evaluations show that while WAP 1.x is optimized for narrowband wireless channels, WAP 2.0 utilizing TCP/IP outperforms WAP 1.x over wideband wireless channels even without compression. The proposed data compression scheme reduces the wireless access time of WAP 2.0 by over 45% in CDMA2000 1XRTT channels, and in low-speed IS-95 channels, substantially reduces access time to give comparable performance to WAP 1.x. The performance enhancement is mainly contributed by the reply content compression, with ROHC offering further enhancements.

**Keywords and phrases:** wireless networks, wireless application protocol, wireless proxy.

## 1. INTRODUCTION

Wireless Internet access is an emerging service that is considered central to the commercial success of the next-generation cellular networks. The wireless application protocol (WAP) is the convergence of three rapidly evolving network technologies: wireless data, telephony, and the Internet. It is the de facto world standard for the presentation and delivery of wireless information services on mobile phones and other wireless terminals. WAP is a result of continuous work to define an industry-wide specification for developing applications that operate over wireless communication networks [1]. The WAP specifications address mobile network characteristics and operator needs by adapting existing network technology to the special requirements of mass-market, handheld wireless data devices and by introducing new technology where appropriate.



WAP 1.x is a standard aimed at optimizing the performance of wireless Internet access under such limitations as low bandwidth, high latency, less connection stability, and bearer availability for wireless networks, and limited screen display area, input facilities, memory, processing, and battery power for the mobile handset. The WAP Forum released version 2.0 of WAP in July 2001. WAP 2.0 brings the wireless world closer to the Internet by adopting the most recent Internet standards and protocols. It also optimizes the usage of emerging wireless networks with higher bandwidths and packet-based connections and maintains compatibility with WAP 1.x compliant contents, applications, and services. A major development of WAP 2.0 is that it provides support for standard Internet protocols such as transmission control protocol (TCP) and hypertext transfer protocol (HTTP), and permits applications and services to operate over all existing and foreseeable air-interface technologies and their bearer services, including general packet radio service (GPRS) and third-generation (3G) cellular standards such as WCDMA and CDMA2000 [2]. In particular, WAP 2.0 utilizes the wireless profiled TCP (WP-TCP) [3] and wireless profiled HTTP

(WP-HTTP) [4] that are optimized for wireless networks and interoperable with TCP and HTTP, respectively.

While some performance evaluations of WAP are found in the literature, they are mainly based on simulations employing theoretical traffic models. WAP performance over GPRS and global system for mobile communications (GSM) networks has been studied and several traffic models developed in [5, 6, 7]. WAP end-to-end security issues have been discussed in [8, 9, 10], and collocating the gateway with the WAP server in the secured enterprise site seems to be the only viable solution that strictly guarantees end-to-end security [8]. All these studies were based on the WAP 1.x protocol stack. There is little work done in evaluating WAP performance in realistic networks using real WAP traffic. Also, since WAP 2.0 has been newly released, there has not been any comparison of the performance of WAP 2.0 stack against WAP 1.x stack in the literature.

Compared with wireline, wireless bandwidth is a scarce resource. However, most data applications and web contents have been developed for wireline networks. To improve bandwidth utilization, data compression schemes can be used when these applications or data are accessed over wireless networks. While many standards exist for the compression of audio and video data [11, 12], and for data transmissions over voice band modems [13], WAP requires the application of data compression over the wireless network at the wireless transaction layer in a manner that is transparent to the wireless data bearer service. In WAP 1.x, a content encoding approach is used at the WAP gateway to compress the data. Although WAP 2.0 is an evolutional step forward, by adopting the HTTP/TCP/IP stack, it also has some disadvantages compared to the WAP 1.x protocol stack employing the wireless session protocol (WSP) and wireless transaction protocol (WTP); for example, the same message is transmitted using a much larger number of bits, and the same session requires more transactions. Therefore, content compression should also be used in WAP 2.0; but suitable compression methods for WAP 2.0 that preserve end-to-end security have not yet been standardized, nor has the performance of data compression in WAP 2.0 been evaluated.

In this paper, a novel proxy architecture employing an advanced data compression scheme is introduced for WAP 2.0 to minimize the air-interface traffic without protocol conversions. It also overcomes the end-to-end security problem in WAP 1.x. The performance of the data compression proxy scheme is compared against the standard WAP 2.0 proxy configuration and WAP 1.x protocol stack through experimental measurements over different emulated wireless networks. Results show that the proposed data compression scheme significantly improves the WAP 2.0 performance in all cases. Our results enable appropriate configuration of the WAP 2.0 protocol stack for various bearer services.

The rest of the paper is organized as follows. In Section 2, we review the WAP proxy architecture and the end-to-end security issue, and describe the proposed data compression scheme for WAP 2.0. In Section 3, we introduce the simulation method and the performance evaluation criteria. The experimental results are presented and discussed in Section 4. Some conclusions are given in Section 5.

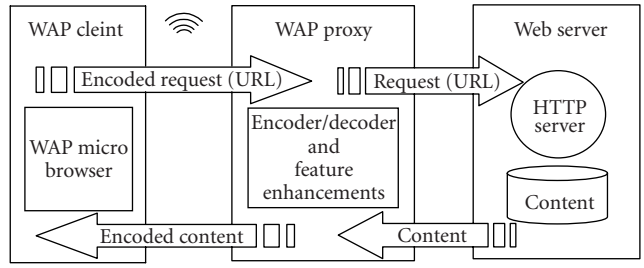


Figure 1: WAP proxy model.

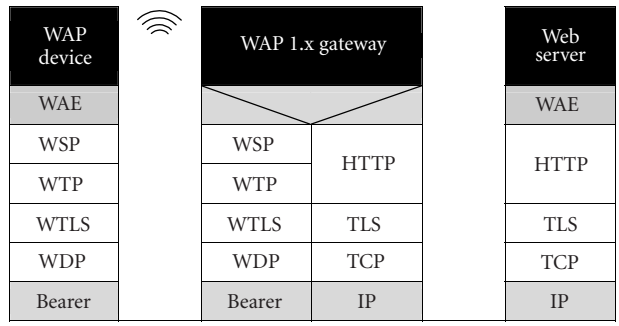


Figure 2: Standard WAP 1.x network configuration.

## 2. WAP PROXY ARCHITECTURE FOR DATA COMPRESSION

### 2.1. *WAP proxy model*

The WAP programming model is an extension of the world wide web (WWW) programming model with a few enhancements such as Push model and support for wireless telephony application (WTA). In WAP 2.0, the WAP proxy is optional, since the communication between the client and server can be conducted using HTTP 1.1. However, deploying a proxy, as shown in Figure 1, can optimize the communication process and may offer mobile service enhancements, such as location, privacy, and presence-based services. In addition, a WAP proxy is necessary to offer Push functionality [1, 2].

In the WAP 1.x configuration (Figure 2), the proxy, also known as WAP gateway, is required to handle the protocol interworking between the client and the content server. A WAP 1.x gateway essentially implements both the WAP 1.x and Internet protocol stacks within the same node. It is used for protocol conversions between these two protocol stacks, and the conversion between text-based wireless markup language (WML) documents in the Internet domain and binary-encoded bytecode in the wireless domain. The WAP gateway communicates with the client using the WAP 1.x protocols: WSP, WTP, wireless datagram protocol (WDP), with data security provided by the wireless transaction layer security (WTLS) protocol, and it communicates with the content server using the standard Internet protocols (HTTP/TCP/IP), with data security provided by the transport layer security (TLS) protocol.

### 2.2. WAP end-to-end security

Although WAP 1.x protocol conversion and content encoding minimizes the air-interface traffic, and WAP 1.x can preserve user data privacy and security using WTLS, WTLS can only protect user data in transit over the wireless network between the WAP gateway and the client at the mobile terminal [14], while TLS is used to protect the user data in transit over the Internet between the gateway and the content server. The gateway, which translates messages from one protocol to another, is a security gray zone for end-to-end applications because the cleartext data is temporarily exposed in its memory during the conversion. Although the conversion happens in the memory of the gateway and is completed quickly, the concept of end-to-end security between the WAP client and the content server is violated. This is not acceptable for applications with strict security requirements, such as bank and financial transactions and e-business, because it is analogous to allowing your ISP to process (and inspect) the data of secure transactions. The only viable solution that strictly guarantees end-to-end security [8] is to collocate the gateway with the WAP server in a protected network that is secured from the Internet, for example, by a firewall. In [8], this alternative configuration was evaluated under various IS-95 wireless links and Internet channel conditions and compared against the standard configuration in which the gateway is located at junction of the cellular network and the Internet. Despite the feasibility of this alternative configuration, its drawbacks are also obvious. Aside from content providers having to invest in the infrastructure and to maintain their own gateways, the WAP clients also have to be configured to switch gateways to access various secure WAP applications. The latter, like having to switch ISPs when accessing different web sites, is cumbersome and undesirable for most users.

As WP-HTTP/WP-TCP are interoperable with HTTP/TCP, there is no complex protocol conversion required between WAP 2.0 and the Internet protocols; therefore the proxy is optional in WAP 2.0 configurations. Even in the presence of a proxy, strict end-to-end security can still be guaranteed by implementing the proxy at the transport layer, which enables it to support end-to-end TLS tunnels between the clients and the WAP servers [15].

Although WAP 2.0 is an evolutionary step forward with respect to WAP 1.x, if the data encoding mechanism employed by WAP 1.x at the gateway were not also employed in WAP 2.0, the transmitted packets in WAP 2.0 would be much larger than the encoded bytecode in WAP 1.x. To minimize the volume of data sent over the air, content coding of the HTTP message body may be employed by the HTTP client in the WAP terminal, and either at the HTTP server or in the WAP Proxy [4]. To support this function, the WAP Proxy must at least provide for deflate coding (data compression) as specified in [16]. Also, an encoding format known as wireless binary extensible markup language (WBXML), similar to WML in WAP 1.x, can be implemented at the WAP proxy [17]. However, these solutions only guarantee end-to-end security if a direct connection exists between the WAP server and WAP client to support TLS tunneling. If the content coding were performed at the WAP proxy instead, a similar end-to-end security problem as in WAP 1.x would still exist.

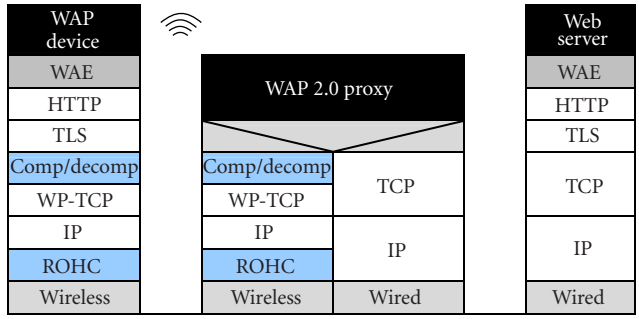


Figure 3: Data compression proxy supporting end-to-end security with TLS tunneling.

### 2.3. Proposed compression scheme with enhanced security

In order to improve the performance of WAP 2.0 while guaranteeing end-to-end security via TLS, a novel proxy architecture, as shown in Figure 3, is proposed. The proxy connects to the WAP server using standard TCP over the Internet and communicates with WAP clients using WP-TCP over the wireless domain to optimize transport layer performance. Thus end-to-end security can be strictly guaranteed by TLS tunneling. To further improve the performance, an advanced data compression scheme is introduced between the proxy and client to reduce the packet size and conserve bandwidth over the air interface. For applications that do not require end-to-end security, the proposed proxy can also work as a HTTP proxy that uses WP-HTTP between the proxy and client above the compression scheme to further improve the performance.

The proposed advanced data compression scheme combines two separate compression processes: TCP content compression and robust header compression (ROHC).

For evaluation purposes, content compression and decompression are accomplished using the deflate algorithm [16], a lossless compression method used in "gzip" that compresses data using a combination of the LZ77 algorithm and Huffman coding. Other lossless compression/decompression algorithms can also be used. The compression and decompression operate in the TCP socket stream using in-memory compression/decompression functions in the "zlib" compression library [18, 19]. Since the content compression works in the transport layer, it compresses all higher-layer headers, including the HTTP header. This compression works better than when only HTTP content compression is employed at WP-HTTP, and results in a maximum content compression for IP packets. There are three options for the content compression: no compression, reply compression, and request and reply compression.

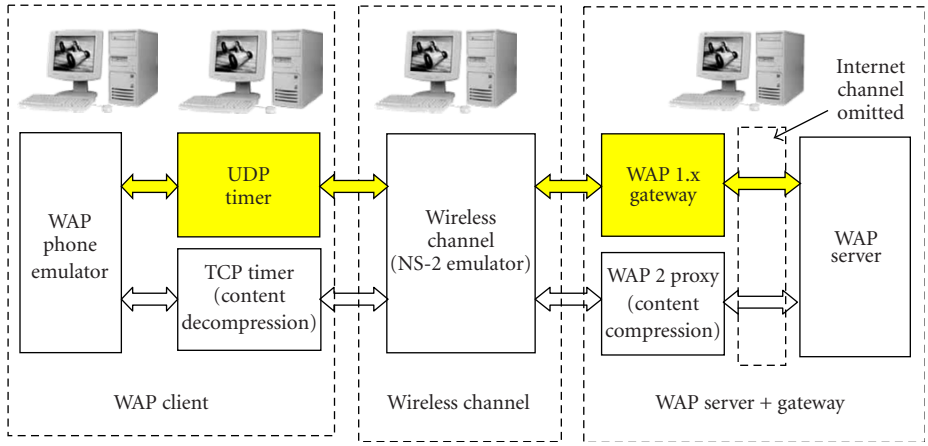


Figure 4: WAP 1.x and WAP 2.0 emulation test bed configuration.

ROHC has been studied extensively in the literature [20, 21, 22], and will be used in all 3G cellular systems, which can substantially improve spectrum efficiency and service quality for IP services such as voice and video over the mobile Internet. For evaluation purposes, ROHC is simulated by applying an appropriate compression ratio. The design and implementation of ROHC are discussed in [23].

## 3. PERFORMANCE EVALUATION METHOD

### 3.1. Test bed configuration

The performance of the proposed proxy architecture is evaluated using the test bed shown in Figure 4, for both versions of WAP (1.x and 2.0), with different combinations of compression methods. Both of the test configurations use a proxy to interconnect the wireless domain and Internet domain, and the Internet section is identical to both. Since different Internet delays resulting from various Internet conditions contribute the same amount of additional delays to both configurations, a differential comparison is more appropriate for performance comparison purposes. Consequently, the Internet domain is not included in the test bed as it is assumed to contribute the same delay to all the test scenarios. Assuming no delay and no packet losses over the Internet allows the performance comparison to focus on the effects of the wireless channels. So the test bed is configured with a WAP server, a WAP proxy (a WAP 1.x gateway for WAP 1.x, the proposed compression proxy for WAP 2.0), a WAP phone emulator as client, and an emulated wireless channel connecting the proxy to the client.

The network simulator 2 (ns-2) [24] is used to emulate the packet level behaviors of over narrowband IS-95 and wideband CDMA2000 1xRTT wireless channels. The emulated wireless channel consists of two nodes, a base station and a mobile client. By attaching the tap agents, the nodes are capable of introducing live traffic into the ns-2 simulator and injecting traffic from the simulator into the live network after the traffic has been subject to appropriate delays and losses. Due to the header added in ns-2 emulation, the Ethernet maximum transmission unit (MTU) is reduced from 1500 bytes to 1400 bytes. The IP packets captured from the live traffic by the node are first fragmented at link layer (LL) and then transmitted to the other node. The received fragments are then defragmented in the node and injected back to live traffic. Each fragment is sent every 20 milliseconds with 168 bits and 3048 bits, respectively, in accordance with the IS-95 and CDMA2000 1xRTT standards. With additional CRC and encoding tail bits, the maximum user data transmission rate in the emulation channel is 9.6 Kbps for IS-95, and 153.6 Kbps for CDMA2000 1xRTT. The wireless channel delay is set to be 1 millisecond, and the TCP options are set based on the mandatory WP-TCP requirements on both the client and the proxy server, for example, window scale option, large initial window, and selective acknowledgement are all supported, and the maximum congestion window size is set as 64 KB. The packet group size for class 2 WTP is set with the default value of 1405 bytes [25]. Since all practical parameter values and typical WML pages from real example sites are used in the simulations, the results closely represent the real-life WAP user experience.

For data services in both the IS-95 and the CDMA2000 networks, data is framed into 20 milliseconds blocks for transmission over the physical layer traffic channel [26, 27, 28]. Therefore, the frame error rate (FER) or block error rate (BLER) are more suitable measures of the link quality as seen by the upper layers than BER, since the use of interleaving and forward error correction coding techniques can lead to the detection and recovery of some bit errors. Several papers have used first-order Markov chains to model block error processes in transmissions over wireless channels [29, 30, 31]. In certain sets of parameters, the Markov chain leads to a unique stationary distribution, which means a uniform FER over time. Therefore, a specific FER is employed as a measure of the transmission quality in the experiments to evaluate performance under each given set of wireless channel conditions. The FER parameter represents the unrecoverable error rate after the FEC decoder. The frame is considered erroneous and needs to be retransmitted when

error occurs that the FEC decoder fails to correct. A selective-repeat (SR) automatic-repeat request (ARQ) error recovery mechanism is employed at the LL for the LL fragments. This provides a reliable connection between the compression and decompression processes such that loss of synchronization due to lost packets is not an issue here. ROHC over the wireless network is simulated by giving the first LL fragment a bigger size than the others. This assumes an average header-compression ratio that is statistically stationary and fixed in a long run.

### 3.2. Performance evaluation criteria

The performance metric considered is the average end-to-end access time or round-trip delay for a sample WML file [32]. WML files are used in our test since we want to compare the WAP 2.0 configurations with WAP 1.x protocol stack. While WAP 2.0 continues its support for WAP 1.x-based WML, the WAP 1.x stack does not recognize the new wireless application environment (WAE) definitions in WAP 2.0, such as extensible HTML (XHTML). In the experiments, several WML files were transferred and the average round-trip delay was obtained. The actual access time (AT) includes the wireless transmission time (WTT, including LL retransmissions), Internet transmission time (ITT, including retransmissions if applicable), and the system processing delay (PD); that is,

$$\mathrm{AT} = \mathrm{WTT} + \mathrm{ITT} + \mathrm{PD}, \tag{1}$$

where PD consists of the queuing delay (QD) at the WAP server and proxy and the processing time (PT) at the server, proxy, and client, given by

$$\begin{aligned}\mathrm{PD} = {} & \mathrm{QD}_{\mathrm{Server}} + \mathrm{QD}_{\mathrm{Proxy}} + \mathrm{PT}_{\mathrm{Server}} \\ & + \mathrm{PT}_{\mathrm{Proxy}} + \mathrm{PT}_{\mathrm{Client}}.\end{aligned} \tag{2}$$

In order to evaluate the performance improvements due to the data compression scheme, differential comparisons are used and the access time differences (ATDiff) is measured for evaluation purposes instead of the absolute AT values.

For WAP 2.0, the AT is the elapsed time between when the client makes a request and when it successfully receives (and decompresses if necessary) a reply at the TCP socket layer. For WAP 1.x, the sessions are based on class 2 WTP transactions, which is the basic request/response and the most commonly used transaction service [25]. It is connection oriented with a reliable invoke message with one reliable result message. WTP is over UDP in our experiments. In this case, the AT is the elapsed time between the invoke and the acknowledgment (ACK), both at the client side. In all cases, ITT is a common element of AT that offsets each other in computing ATDiff.

To facilitate the evaluation, wireless access time (WAT) is defined as AT less ITT, or the sum of the wireless transmission time and the processing delay, that is,

$$\mathrm{WAT} = \mathrm{WTT} + \mathrm{PD}, \tag{3}$$

the WAT of WAP 2.0 configuration without compression is used as the basis for comparison purposes. All other configurations are evaluated by comparing the ATDiffs, which are the WATs of other configurations minus the WAT of uncompressed WAP 2.0.

$$\begin{aligned}\mathrm{ATDiff} &= \mathrm{AT}_{\mathrm{otherconf}} - \mathrm{AT}_{\mathrm{nocomp\,WAP\,2}} \\ &= \mathrm{WAT}_{\mathrm{otherconf}} - \mathrm{WAT}_{\mathrm{nocomp\,WAP\,2}}.\end{aligned} \tag{4}$$

### 3.3. Assumptions and limitations

The wireless channel implemented in ns-2 closely simulates the link layer behavior of IS-95 and CDMA2000 1xRTT with a specific FER. However, due to the constraints of our test bed configuration and wireless channel emulation, our experiments are subject to certain assumptions and limitations. They are summarized as follows.

(i) The emulated wireless channel is used solely by the WAP application during the experiments. There could be other applications sharing the channel in real life. Extra delay would be incurred if the channel was shared with other traffic streams.

(ii) It is assumed that only one WAP session is in progress at any given time; that is, no new WAP request is generated until the result from the former request is received. This results in no queuing delays at the WAP server or the WAP proxy.

(iii) A fixed link layer FER is assumed on both the uplink and downlink. In real life, the traffic and propagation conditions in the CDMA channel may cause fluctuations in noise and interference levels and hence the FER, and the uplink and downlink may have different FERs.

(iv) The content compression and decompression are implemented on Pentium PCs. The processing time could be lower in real gateways employing more powerful processors, and higher in mobile terminals with less powerful processors.

(v) Not all optimizations suggested by WP-TCP are implemented due to the constraints of the test bed environment. Handoff delays have not been considered.

## 4. PERFORMANCE RESULTS

### 4.1. WAP enhancement with compression scheme

The performance is evaluated by comparing the ATDiffs of different compression options under various wireless channel conditions. WAP 1.x was also tested as a comparison and as an indication of the performance of WAP binary XML (WBXML) in WAP 2.0 since WBXML employs similar encoding and decoding method as binary WML. The WAT of WAP 2.0 without compression ($\mathrm{WAT}_{\mathrm{nocomp\,WAP\,2}}$) for both IS-95 and CDMA2000 1xRTT wireless channels, which will be used as the basis for further comparisons, are shown in Figure 5.

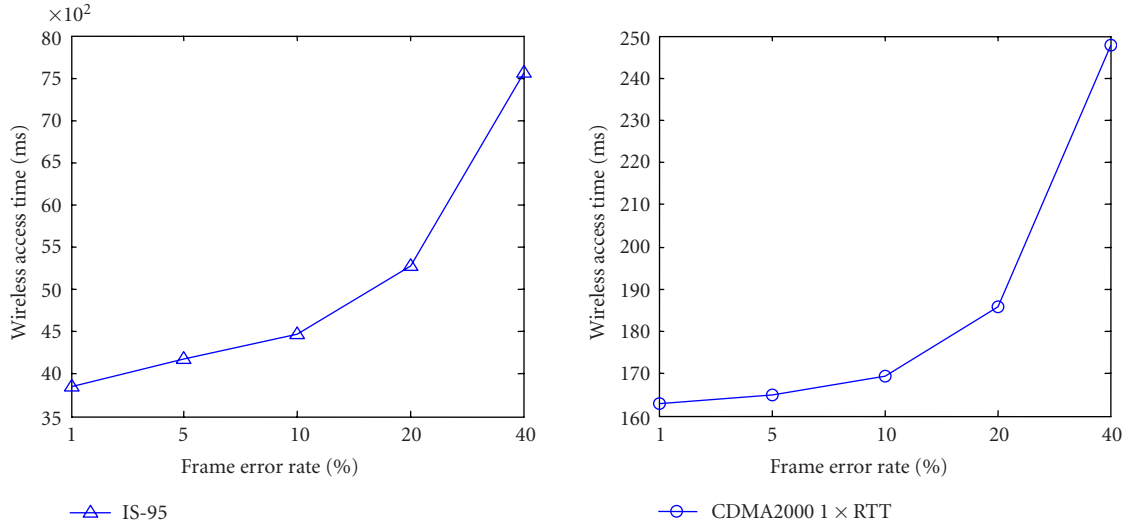


Figure 5: Wireless access time of WAP 2.0 without compression.

Table 1: WAP processing delays.

| WAP 1.x | WAP 2.0 (HTTP, TCP/IP) | | |
|---|---|---|---|
| | No compression | Reply compression | Request & reply compression |
| 180.05 ms | 4.02 ms | 6.69 ms | 11.51 ms |

The transmission times over the LAN interconnecting the test bed computers were measured and the processing delays (PD) (Table 1) were estimated by subtracting the transmission times from the total delays. The PD in WAP 1.x comes mainly from the protocol conversion and data encoding and decoding at the gateway and client. Result shows that in good conditions the PD of WAP 1.x is much larger than that of WAP 2.0, which implies that WAP 2.0 HTTP/TCP/IP stack is more efficient. The compression process only causes a processing delay of several milliseconds.

Since IP is supported in IS-95 but not in GSM and most other narrowband network bearers on which WAP 1.x protocol stack employing WTP and WDP has to be used, WAP 2.0 with TCP/IP support is unlikely to be used in narrowband networks. We briefly compare the results in IS-95 as an indication of the effectiveness of WAP 2.0 data compression scheme for IP enabled narrowband wireless networks and focus our results on the CDMA2000 1xRTT wireless channel.

Results for an emulated IS-95 channel with a maximum bandwidth of 9.6 Kbps presented in Figure 6 show that the WAT of WAP 1.x is 2.72–6.04 seconds less than that of WAP 2.0 employing TCP/IP when no data compression is used, which corresponds to 71%–80% less than $\text{WAT}_{\text{nocomp WAP 2}}$ from Figure 5. This clearly shows the advantage of WAP 1.x over TCP/IP in low-bandwidth networks. Figure 6 shows that by applying the proposed advanced data compression scheme, the performance of WAP 2.0 can be improved to match that of WAP 1.x.

Over an emulated CDMA2000 1xRTT channel with maximum bandwidth of 153.6 Kbps, WAP 2.0 outperforms WAP 1.x even if no compression is applied, with WAT reduced by 78.4 milliseconds and 25.7 milliseconds at 1% and 40% FER, respectively (Figure 7), corresponding to 32.5% and 9% improvement on WAT compared with $\text{WAT}_{\text{nocomp WAP 2}}$. This is due to the long processing delay for protocol conversions in the WAP 1.x gateway and client. The lower processing time of HTTP/TCP makes them more appropriate for high-speed networks. Since the transmitted data traffic is much higher than that in WAP 1.x, the WAP 2.0 performance degrades when FER is high due to more packet retransmissions.

The content compression brings the most performance enhancements by reducing the transmission delays by 73.6–117 milliseconds or 44%–46% less than $\text{WAT}_{\text{nocomp WAP 2}}$ at 1% and 40% FER (Figure 8) because text-based WML (or XHTML) files yield high compress ratios. The compressed packets need much fewer LL fragments to transfer. When ROHC is employed, another 3–7 milliseconds or 3% reduction in WAT can be achieved over content compression. Results also show that reply compression works even better than the combined request and reply compression. This can be explained as follows: because the request packet is quite small, therefore the data compression scheme does not give much gain, and the transmission time saved from the reduced size is smaller than the processing delay introduced by the request compression. The results show that WAP 2.0 is more suitable for the high-speed wireless networks, and the compression scheme can reduce WAT by over 76 milliseconds at 1% and 120 milliseconds at 40% FER, corresponding to over 45% improvement in WAT, but request compression is not appropriate for use over a high-speed wireless network.

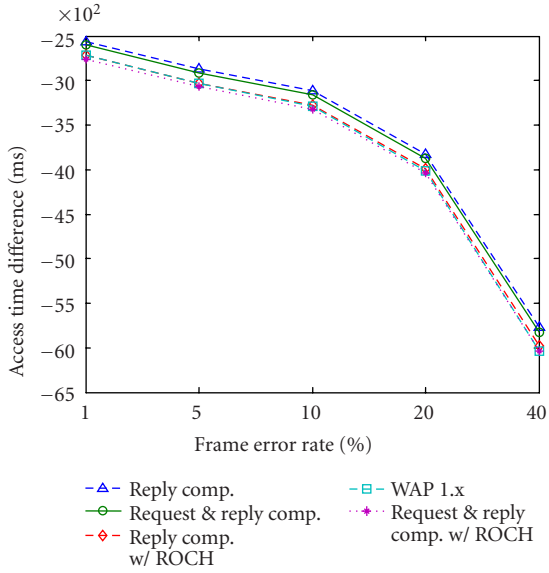


Figure 6: WAP performance in IS-95.

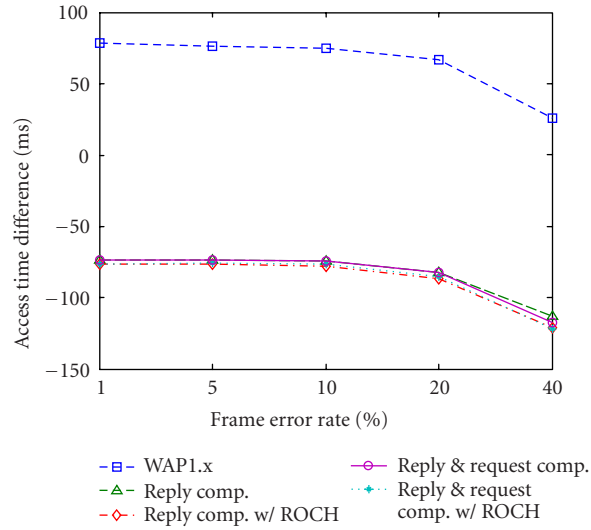


Figure 7: WAP performance in CDMA2000 1xRTT.

The above results are obtained based on data compressions and protocol conversions at the processing speeds of the test bed computers. With a less powerful mobile terminal processor, there will be some extra processing delay for both protocol conversion and content encoding in WAP 1.x and the proposed data compression and decompression process in WAP 2.0. Considering the complexity of WAP 1.x protocol conversion, it is reasonable to assume that this has a higher processing delay than the WAP 2.0 content compression and decompression process. Furthermore, in WAP 2.0, the extra processing delay for content compression and decompression is generally much smaller compared with the reduction in transmission time made possible by content compression. Therefore, although the numerical results are specific to the test bed equipment, our general observations regarding the effectiveness of the proposed proxy architecture supporting data compression, based on the experimental results, remain valid.

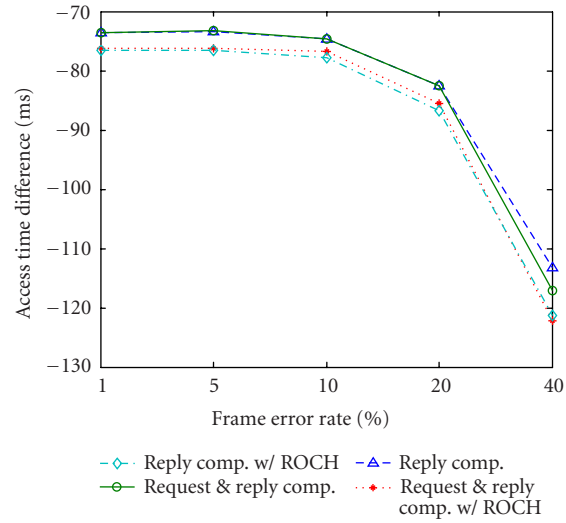


Figure 8: The performance of WAP 2.0 with compression in CDMA2000 1xRTT.

### 4.2. *WAP 2.0: proxy versus direct connection*

In WAP 2.0, a direct TCP connection can be used to provide an end-to-end HTTP/1.1 service. However, using a proxy, which leads to a split-TCP approach [33], can optimize and enhance the connection between the wireless domain and the Internet domain.

In the case of a direct connection, the optimizations provided by WP-TCP and WP-HTTP over wireless links may not be available as the wireless profiled options for the respective protocols may not be implemented at the servers. In the mobile networks with bursty errors and high bit error rates, relatively long delays and variable bandwidth, the congestion control mechanism of standard TCP adversely affects its performance, for example, packet error is regarded as congestion, which leads to reduction of congestion window and slow recovery. In addition, two major factors also contribute to the increase in the access time.

First, the split-TCP approach using a proxy shields problems associated with wireless links from the wireline Internet and vice versa. The direct connection causes error propagation between the Internet and wireless domains. It can be proved by a simple calculation. Let the packet drop rates and transmission times over the Internet and wireless domain be $\varepsilon_1$, $t_1$ and $\varepsilon_2$, $t_2$, respectively, in the forward direction and assume perfect feedback with no packet drops in the reverse direction. In a direct connection, the overall access time (AT) is the process delay (PD) plus direct transmission time ($t_{\text{direct}}$):

$$\text{AT}_{\text{direct}} = \text{PD} + t_{\text{direct}} = \text{PD} + \frac{t_1 + t_2}{(1-\varepsilon_1)(1-\varepsilon_2)}. \quad (5)$$

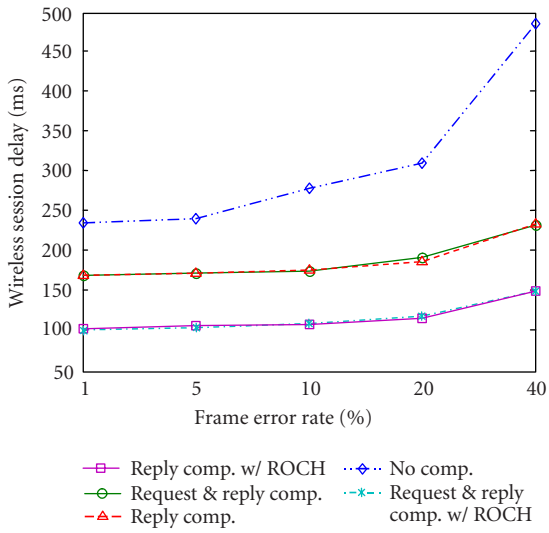


Figure 9: Direct connection wireless session delay in IS-95.

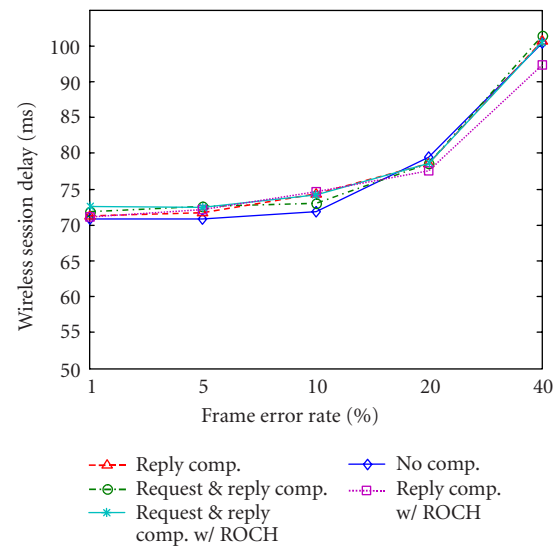


Figure 10: Direct connection wireless session delay in CDMA2000 1xRTT.

If a proxy is present, the AT is given by

$$\begin{aligned} \mathrm{AT}_{\mathrm{proxy}} &= \mathrm{PD} + \mathrm{ITT} + \mathrm{WTT} \\ &= \mathrm{PD} + \frac{t_1}{1-\varepsilon_1} + \frac{t_2}{1-\varepsilon_2} \\ &= \mathrm{PD} + \frac{t_1 + t_2 - (\varepsilon_1 t_2 + \varepsilon_2 t_1)}{(1-\varepsilon_1)(1-\varepsilon_2)}. \end{aligned} \tag{6}$$

It is obvious that $\mathrm{AT}_{\mathrm{proxy}} < \mathrm{AT}_{\mathrm{direct}}$. The proxy facilitates independent error recovery over the wireless and Internet domains so that error-free data is always passed from one domain to the next. Thus no retransmission needs to pass through both domains.

Second, TCP is a reliable, connection-oriented transport layer protocol. For each TCP session, there is a 3-way handshaking for the TCP connection establishment and 4-way data exchange for the TCP connection termination process. If there is no proxy, the WAP client has to establish a separate TCP session for each different WAP server. With a proxy, the WAP client can maintain a long-lived socket with the proxy, thus eliminating extra connection and termination delays in the wireless domain. The wireless session delay (WSD) is used to represent these delays in our experiments. The WSD is defined as the time delay due to TCP connection establishment and termination in the wireless networks.

Over the low-bandwidth IS-95 channel supporting TCP/IP, experiment results in Figure 9 show that the WSDs are quite high if direct connections are used, 234 milliseconds at 1% FER and 483 milliseconds at 40% FER, respectively. If content compression is employed, the WSD is reduced by 28% at 1% FER and 52% at 40% FER. Figure 9 also indicates that ROHC in the wireless domain can give a 40% reduction in WSD over the content compression scheme due to the reduced header size in the handshaking packets.

With the WAP 2.0 TCP/IP stack, in CDMA2000 1xRTT wireless channels, WSDs are around 70 milliseconds at 1% FER and 95 milliseconds at 40% FER (Figure 10). WSDs are almost the same with or without the data compression scheme. Also, ROHC gives nearly no benefit in reducing the WSD. This is because the TCP handshaking and termination packets are quite small, and they can be transmitted in one LL fragment in all cases.

Note that the WSDs presented above are obtained over an error-free Internet in our test environment. In practice, packets may be dropped over the Internet due to congestion, in which case the WSD of direct connections will become even higher due to possible retransmissions of handshaking packets caused by the Internet and wireless losses. These results clearly illustrate the performance enhancements provided by the proxy made possible by setting up long-lived connections, especially when the clients frequently switch applications hosted on different WAP servers.

Since a proxy is usually maintained by a wireless service provider, beside the above-mentioned advantages, a proxy is required for WAP Push operations and may offer location, privacy, and presence-based services to mobile users. Furthermore, the caching capability at the proxy can provides better service experience to end users, especially for low-end WAP phones.

## 5. CONCLUSIONS

We have presented a novel proxy architecture employing advanced data compression schemes to minimize air-interface traffic thus significantly improving the access time performance of WAP 2.0, while ensuring that end-to-end security can be strictly guaranteed using TLS tunneling. Most of the access time reduction is contributed by the reply content compression, while ROHC can offer further improvements. Experimental results show that WAP 1.x is optimized for narrowband networks. However, in narrowband IS-95 networks with IP support, the proposed scheme can reduce

WAP 2.0 access time to the same level as WAP 1.x. In wideband CDMA2000 1xRTT networks, WAP 2.0 outperforms WAP 1.x in access time even without data compression, and the advanced compression schemes can reduce access time by 75–120 milliseconds in the test bed network, corresponding to over 45% improvement on WAT. Although optional in WAP 2.0, the proxy not only prevents the error propagations between wired and wireless domains, but also significantly reduces the wireless session delays due to TCP connection establishments by enabling long-lived connections to be set up between the proxy and wireless terminals. With the deployment of IP-enabled high-speed 2.5*G* and 3*G* networks, WAP 2.0 will facilitate further convergence between wireless networks and the Internet, and the proposed data compression scheme can bring huge performance benefits.

## ACKNOWLEDGMENTS

This paper is based in part on a paper presented at IEEE WCNC, New Orleans, Louisiana, March 2003. This work was supported by grants from TELUS Mobility and the Advanced Systems Institute of British Columbia, and by the Canadian Natural Sciences and Engineering Research Council under Grant no. CRD 247855-01.

## REFERENCES

[1] WAP Forum, "Wap architecture specification," version 12-July-2001, http://www.wapforum.org/what/technical.htm.

[2] WAP Forum, "Wap 2.0 technical white paper," January 2002.

[3] WAP Forum, "Wireless profiled TCP," version 31-March-2001.

[4] WAP Forum, "Wireless profiled HTTP," version 29-March-2001.

[5] P. Stuckmann and C. Hoymann, "Performance evaluation of WAP-based applications over GPRS," in *Proc. IEEE International Conference on Communications (ICC '02)*, vol. 5, pp. 3356–3360, New York, NY, USA, May 2002.

[6] A. Andreadis, G. Benelli, G. Giambene, and B. Marzucchi, "Performance analysis of the WAP protocol over GSM-SMS," in *Proc. IEEE International Conference on Communications (ICC '01)*, vol. 2, pp. 467–471, Helsinki, Finland, June 2001.

[7] S. Lee and N.-O. Song, "Experimental WAP (wireless application protocol) traffic modeling on CDMA based mobile wireless network," in *Proc. IEEE VTS 54th Vehicular Technology Conference (VTC '01)*, vol. 4, pp. 2206–2210, Atlantic City, NJ, USA, October 2001.

[8] S. Sheoran and V. C. M. Leung, "Evaluation of WAP network configuration supporting enhanced security," in *Proc. International Conference on Consumer Electronics (ICCE '02)*, pp. 78–79, Los Angeles, Calif, USA, June 2002.

[9] P. Ashley, H. Hinton, and M. Vandenwauver, "Wired versus wireless security: the internet, WAP and iMode for E-commerce," in *Proc. 17th Annual Computer Security Applications Conference (ACSAC '01)*, pp. 296–306, New Orleans, La, USA, December 2001.

[10] E.-K. Kwon, Y.-G. Cho, and K.-J. Chae, "Integrated transport layer security: end-to-end security model between WTLS and TLS," in *Proc. 15th International Conference on Information Networking (ICOIN '01)*, pp. 65–71, Beppu City, Oita, Japan, January 2001.

[11] ISO/IEC JTC1/SC29/WG11 N4668, "Overview of the MPEG-4 standard," March 2002.

[12] ITU-T Recommendation H.263, "Video coding for low bit rate communication," February 1998.

[13] ITU-T Recommendation V.42bis, "Data compression procedures for data circuit-terminating equipment (DCE) using error correction procedures," January 1990.

[14] WAP Forum, "Wap transport layer end-to-end security," version 28-June-2001.

[15] WAP Forum, "Wap TLS profile and tunneling," version 11-April-2001.

[16] P. Deutsch, "Deflate compressed data format specification," version 1.3, May 1996, RFC1950.

[17] WAP Forum, "Binary XML Content Format Specification," Version 1.3, July 25, 2001.

[18] P. Deutsch, "ZLIB compressed data format specification," version 3.3, May 1996, RFC1950.

[19] Zlib version 1.1.4, http://www.gzip.org/zlib/, March 2002.

[20] C. Bormann, C. Burmeister, M. Degemark, et al., "Robust header compression (ROHC)," July 2001, RFC3095.

[21] M. Degermark, "Requirements for robust IP/UDP/RTP header compression," June 2001, RFC3096.

[22] M. Degermark, H. Hannu, L. Jonsson, and K. Svanbro, "Evaluation of CRTP performance over cellular radio links," *IEEE Pers. Commun.*, vol. 7, no. 4, pp. 20–25, 2000.

[23] L.-E. Jonsson and P. Kremer, "ROHC implementer's guide," IETF Internet draft, <draft-ietf-rohc-rtp-impl-guide-04.txt>, September 23, 2003.

[24] The network simulator ns 2, "version 2.1b9," http://www.isi.edu/nsnam/ns.

[25] WAP Forum, "Wireless Transaction protocol," version 10-July-2001.

[26] TIA/EIA Interim Standard-95, "Mobile station—base station compatibility standard for dual-mode wideband spread spectrum cellular system," July 1993.

[27] 3GPP2 C.S0002-C, "Physical layer standard for cdma2000 spread spectrum systems, Release C," Version 1.0, May 28, 2002.

[28] 3GPP2 C.S0003-C, "Medium Access Control (MAC) Standard for cdma2000 Spread Spectrum Systems, Release C," Version 1.0, May 28, 2002.

[29] C. C. Tan and N. C. Beaulieu, "On first-order Markov modeling for the Rayleigh fading channel," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 2032–2040, 2000.

[30] M. R. Hueda, "On first-order Markov modeling for block errors on fading channels," in *Proc. IEEE 55th Vehicular Technology Conference (VTC '02)*, vol. 3, pp. 1336–1339, Birmingham, Ala, USA, May 2002.

[31] M. R. Hueda, "On the Markovian approximation for block-errors in DS-CDMA transmissions over slow fading channels with multicarrier transmit diversity," in *IEEE International Conference on Communications (ICC '02)*, vol. 2, pp. 737–741, New York, NY, USA, May 2002.

[32] B. Eged, T. Dezso, and F. Egedi, "Server side round-trip delay measurements in WAP environments," in *Proc. 18th IEEE Instrumentation and Measurement Technology Conference (IMTC '01)*, vol. 1, pp. 525–529, Budapest, Hungary, May 2001.

[33] A. V. Bakre and B. R. Badrinath, "Handoff and systems support for indirect TCP/IP," in *Proc. 2nd Symposium on Mobile and Location-Independent Computing*, pp. 11–24, Ann Arbor, Mich, USA, April 1995.

**Zhanping Yin** received his B.Eng. and M.Eng. degrees in optical instrument from Tianjin University, Tianjin, China, and the M.A.Sc. degree in electrical engineering from the University of British Columbia, Vancouver, Canada, in 1992, 1995, and 2002, respectively. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada. His current research interests are in wireless communications protocols including WAP, WLAN, WPAN, UWB, and cross-layer design.

**Victor C. M. Leung** received the B.A.Sc. (with honors) and Ph.D. degrees, both in electrical engineering, from the University of British Columbia (UBC) in 1977 and 1981, respectively. He received the APEBC Gold Medal as the Head of the Graduating Class in the Faculty of Applied Science, and the Natural Sciences and Engineering Research Council Postgraduate Scholarship. From 1981 to 1987, Dr. Leung was a Senior Member of Technical Staff at MPR Teltech Ltd. In 1988, he was a Lecturer in the Department of Electronics, the Chinese University of Hong Kong. He returned to UBC in 1989 as a faculty member, where he is a Professor in the Department of Electrical and Computer Engineering, holder of the TELUS Mobility Industrial Research Chair in Advanced Telecommunications Engineering, and Associate Head for Graduate Affairs. His research interests are in the areas of architectural and protocol design and performance analysis for computer and telecommunication networks, with applications in satellite, mobile, personal communications, and high-speed networks. Dr. Leung is a Fellow of IEEE, a Member of ACM, an Editor of the IEEE Transactions on Wireless Communications, and an Associate Editor of the IEEE Transactions on Vehicular Technology.