IEEE
COMMUNICATIONS
SURVEYS
The Electronic Magazine of
Original Peer-Reviewed Survey Articles

ENHANCING WIRELESS INTERNET PERFORMANCE

FEI HU, ROCHESTER INSTITUTE OF TECHNOLOGY
NEERAJ K. SHARMA, CLARKSON UNIVERSITY

ABSTRACT

This article surveys wireless Internet technologies whose goals are to enhance networking performance. These technologies are organized into seven categories: power saving, mobile performance, Multimedia Quality-of-Service, application performance, transport-layer characteristics, data-link layer, and non-TCP options. For each category, the main technical characteristics are outlined, the architectural aspects are discussed, and the advantages and disadvantages are analyzed. The objective of this article is to contribute to the overall understanding of the technologies available for constructing the forthcoming wireless Internet infrastructure.

The convergence of wireless access and the Internet will be the next wave in the information industry [1]. Currently, network providers try to offer services over wireless Internet infrastructures such as GPRS [2]. However, numerous studies remain to be carried out on the enhancement of wireless Internet performance. Some major obstacles, such as the high error rate of wireless links, frequent handoff, limited power, long delay, and limited wireless bandwidth, could inhibit its widespread use. Although some papers have introduced some enhancement schemes, they did not cover the *multiple* network protocol layers and focused on a *specific* aspect of wireless Internet, such as Transport Control Protocol (TCP) modifications for wireless communications. This article will survey various proposed approaches to mitigate the effect of the abovementioned obstacles. Our discussion of wireless Internet enhancements is based on the classification as shown in Fig. 1. Broadly speaking, we can classify the enhancements into seven categories (refer to the tan shaded rectangles in Fig. 1):

Saving Power: In the past, energy efficiency research centered around the *physical layer,* only because the power consumption in a mobile computer was a direct result of the hardware. Now people have realized that the goal of saving power should be implemented in *multiple* wireless Internet protocol layers instead of the physical layer only.

Improving Mobile Performance: In a wireless Local Area Network (LAN), mobility performance is not a major concern since a mobile host could visit the Internet through a wireless outlet and their mobility is limited to the local area. However, in GSM or other cellular networks, the handoff could occur frequently between the neighboring cells. Therefore, we need enhancement strategies for guaranteeing the smoothness of the handoff communications.

Guaranteeing Quality-of-Service (QoS): QoS requirements such as end-to-end delay and bandwidth-per-flow should be met in multimedia wireless networks. Networking experts proposed some enhancement mechanisms such as priority-queue and differentiated services for guaranteeing QoS.

Application-Layer Performance: The performance of the Internet *applications* such as FTP could be optimized to overcome the obstacles of wireless links. For example, local caching could save a large amount of wireless bandwidth in the Web applications.

TCP-Layer Performance: Although TCP has been greatly improved in its ability to adapt to high-speed links, many versions of TCP over wireless links still cannot keep the comparative throughput as can TCP in wired networks [3]. The main disadvantage is that traditional TCP assumes that all packet losses are due to network congestion. This assumption needs significant modification in wireless Internet applications because most packet losses are due to wireless link errors [4]. This article will survey typical TCP enhancements for wireless environments.

Data-Link Layer Performance: Wireless Internet could produce many symbol interferences, multi-path fading, and bit errors in the data-link layer. Forward error correction (FEC) and retransmission can be used to overcome the link errors.

Non-TCP Enhancements: This basically refers to the adoption of non-TCP protocols in the transport layer of the Internet in order to optimize the *specific* networking performance, such as in wireless LANs. However, it should provide a seamless interface for interoperation with the *regular* TCP standard.

It should be noted that these categories could be further classified. For example, application-layer enhancements could be based on a mediator (sometimes called *proxy*) or not. Some data-link-layer enhancements need to interact with the TCP layer, while others do not.

The remainder of this article will summarize the main mechanisms proposed in recent literature for improving wireless Internet performance. Each of the next seven sections will be devoted to each of the abovementioned categories. For

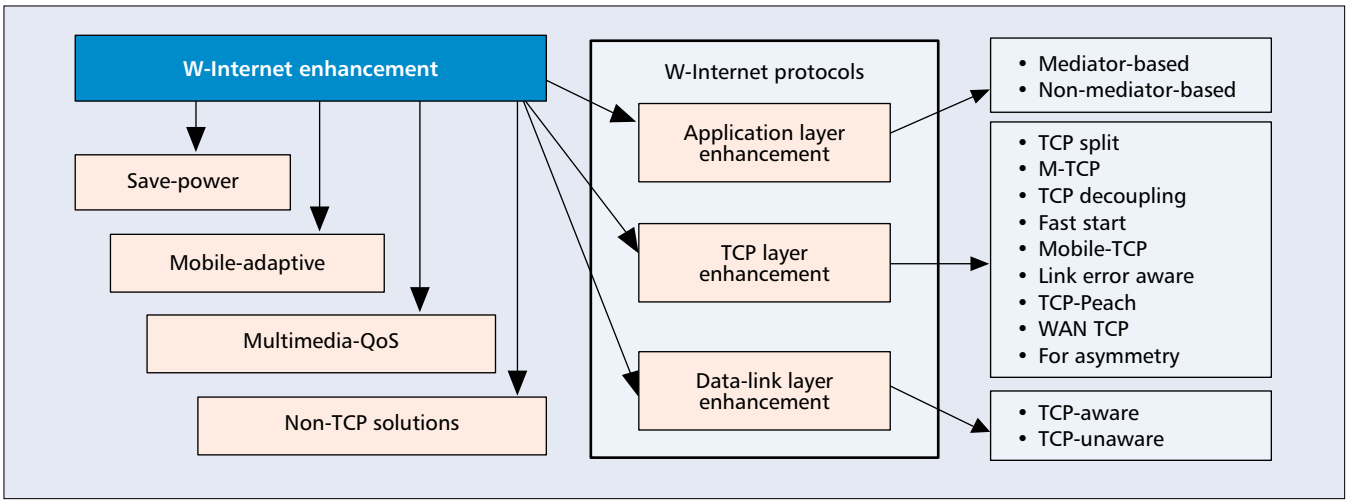

FIGURE 1. *Classification of wireless Internet enhancements.*

each category, the main technical characteristics are outlined, the architectural aspects are discussed, and the advantages and disadvantages are analyzed. We later list the proper scenario for each enhancement technology. Finally, the conclusions are draw.

APPLICATION-LAYER ENHANCEMENT (ALE)

In this section we discuss some typical enhancing technologies for adapting Internet applications to the wireless environment. We define each of these techniques individually, and we identify application environments where these techniques have been applied, and comment on the impact they have on performance. Finally, we summarize our findings in a table.

TECHNIQUE 1: PROTOCOL REDUCTION

In the wireless Internet Application Layer, traditional HTTP has inefficiencies such as large connection overhead, redundant transmission capabilities, and verbose protocol setup. To save the limited wireless bandwidth, we can enhance the traditional HTTP through the following two approaches:

- Simplify the protocol headers. HTTP *request headers* contain lists of MIME content-types that can be hundreds of bytes in length. The server and the client could negotiate with each other on the reduction of the HTTP Request Headers.
- Reduce the transmission rounds. If Internet users are requesting the same content as the last request, the cached local information could be used.

TECHNIQUE 2: CACHING

Caching strategy is based on the Least Recently Used (LRU) algorithm and can be adopted on both the client and the server side. The caching objects in wireless environments should be consistent across each browser session and may have a digital signature for security. Caching could efficiently save wireless bandwidth.

New caches have to be reconstructed on the client side or the server side when the handoff occurs across the cell boundary. In [5], a *cache relocation* solution was proposed to enhance the roaming performance of the mobile hosts. Cache relocation assumes the existence of user *profiles* stored in specialized nodes within the user's *home network* (HN). When the mobile host moves into a new sub-network different from its HN, the registered profile database is queried and the relevant information is forwarded to the new sub-network. A path *prediction* algorithm is used to determine which cell the mobile user is likely to move to.

TECHNIQUE 3: DYNAMIC URL

Current Web infrastructure cannot accommodate mobile clients because it transmits the dynamic information *explicitly* through form-based interfaces that require the users' input in the client machine. In other words, it cannot support *automatically* updated documents.

To address this problem, *dynamic uniform resource locators* (D-URLs) are proposed in [6] to allow a single URL to return different documents corresponding to different mobile locations. For example, the following D-URL could be used to return the HTML document that describes the mobile user's current location (Fig. 2): http://own server/places/$(Location).html.

In Fig. 2, the actual document URL name is assigned to the D-URL variable [6].

TECHNIQUE 4: DIFFERENCING

The principle of differencing is to take advantage of the fact that different replies from the same program are usually very similar. For example, successive replies from a stock-quote server contain significant redundancy, such as graphic art. The differencing operation is typically implemented by inserting two components into the data path between the client and the server: client side intercept (CSI) and server side intercept (SSI). The SSI performs the differencing operation, the result being a *difference stream*, which represents the difference between the received report and the initially established CGI objects. Then, the *difference stream* is transmitted between CSI and SSI. Differencing techniques are implemented in some applications, such as IBM WebExpress [7].

TECHNIQUE 5: SPLIT PROXY

In the *split proxy* scheme, two peer *mediators* are introduced, i.e., the *agent* on the client side and the *proxy* on the server side. The goal of using *split proxy* is to hide the influences of high-error-rate wireless links from the visible Internet applications. Some enhancing techniques can be adopted on the *tunnel* between those two peer *mediators*. For example, *caching* can be carried out in the proxy *tunnel*. For each cached object, we can define a *lease* as a time period after which the server will assume that the client has terminated the connection

unless the client renews the *lease*. The *mediator* keeps the server alive by periodically renewing the leases [8]. In addition, *differencing* and *data compression* can also be performed in this *tunnel* in a transparent way.

Technique 6: Dynamic Documents

The dynamic documents scheme adopts a type of program written in the portable Tcl scripting language, which is then transmitted to the browser and interpreted there. The client may apply different *presentation policies* on a per-document basis. For example, if the network bandwidth is not sufficient, the system can delay the fetching of large multimedia resources or just execute the Tcl script in the background. *The dynamic documents* scheme needs help from the *interface processing policies* (IPP) [9]. A typical example of IPP is the processing of fill-out form contents in many wireless applications. This operation can be restricted to the client only instead of relaying the request to the server. Communication with the sever too frequently could largely consume wireless bandwidth.

Technique 7: Pipelining

Pipelining means the fetching of several files simultaneously over a single Internet connection. The benefit of adopting *pipelining* is to reduce the transmission latency and to allow the file transfer to take advantage of the available bandwidth. HTTP 1.1 with *pipelining* has been standardized and is now in use. A non-HTTP protocol based on *pipelining* technology was proposed in [10]. A useful extension made in *pipelining* is the ability to resume file transfers when a connection is broken.

Technique 8: GUI Enhancement

Due to the small size of handheld devices, developing interactive graphical user interface (GUI) Web applications for wireless transmission requires thoughtful planning. Here are some good hints for optimizing enterprise-wide HTML applications for small mobile display devices [11]:

- Choose only the most essential contents to display.
- Organize information effectively.
- Try to avoid graphics that need large amounts of transmission bandwidth.

In Table 1 we list the main techniques for enhancing wireless Internet application performance and the corresponding examples.

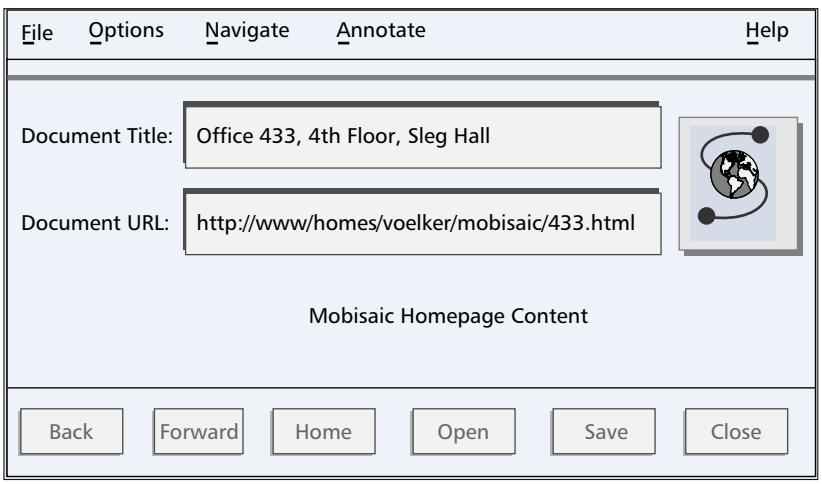

FIGURE 2. *Dynamic URLs [6].*

Transport-Layer Enhancement (TLE)

Wireless TCP, which is defined as the TCP version in a network scenario with wireless sections, can present many challenging issues, such as failing to respond to burst errors and entailing wasteful effort of retransmissions [14]. Current mechanisms to enhance *Wireless TCP* performance are based on the modification of the regular TCP algorithm that was originally designed for *wired* links.

Recommendations From IETF on Enhancing TCP performance Over Wireless Links

Recently the Performance Implications of Link Characteristics (PILC) group in IETF suggested some valuable TCP modifications for enhancing the end-to-end performance of wireless links (please refer to RFC 3155 [15] and Internet Draft [16]). The main conclusion from PILC was that avoiding congestion in wireless scenarios has to take precedence over quickly repairing transmission errors, since congestion affects all traffic while transmission loss affects only the specific traffic that encounters errors. Thus, a good TCP scheme for wireless scenarios should decrease the amount of time spent unnecessarily in the congestion avoidance phase.

TCP Enhancement for Satellite Networks

The characteristics of satellite channels, such as large latencies, path asymmetries, and occasionally high error rates, provide Satellite TCP operation with a challenging environment [17]. Satellite TCP performance enhancement is required mainly due to two problems:

- The regular TCP version needs a long time (generally more than three seconds) to reach the full recommended window size in satellite links, which can waste satellite bandwidth in the *slow-start phase* of TCP evolution [18]
- The *steady-state* behavior of regular TCP cannot fully utilize the bandwidth provided by T1 satellite channels [19]

Satellite Transport Protocol (STP) — The key to Satellite Transport Protocol (STP) [17] is *saving the reverse-satellite-channel bandwidth* to overcome the asymmetry of satellite networks. First, it suggests the use of a "split-TCP" scheme, i.e., the TCP connection is split at the gateway between the satellite network and the Internet. In addition, a new transport protocol is adopted over the satellite portion. The idea of split-TCP could shield high-latency satellite links from Internet applications. Second, STP adopts a different type of data acknowledgement compared to regular TCP: the transmitter periodically requests the receiver to acknowledge only successfully received packets, which is different from regular TCP, in which the receiver typically sends an ACK for every received TCP segment. Thus STP could save the reverse-channel bandwidth, which is beneficial for overcoming the asymmetry of satellite networks.

TCP-Peach — The key concept of TCP-Peach [20] is Sudden Slow-Start, an enhanced TCP algorithm for fully utilizing the wide satellite channel resource in the forward direction (i.e, from the Internet server to the client). TCP-Peach uses the dummy segments to probe the availability of network resources. The dummy segment does not carry valid data information to the receiver. Besides the traditional TCP algorithms, i.e., Congestion Avoidance and Fast Retransmit, TCP-Peach introduced two new algorithms: Sudden Start (SS) and Rapid Recovery (RR). The basic idea of SS is that the sender sets the congestion window size to 1

No.	Technique names	The advantages they have on enhancing wireless Internet performance	Application examples
1	Protocol reduction	Can save wireless bandwidth since the protocol contents transmitted over the wireless links are decreased.	[12]
2	Caching	Increase local browser response speed and save wireless bandwidth.	[5]
3	Dynamic URLs	Meet the location-mobility requirements.	[6]
4	Differencing	Decrease the transmission of requested server contents since only the difference between neighboring contents is sent.	[7]
5	Split proxy	Hide the harmful impacts of wireless links such as high-error-rate and limited bandwidth from the visible applications.	[13]
6	Dynamic documents	Allow the application of different presentation policies on a per document basis for adapting to mobility requirements.	[9]
7	Pipelining	Allow the fetching of several files simultaneously over a single connection and thus save wireless bandwidth.	[10]

Table 1. *Typical techniques for application-layer enhancements.*

in the beginning of a connection, but after the first data segment it will suddenly transmit (*MaxWin-1*) segments in one round trip time (*RTT*) [20]. Thus after one RTT, the congestion window size increases much more quickly than the regular TCP Slow-Start algorithm. *RR* is used to replace the Fast Recovery algorithm in the regular TCP version. As shown in Fig. 3, after the detection of a packet loss, the dummy segments can tell the sender the type of loss, i.e., whether it is a congestion loss or a link-error loss. If it is a wireless link-error loss, TCP-Peach will quickly recover its congestion window size to the threshold value, but a regular TCP version, such as TCP-Reno, can only increase one segment per RTT. Thus, TCP-Peach improves the satellite TCP throughput.

BA-TCP — The key technique of Bandwidth Aware TCP (BA-TCP) [21] is to guarantee fair bandwidth sharing among several competing connections with round trip times that may differ by more than two orders of magnitude. In BA-TCP, the IPv6 optional fields are used to convey delay and available bandwidth measurements to the TCP receivers in order to compute a proper advertised window size. In BA-TCP, a satellite connection can fairly share a bottleneck link with wired connections.

Indirect TCP (I-TCP)

The key concept of Indirect TCP (or I-TCP) is TCP-Split. I-TCP splits an end-to-end TCP connection between a fixed host and a mobile host into two separate connections:

- A wired TCP connection between the fixed host and the mobility support router (MSR) currently serving the mobile host.
- A wireless TCP connection between the MSR and the mobile host [22].

A wireless protocol is built for communicating between the mobile host and the MSR. If the MH switches cell areas during the lifetime of an I-TCP connection, a new MSR will be assigned to become the bridge between the wired connection and the wireless connection. The fixed host (FH) is completely unaware of the indirection and will not change its communication mode even when the MH switches cell areas.

M-TCP

The key technique of M-TCP [23] is to split TCP while guaranteeing the TCP end-to-end semantics. It was proposed to work well in the presence of frequent disconnection events and in low bit rate wireless links subject to dynamically changing bandwidth. When the M-TCP side at the mobile host (MH) is notified that the connection has been lost, it freezes all M-TCP timers (Fig. 4). This essentially ensures that disconnections do not cause the MH's M-TCP to invoke the congestion control algorithm.

TCP Decoupling

The key concept of TCP decoupling is to decouple TCP congestion control from TCP error control, and allow them to be performed separately and independently [24]. As shown in Fig. 5, the guaranteed minimum bandwidth (GMB) sender sends data packets to the TCP circuit. At the same time, a control TCP connection is set up to probe for the available bandwidth beyond the allocated GMB. By this decoupling scheme, data packets are transmitted as independent packets from header packets, and the data packet stream does not suffer from TCP's congestion control mechanism that is applied only to the header packets stream.

A big advantage of TCP decoupling is that a corrupted or lost data packet will not trigger the TCP congestion control algorithm, which could reduce the sending rate upon packet corruption. Because the error probability for these tiny header packets is much smaller than that of full-size packets carrying data payload, the probability of mistakenly triggering the TCP congestion control algorithm is significantly reduced. Thus it provides a reliable and high-throughput data transfer over the wireless network links while using TCP congestion control to avoid network congestion.

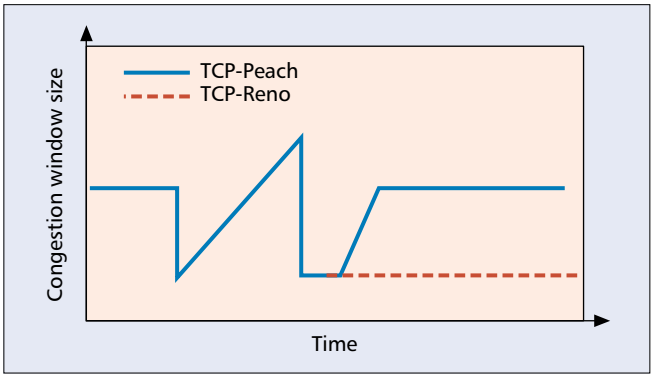

Figure 3. *The benefit of TCP-Peach over regular TCP-Reno [20].*

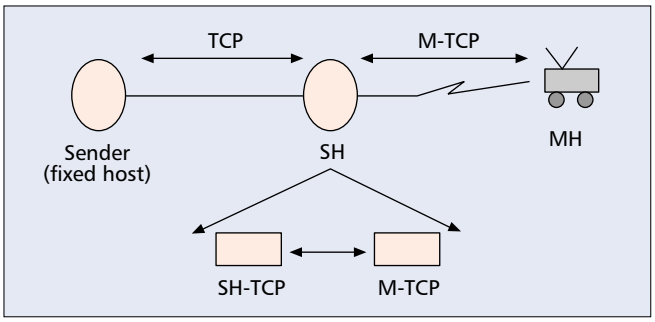

FIGURE 4. *Architecture of M-TCP [23].*

FAST START

The key technique of Fast Start [25] is to optimize the slow-start behavior of short TCP connections. A wealth of evidence suggests that most TCP connections today are short-sized, usually under 10 kb [26]. This means that the common assumption about TCP evolution, i.e., slow-start phase plus congestion avoidance phase, may not be valid [25]. As shown in Fig. 6, a short TCP connection often cannot enter congestion avoidance phase. It consists of unfinished slow-start[1] sections and a series of timeouts.

The slow-start performance for short TCP connections needs to be improved since regular slow-start behavior starts from window size 1 instead of a larger value (>1), which can waste the wireless resource. A typical proposal is stated in [27], called Fast-Start, which uses a double-queue model to derive the optimized initial TCP parameters such as threshold window size (ssthresh) and the starting window size of slow-start phase (cwnd). They deduced that the choice of cwnd is based on the file transfer size and the bandwidth-delay product of the wireless link. A technology called shift enhancement is used in [27] to adapt to the variable bandwidth-delay product of the wireless link. This idea is useful since actual wireless characteristics could vary frequently due to multi-path fading and symbol-interference.

[1] *This basically means that the slow-start phase of short TCP connections could not reach its maximum Window Size, which generally is a threshold value advertised by the receiver.*

LINK ERROR AWARE (LEA)

The idea of link error aware (LEA) is based on the observation that usually losses in the wired section are due to networking congestion, while losses in the wireless section are related to link errors. LEA can distinguish and handle efficiently these two types of losses, i.e., congestion loss and error loss [28]. In LEA, the base-station is responsible for generating and transmitting a special ICMP (Internet Control Message Protocol) message, called ICMP-DEFER, to the sender, when the first attempt at transmitting the packet on the wireless side is unsuccessful. This policy ensures that, within one round trip time, the TCP sender will receive either an acknowledgment for the packet or an ICMP message. A lack of both indicates a networking congestion loss instead of a wireless link error loss. A LEA scheme can ensure that end-to-end retransmissions do not start while link-layer retransmissions are going on.

WIRELESS WIDE–AREA–NETWORK TCP (WTCP)

The main goal of Wireless Wide-Area-Network TCP (WTCP) is to address the problem of rate control and reliability over commercial wide-area wireless networks [29]. WTCP performs rate control at the receiver, and uses inter-packet delays as the primary metric for rate control. WTCP uses the average per-packet separation to distinguish networking congestion losses from wireless error losses. This is similar to the approach in [30]. If it is a wireless error loss, the sender will not decrease its sending rate. Because the value of round-trip-time is large in the wireless wide area networks and many data transmissions are short-sized, WTCP attempts to compute the appropriate transmission rate for a connection immediately upon startup rather than going through the slow-start phase. Thus, WTCP can save wireless bandwidth.

TCP FOR ASYMMETRIC WIRELESS ENVIRONMENTS

In some wireless networks such as cable modem Internet, a faster downstream (>10Mb/s) and a slower upstream (<56kb/s) typically characterize a bandwidth asymmetry. Other asymmetric wireless cases include:

- Latency asymmetry (such as in satellite networks).
- Media access asymmetry (such as in IEEE 802.11-based wireless LAN).
- Packet error asymmetry (as in some cellular networks, the power from the base station to MH may be much larger than the reverse link, resulting in the uplink and

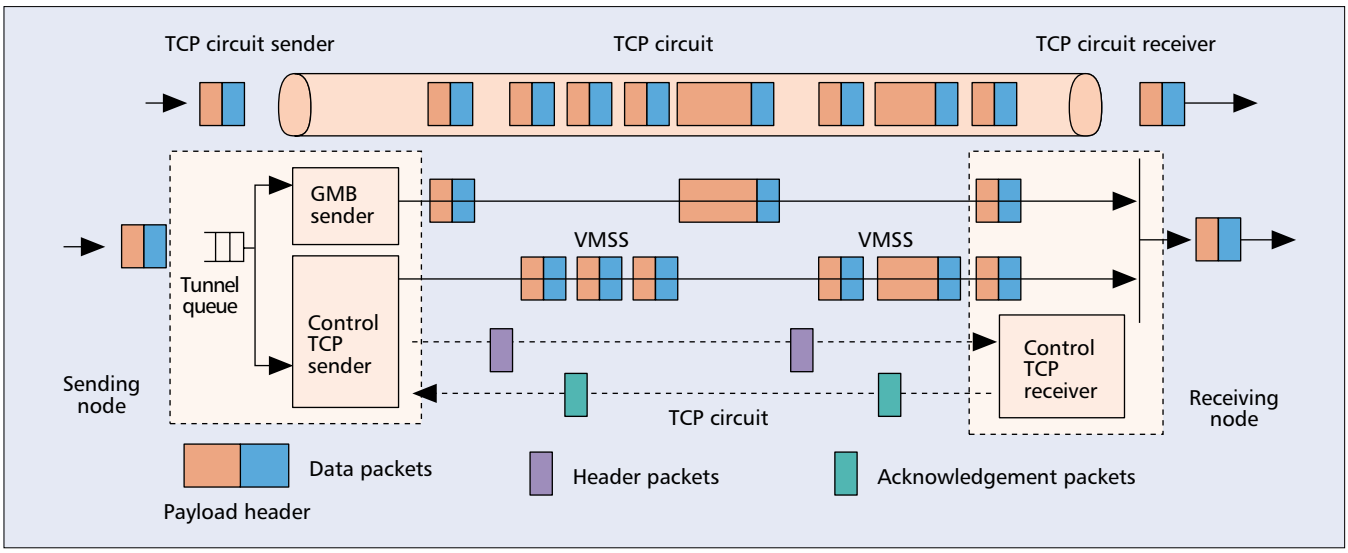

FIGURE 5. *TCP decoupling [24].*

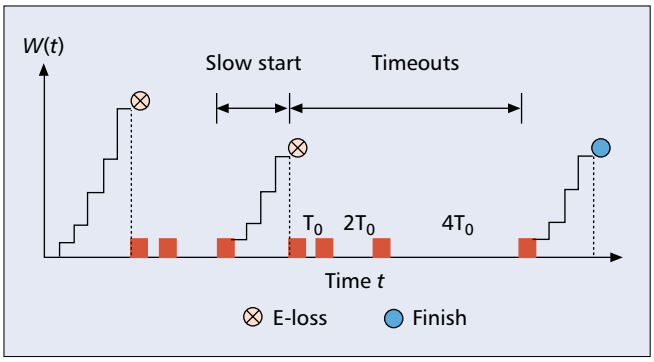

FIGURE 6. Short TCP *evolution over wireless links.*

downlink channels exhibiting different packet error rates) [31].

Good solutions to mitigate the effect of the asymmetry on TCP performance include TCP/IP header compression, ACK congestion control, ACK filtering [32], ACK prioritization [33], and other methods. Generally, the performance of TCP in the high bandwidth direction can be severely reduced due to the delay of ACK packets in the reverse link.

In order to overcome bandwidth asymmetry, a good solution called ACE (acknowledgment based on cwnd[2] estimation) was proposed in [34]. The idea of ACE is to let the number of packets per acknowledgment vary according to the sender's cwnd (congestion window-size). When cwnd is small, the number of packets per acknowledgment will be small. Thus, we can help speed up initial transfer and build up the cwnd. When cwnd is larger, the number of packets per acknowledgment is larger, and thus the number of acknowledgments sent on the narrow bandwidth link is reduced without much impact to the sender.

Delayed Duplicate Acknowledgments (DDA)

The key concept of delayed duplicate acknowledgment (DDA) [35] is to inactivate the TCP congestion avoidance algorithm by delaying transmission of duplicate acknowledgments. DDA is an extension of the SNOOP scheme proposed in [36]. However, unlike SNOOP, DDA does not need to look at the TCP headers, which makes DDA efficient for encryption. Like SNOOP, DDA also uses link-level retransmission. But DDA attempts to reduce the interference between TCP-layer retransmissions and link-level retransmissions by delaying the third and subsequent duplicate packets for an interval of d.[3] Specifically, when out-of-order packets are received, the TCP receiver responds to the first two consecutive out-of-order packets by sending duplicate packets immediately. By setting the proper value of d, DDA can imitate SNOOP in case of wireless error loss.

Comparison of Different TLEs

Within the above discussion, we did not mention the performance of different TCP versions such as TCP-SACK, TCP-Reno and TCP-Tahoe because these TCP versions are not designed for enhancing wireless performance. Interested readers may refer to [37], which provides a good analysis of the performance of these TCP versions in wireless LAN.

[2] *The size of congestion window (number of TCP packets).*

[3] *As we know, when more than three duplicate ACKs are received, the sender will assume that the corresponding packet is lost and will try to retransmit it. DDA attempts to hide this from TCP sender and thus avoids retransmission.*

In Table 2 we compare different wireless TCP proposals from the following nine factors:

- Adopting a proxy to split TCP into two parts: the wired part and the wireless part.
- Maintaining the end-to-end TCP semantics.
- Distinguishing error loss from congestion loss.
- Handling handoff disconnections.
- Optimizing the RTT (round trip time) delay to avoid long-period time-outs.
- Saving wireless bandwidth (e.g., fully utilizing wireless bandwidth through probing).
- Needing an extra type of packets for probing available bandwidth or identifying wireless loss.
- Modifying the existing TCP algorithm.
- Enhancing the performance of short-TCP connections.

Based on Table 2, we can see that only Indirect-TCP does not maintain the end-to-end semantics. Other methods, such as M-TCP, Mobile-TCP, and DDA, can all overcome this drawback, although they also adopt the split-TCP approach. Basically, they achieve this by hiding wireless error loss from the TCP sender and not blocking the regular TCP ACK flows.

Only M-TCP and Mobile-TCP take into consideration the serious effect from frequent handoff disconnections through the relocation of TCP connection information.

Only TCP-Peach and Fast Start consider the improvement of short-TCP performance. Generally, short-TCP may not reach the congestion avoidance phase and stay in the slow-start phase only. TCP-Peach attempts to probe the wireless bandwidth within one RTT, and suddenly adjusts to the maximum window size (called rapid recovery).

In Table 2, modifying regular TCP could include the following approaches:

- Changing initial slow-start window size (such as in Fast Start).
- Adding new phases (such as rapid recovery in TCP-Peach).
- Adopting rate-based flow control instead of window-based flow control (such as in TCP-decoupling).
- Adding new TCP operations (such as identifying ICMP message in LEA).

It is very useful to differentiate error loss from congestion loss since TCP does not need to decrease its window size when responding to error loss. Table 3 lists some common methods adopted by four wireless TCP schemes to distinguish those two types of losses.

When choosing proper TCP enhancements for practical wireless Internet design, drawbacks of each approach should be considered. We list the disadvantages of some wireless TCP schemes in Table 4.

Data-Link-Layer Enhancement (DLE)

Unlike TCP approaches, which focus on end-to-end networking behaviors and ignore the details of immediate nodes, data-link layer enhancement attempts to modify each immediate node to hide the wireless errors from the TCP layer. Thus, the sender will trigger its congestion avoidance algorithm under the assumption that each lost packet is due to networking congestion instead of the wireless link error.

The PILC group provides suggestions on adopting performance enhancing proxies (PEPs) to mitigate the link-related degradations (please refer to RFC 3135 [38]). However, it does not advocate the use of PEPs in general cases since PEPs could interfere with the end-to-end usage of the IP mechanism.

Technique 1: Utilizing the Coarse RTT Timeout

The key technique used in the Transport Unaware Link Improvement Protocol (TULIP) [39] is utilizing the coarse

Technology features / Types of W-TCP	Split TCP (proxy-based)	Maintain end-to-end semantics	Distinguish two types of loss	Handle handoff disconnection	Optimize RTT delay	Save wireless bandwidth	Extra probing packets sent	Modify regular TCP algorithm	Enhance short-TCP
I-TCP	√								
M-TCP	√	√		√		√			
TCP-Decoupling		√	√				√	√	
TCP-Peach		√	√		√	√	√	√	√
WTCP		√	√			√		√	
Fast-Start		√				√		√	√
LEA		√	√				√	√	
Mobile-TCP [70]	√	√		√		√			
DDA	√	√			√	√			

Table 2. *Techniques adopted in different wireless-TCP proposals.*

RTT timeout. Some papers such as [40] conclude that introducing reliability at the data-link layer introduces unnecessary and redundant retransmissions because of the competing retransmission strategies between the transport layer and the data-link layer. However, TULIP disagrees with this opinion. When wireless error loss occurs, TULIP carries out necessary retransmission in the data-link layer within the long timeout period whose value can be calculated by the TCP algorithm (usually multiple times of 500ms).

Technique 2: Using FEC to Reduce Wireless Error Loss

This technique in effect decreases the wireless error rate in the data-link layer by using forward error correction (FEC). Because each erroneous packet will be discarded by the receiver and considered a loss that will trigger the TCP congestion avoidance algorithm, this technique could efficiently overcome packet errors and avoid the activation of the TCP congestion avoidance algorithm.

A typical example that uses this technique is AIRMAIL (Asymmetric Reliable Mobile Access In Link-layer) [41], which is an asymmetric link-layer protocol for reducing the processing load at the mobile host that has limited power compared to the base station, and establishes the reliability by a combination of automatic repeat request (ARQ) and FEC. The key ideas of AIRMAIL include:

- Placing most intelligence in the base station side instead of placing it evenly on both base station and mobile hosts. This can result in as much as a one-third reduction of compiled code.
- Requiring the mobile terminal to combine several ACKs into a single ACK to conserve power.
- Modifying regular FEC to incorporate three levels of channel coding and combing the FEC and link-layer retransmission to obtain better performance in terms of the end-to-end throughput and latency.

Technique 3: Invoking Multiple Flows for Each TCP Connection

This technique of invoking multiple flows for each TCP connection is adopted in the multi-service link layer (MSLL) scheme that can provide multiple simultaneous services at the link layer [42]. In the MSLL scheme, an internal packet classifier distributes incoming packets to appropriate services. The allocation measurement tracts the fraction of data allocated to each service by the classifier. A self-clocked fair queueing (SCFQ) scheduler strictly enforces the desired bandwidth allocation. The goal of the link-layer scheduler is not the provision of end-to-end QoS but the preservation of higher layer scheduling decisions. The scheduler decides which packet to send next based on the measured bandwidth allocations, labels it with a service number, and hands it to the MAC layer for transmission.

Technique 4: Link Shaping

The key idea of link shaping [43] is to use a suitable link-layer error recovery mechanism to hide the fluctuations of the wireless medium from the TCP layer. TCP throughput can deteriorate significantly if the loss rate is above a threshold that scales as the inverse square of the bandwidth-delay product [43]. Link shaping can control the wireless link loss below a threshold, and thus prevent the degradation of TCP throughput. Link shaping suggests that each TCP packet that enters the buffer in the wired/wireless interface is ultimately delivered to the destination of the wireless link without invoking TCP error-recovery mechanism, and the only losses seen by TCP are those due to buffer overflow at the wireless link. Given the packet-loss model in [43], link shaping chooses a proper size of the buffer (B) at the wired-cum-wireless interface so that the buffer overflow probability is less than a warning threshold.

Type	Method to distinguish the e-loss from congestion loss
TCP-Decoupling	Sending TCP-header packets in the TCP circuit besides TCP-data packets.
TCP-Peach	Transmitting dummy packets besides regular TCP packets.
WTCP	Measuring the inter-packet interval in the receiver.
LEA	An ICMP-DEFER message is sent back to the sender.

Table 3. *How some wireless-TCP proposals distinguish between congestion loss and wireless error loss.*

Name	Disadvantages
I-TCP	• Cannot maintain end-to-end semantics; not applicable for encrypted traffic. • Large processing overhead in base station (BS), which can cause resource shortage in the case of many parallel TCP connections established between mobile hosts and BS at a time. • Not applicable for asymmetric networks where ACK flows back in different path.
M-TCP	• Need large buffer size and high-speed processors in BS because BS should maintain all state information for each TCP connection. • Need re-packetization at the TCP sender.
TCP-Decoupling	• Need large modification to current TCP due to the introduction of "circuit." • Could cause extra network load due to the transmission of TCP-header packets. • High implementation cost.
LEA	• BS should send back special message to the sender, which requires the modification to the sender's TCP algorithm.
WTCP	• It adopts rate-based flow control rather than window-based flow control, thus applicable only to specific environments such as wide-area network, and not compatible with regular Internet. • Need to measure the inter-packet interval in the receiver and the sender, which needs further intelligence in MH.
TCP-Peach	• May not be applicable to non-satellite links where propagation delay is not long enough for probing network bandwidth. • May cause extra overhead due to transmission of extra probing packets. • Need large modifications to regular TCP due to the introduction of two new phases: sudden start and rapid recovery.
Fast-Start	• Usable only if a recent value of the congestion window for the same path is available at the sender, which may not be practical in some cases. • Hard to determine the initial slow-start window size.

Table 4. *Drawbacks of some Wireless-TCP proposals.*

Summary of Data-Link-Layer Enhancements

Based on the above discussions of data-link-layer enhancements, the following factors should be considered for the practical choice of a proper technique:

- For current wireless Internet that uses coarse timeouts (generally 500ms), through proper retransmissions of lost packets in the wireless hop (between base station and mobile host), data-link-layer enhancements can avoid the frequent trigger of TCP congestion control due to error loss.
- The choice of the buffer size value in the wired-cum-wireless interface is crucial since the buffer needs to deal with the wireless channel's time variations, i.e., the variations of the bandwidth-delay product.
- The combination of FEC and auto retransmission quest (ARQ) could be adopted in the wireless hop. FEC facilitates the correction of packets with errors before they are delivered to the TCP layer. ARQ can utilize the remaining wireless bandwidth to carry out retransmission. However, the retransmissions should not interfere with the delay-sensitive traffic types such as CBR (constant bit rate).[4]
- Data-link-layer enhancements can carry out error handling and dispatch differentiated-served frames through priority-based queues.

Mobile Performance Enhancement (MPE)

Before discussing enhancement techniques for mobility performance, we need to point out that there are major differences between wireless access and mobility [44]. Wireless access does not mean that mobility should be supported. For example, a fixed-location user might use a wireless interface via a LAN to visit the Internet while sitting in his office. On the other hand, mobile users do not necessarily use wireless interfaces. For instance, a mobile user can simply connect to a fixed network using a wired interface as he moves to a new place. In wireless Internet, mobility performance enhancement is often an important factor.

In many cases people attempt to modify existing Internet protocols to improve mobility performance. Most modifications are based on the concept of Mobile IP [45]. However, mobile performance enhancement (MPE) need not be implemented in the IP layer only. That is, there is a need to implement it in multiple protocol layers to complement each other. This section will introduce several typical MPEs that exhibit good mobility performance in wireless Internet.

Dynamic Host Configuration (DHC) from Microsoft Research

The goal of dynamic host configuration (DHC), which has recently been implemented in the Microsoft wireless Internet called the CHOICE network [46], is to make the mobile host auto-configurable when it moves between cell areas instead of relying on the base station's location management. The crucial protocol that enables the features of the CHOICE network is the Protocol for Authorization and Negotiation of Services (PANS) [47], which supports the global authentication of mobile users and device auto-configuration. DHC addresses the mobility problems by configuring the mobile host to migrate between public and private networks, while regular Mobile IP is primarily concerned with locating the mobile host and re-routing packets to the mobile host's current destination.

Enhanced Mobile IP (EMIP)

Enhanced Mobile IP (EMIP) [48] is the extension of Mobile IP. EMIP performs routing-enhancement in an attempt to implement seamless handoff.[5] Compared to Mobile Ipv6,

[4] *CBR has the top priority and should be transmitted whenever bandwidth is available. When there is bandwidth left, ARQ could be performed for retransmitting the packets with error. However, ARQ cannot use the bandwidth that is reserved for CBR traffic such as continuous voice.*

[5] *The handoff is specified as seamless when the MH does not notice any disruption in the quality of the received application data stream, i.e., the mobile IP duration and the IP packet loss during handoff are minimized.*

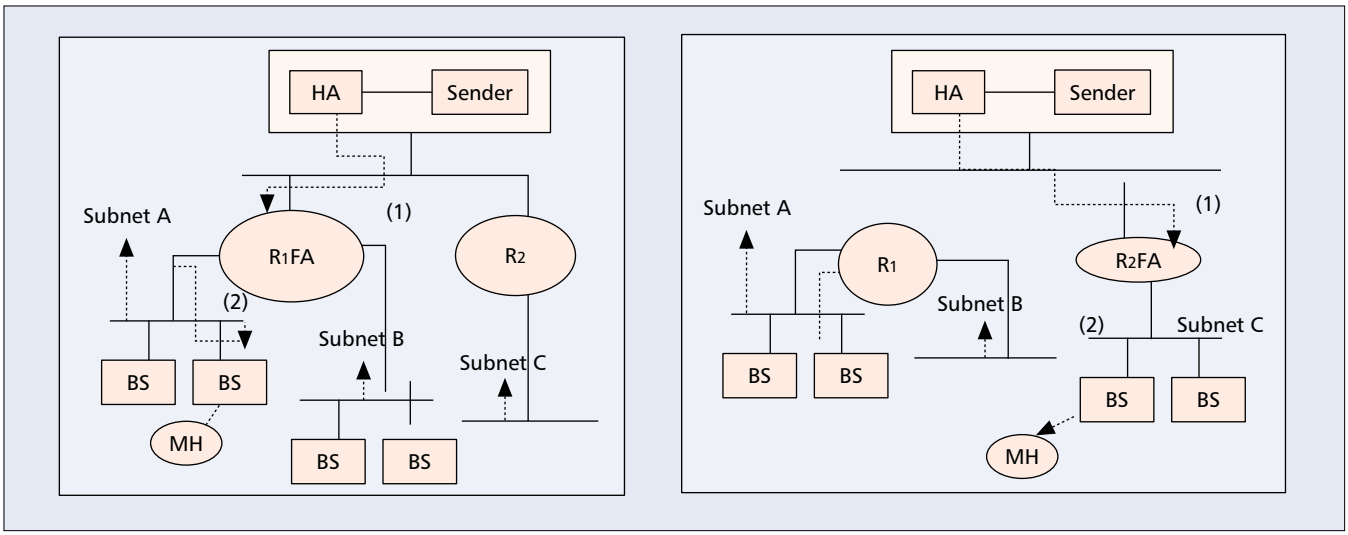

FIGURE 7. *Routing scheme in fast-routing [52].*

EMIP can be performed faster, and can reduce the number of packets that are lost during the handoff. EMIP is more efficient than Mobile Ipv6 since it does not require any additional network bandwidth during the handoff.

FREEZE-TCP (F-TCP)

The key concept of Freeze-TCP (F-TCP) is to change the TCP algorithm in the mobile host so that the base station can be prevented from sending packets during handoff. We outline the F-TCP scheme here because the main goal of F-TCP is to overcome the handoff disconnection instead of wireless link errors.

The idea of F-TCP is to move the onus of signaling an impending disconnection to the mobile host side instead of the base station side [49], because a mobile host can easily monitor signal strengths and detect an impending handoff, and even predict a temporary disconnection. If a handoff occurs, the mobile host advertises a zero window size to force the sender into frozen mode and to prevent it from dropping its congestion window size. This "cheating" behavior is implemented based on one of TCP's characteristics, i.e., the sender will not change its window size if one of the zero window probes is lost.

CELLULAR IP (CIP)

The key goal of Cellular IP is to implement global mobility. Cellular IP is a new approach to support local mobility, and it can also inter-work with Mobile IP to provide wide-area mobility support [50]. Two crucial components of CIP are:

- **Paging caches**, which maintain a distributed cache for location management and can quickly pinpoint idle mobile hosts that wish to engage in active communications. Paging caches can accommodate a large number of mobile hosts without overloading the location-management system.
- **Routing caches**, which maintain the positions of active mobile hosts in the service area and dynamically refresh the routing state in response to the handoff of active mobile hosts.

MOBILE APPLICATION SUPPORT ENVIRONMENT (MASE)

Mobile Application Support Environment (MASE) [51] is implemented in the mobile host as well as in the mobility gateway that is a node between the mobile host and the fixed host. MASE attempts to make the mobile operations transparent by masking the link outages and disconnections due to handoff from the applications. The key component of MASE is the communication manager. The session manager embedded in the communication manager could perform typical session-management functions such as initiating, maintaining, terminating, and recovery of TCP connections.

FAST-ROUTING (FR)

The key concept of Fast-Routing (FR) [52] is routing domain, which is defined as follows:

- All route changes within the routing domain are accomplished with the help of route table changes.
- All route changes between the routing domains are accomplished by Mobile IP.

A routing-domain can be any management-specified domain. In Fig. 7, subnet A and subnet B can form a routing-domain, while subnet C forms another routing domain. A mobile host has a home agent (HA) that is in the same subnet as the sender. The foreign agent (FA) acts as a gateway to a particular routing domain. Two routers in Fig. 7, i.e., R1 and R2, are FAs. The left section of Fig. 7 shows that the mobile host is in the vicinity of subnet A. Routing from HA to the FA (i.e., R1) is carried out through Mobile IP. All mobility within subnets A and B is taken care of by local route changes. When the MH moves into the region of subnet C (another routing-domain), the HA will choose R2 as the FA.

We can see that the routing domain is a useful idea since local mobility can be achieved by changing the local routing table when a mobile host does not move out of the same routing domain. This local routing change can be fast since no global routing change is involved.

FAST-RETRANSMIT

The key concept of Fast Retransmit [53] is to make the base station avoid going through the state of time-out during handoff disconnection since time-out can greatly sacrifice TCP throughput. During the long period of a handoff process, a whole window worth of data may be lost. As shown in the left side of Fig. 8, ACKs cannot be received by the sender after the handoff is complete. Thus, the sender will time-out and retransmit the lost packets. During the retransmission of lost packets, another time-out could occur if handoff is still not complete.

Fast Retransmit suggests that the mobile host can immedi-

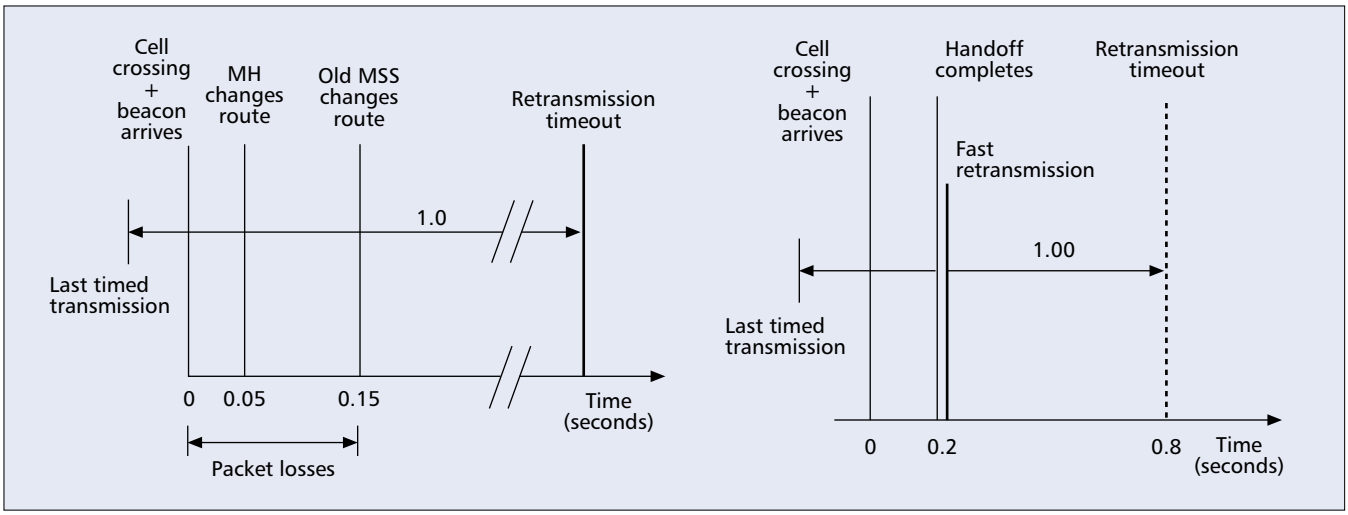

FIGURE 8. *(Left) Handoff problem; (Right) Using Fast Retransmit [53].*

ately send back three duplicate ACKs to the sender when it finishes a handoff (right side of Fig. 8). This will trigger the sender to enter fast retransmit instead of waiting for the occurrence of a time-out. Thus, TCP throughput can be improved during handoff disconnection since the time-out event is avoided.

MULTIMEDIA-QOS ENHANCEMENT

This century will see more and more multimedia applications in wireless environments [54]. The most important issue for wireless multimedia communications is to maintain mobile-QoS requirements, such as handoff delay, disconnection loss rate, and desired bandwidth [55]. The main policies of multimedia-QoS enhancement include intelligent resource ReSerVation Protocol (RSVP), adaptive resource management, and Class-Based Queuing (CBQ).

Recently, the Next Steps In Signaling (NSIS) working group discussed the requirements for QoS signaling protocols in an Internet Draft [56]. One important point from NSIS is that, in wireless environments, it is essential to define interoperability between multimedia traffic, mobile behavior, and QoS management [56].

DISTRIBUTED FAIR SCHEDULING (DFS) SCHEME FROM MICROSOFT RESEARCH

The Microsoft Communications Projects Research Group proposed the concept of Distributed Fair Scheduling (DFS) to allocate wireless resources in proportion to the priorities of the packet flows sharing the wireless channels [57]. DFS happens in the media access control (MAC) layer of wireless networks. A key technique adopted by DFS is the Exponential Mapping Scheme, which calculates the value of the backoff interval[6] based on an exponential function instead of a simple linear function that is currently adopted by the IEEE 802.11 standard [57]. DFS overcomes the shortcomings of the IEEE 802.11 standard that could not perform fair resource allocation among different multimedia flows on short time scales.

[6] *In Wireless MAC layer, when two mobile hosts want to send their packet simultaneously, a collision will occur. To avoid the collision, the senders will wait for a Backoff Interval whose value is uniformly distributed over a range.*

MODIFIED RSVP

The key idea of Modified RSVP is to modify the regular RSVP to adapt to the wireless multi-user sharing environment. Regular RSVP is a signaling mechanism for carrying the QoS parameters from the sender to the receiver, and also for making reservations along the communication path.

Because the wireless link is shared among multiple mobile hosts in uplink and downlink directions, it is necessary to adopt a link-sharing mechanism such as Class-Based Queuing (CBQ) that can meet the QoS requirements of the mobile hosts. Modified RSVP is thus based on the combination of regular RSVP and the CBQ scheme [52].

MOBIWEB

The key idea of MobiWeb is to adopt a proxy-based architecture to enhance multimedia performance [58]. It includes a priority scheme that preserves media smoothness despite short-term fluctuations of the wireless link, as well as an adaptation mechanism applied to streams to respond to long-term link changes and handoffs.

The adaptation mechanism works as follows: MobiWeb adopts a set of timers to increase the utilization of the resources when the resources are abundant, but forces enhanced streams to back off to more moderate resource usage when resources are scare.

The priority scheme consists of two aspects:

- An admission control scheme is used to reserve resources for each stream in order to guarantee at least the minimum acceptable level of quality in the stream's performance.
- A dynamic prioritization scheme is used to protect the higher-priority streams from losing their shares of resources to streams with lower priority.

IP-QOS

The key idea of IP-QoS is to integrate the existing QoS approaches, such as RSVP over IP layer, with the protocols supporting mobility [59]. It optimizes the existing differentiated services architecture standard based on three levels of protocol implementations:

- Session Layer Negotiation. The QoS-aware applications are able to specify their own traffic and QoS requirements, and a QoS API is then used to map the application service profile to an understandable form for the underlying resource manager (RM).

Layer	Efficient schemes on power-saving
Application layer	Some APIs, such as the Advanced Power Interface invented by Intel Corporation, could help software developers create programs that are more power-conserving.
Transport layer	Some studies demonstrate that error correlation could significantly affect the energy performance of TCP. The TCP congestion control algorithm could actually allow for greater energy savings by backing off and waiting during error bursts [64].
Network layer	We should not use shortest-delay or shortest-hop as the only metric for routing efficiency. Energy consumed per packet is an important metric, since many studies show that the usage of power-aware metrics can result in no extra delay over the traditional shortest-hop metric [64].
MAC layer	Efficient mechanisms in the MAC layer include: 1) turning off the transceiver whenever the node determines that it will not receive data for a period of time; and 2) allocating contiguous slots for transmission or reception to reduce turn-around that could consume power.
Data-link layer	There should be a careful trade-off between 1) the retransmission of packets when responding to loss and 2) the power consumption resulting from data transmission.
Physical layer	Low-power design focuses on the hardware techniques such as a CPU with variable clock speed and flash memory. Efficient encoding/decoding in the physical layer is also important for power-conservation.

Table 5. *Efficienct schemes for saving power in protocol layers of W-Internet.*

- End-to-End QoS Resource Reservation. RM performs the resource allocation and admission control for the core network.
- Local Inquiry. It is used for local resource inquiry and can communicate with the global network RM.

Selective Packet Prioritization (SPP)

The main goal of Selective Packet Prioritization (SPP) [60] is to enhance the performance of Wireless Voice IP (VoIP) service, which is a typical wireless multimedia application and is gaining great attention today. The SPP architecture assumes a hot-spot network scenario, such as an IEEE 802.11 wireless LAN where the interface above the MAC layer is indistinguishable from the IP interface on a fixed network.

SPP attempts to develop a QoS-support scheme that selectively marks the packets to a higher network priority at the sender based on the properties of the speech signal. Differentiated services packet priorities are mapped to the QoS control mechanism of the lower protocol layers. Thus, prioritized packets are protected against other flows using the shared medium and also against channel errors.

Alternative Non-TCP Enhancements

For specific wireless scenarios, such as last-hop-wireless or some high-speed campus wireless LAN, more efficient wireless protocol suites may be used to optimize the local wireless performance.

A typical example is Wireless Application Protocol (WAP), which defines an industry-wide specification for wireless networks [61]. A key aspect of WAP is that all connections between MH and FH will pass through a proxy, i.e., WAP gateway. Thus, WAP is a split-connection proposal.

If the last hop is a wireless link, another efficient non-TCP protocol, called Mobile-End Transport Protocol (METP), replaces TCP/IP in wireless Internet by a simpler protocol with smaller headers [62].

SWAN (Seamless Wireless ATM Network) [63] is one of the earliest systems to realize the concept of the mobile ATM network. It uses room-sized pico cells and mobile multimedia end-points.

Power-Consumption Enhancement (PCE)

One of the hurdles to the extension of wireless Internet is the high power consumption. This obstacle has not been overcome efficiently today. The value of power consumption depends on parameters such as the amount of avoidable extra traffic and the total duration of the TCP connection. The key to saving power in wireless Internet lies within the higher levels of the wireless protocol stack instead of **only** the physical hardware.

Table 5 summarizes the typical schemes in the protocol layers of wireless Internet for saving power.

Discussions

Methodology for the Choice of Proper Enhancements

The choice of enhancements for wireless Internet should be based on different networking architectures and performance requirements. For example, since wireless LAN usually forms an independent Intranet, ATM-based or other non-TCP/IP architectures could be adopted. When one is choosing a proposal, trade-offs should be made between performance efficiency and implementation cost. For instance, because TCP is an end-to-end protocol, the requirements of modifying all immediate routers may not be practical in global mobile networks. Table 6 lists the appropriate environments of some enhancements.

Other W-Internet Layer Enhancements

Apart from the abovementioned enhancements, there could be other enhancements that can be applied to other wireless Internet protocol layers, including the session layer, the MAC layer, and the physical layer.

A hot research area for the session layer of the cellular Internet (one type of wireless Internet) is the Connection Admission Control (CAC), which focuses on allocating wireless bandwidth to handoff calls and new calls efficiently [65]. We have proposed a novel CAC algorithm for multi-class wireless applications in [66]. For MAC layer performance enhancement, [67] lists some typical MAC proposals. One of the main goals on the MAC layer is saving power. On physical-layer enhancements, coding and retransmission are the main strategies. An efficient error-control architecture for wireless ATM was recommended in [68] (Fig. 9). In addition, a high-rate

Class	Sub-class	Appropriate environments
Application-layer enhancement	Dynamic URLs	For wireless environments that have frequent mobility handoffs.
	Protocol reduction	Wireless links are limited to a specific local environment that does not have strict end-to-end semantics guaranteeing requirements.
	Cache relocation	In an environment where the following assumptions can be made: 1) the system has user profiles stored in the home network; 2) the mobile host can perform path-prediction somewhat accurately
	Differencing	It is based on the adoption of an intermediate layer below the application layer called intercept layer.
TCP-layer enhancements	TCP-Peach	Appropriate for the wireless systems with long propagation delays and high link error rates such as in satellite links.
	WTCP	Useful when the system has the following features: very low and variable bandwidths; very high and variable delays; heavy error-loss; asymmetric uplink and downlink channels; and occasional blackouts, such as in wireless wide-area-networks (CDPD, etc.)
	M-TCP	Useful when the system wants to maintain the end-to-end semantics, and has frequent mobile disconnections and long-time disconnections due to heavy multipath fading (such as in cellular networks).
	I-TCP	Only useful in specific sub-networks where the base-station can be extended as a mediation that hides the error-loss in wireless links from the fixed host.
	Short-TCP	For wireless Internet scenarios where most TCP connections are short-sized.
	TCP-Decoupling	For wireless Internet where it is possible to modify traditional TCP algorithms. When it is necessary to consider TCP-friendly multimedia stream and ATM virtual circuits.
Link-layer enhancements	TULIP	It uses coarse and generous timeout values instead of accurate RTT value. Useful in half-duplex radio links. It needs the support from the MAC layer.
	MSLL	Used when there is a need for QoS support. Useful when there is non-TCP traffic.
	AIRMAIL	Used when implementation costs permit the combination of FEC and ARQ. BS is highly intelligent. Useful in very error-prone links, and when the mobile host has very limited power.
Mobility enhancements	Cellular IP	Used in environments where mobile hosts migrate frequently. Cellular networks that request wide-area mobility and low handoff delay.
	Fast routing	Useful when there is a need for small routing table size; in cases with frequent handoffs. Useful when it is permitted to define the local routing tables, such as in a wireless campus network.
	Freeze-TCP	Useful when there are frequent handoff disconnections. When there is a need to handle encrypted traffic, keep end-to-end TCP semantics. When implementation cost is a high priority.
Multimedia-QoS enhancement	SPP	Used in "hot-spot" network scenario such as an IEEE 802.11 WLAN where the interface above the MAC level is indistinguishable from an IP interface on a fixed network.
	MobiWeb	Useful when real-time applications need to adapt to the high-varying wireless links. Only applicable between the base station and the mobile host instead of in an end-to-end scenario.
	IP-QoS	It needs the use of Mobile Ipv6 and an end-to-end implementation.
Non-TCP/IP enhancements	METP	Used for last-hop-wireless only. Highly bursty wireless link. When hand-off latency is not a major concern and the BS is well connected, and mobile host does not have many connections at any time.
	SWAN	Used in indoor wireless network. Based on room-sized pico-cells and mobile multimedia mobile hosts such as PDA. With wired-ATM backbone network and wireless ATM last-hops to the mobile hosts. Using 2.4 GHz ISM band radios.
	WAP	It needs inter-cooperation from other companies. When there is low bandwidth and high-latency. When good security is needed.

Table 6. *Appropriate environments for some enhancements.*

extension of the IEEE 802.11 physical-layer protocol is proposed in [69] for enabling 22 Mb/s wireless transmission.

Conclusions

This article reviewed the modern approaches for improving wireless Internet performance from the following seven aspects:

- Application-layer
- TCP-layer
- Data-Link-layer
- Multimedia-QoS
- Mobility performance
- Power consumption
- Other non-TCP/IP schemes

We discussed the essential features for each enhancement technology, and compared the advantages and disadvantages of these approaches. It was pointed out that the selection of different enhancement schemes should be based on specific

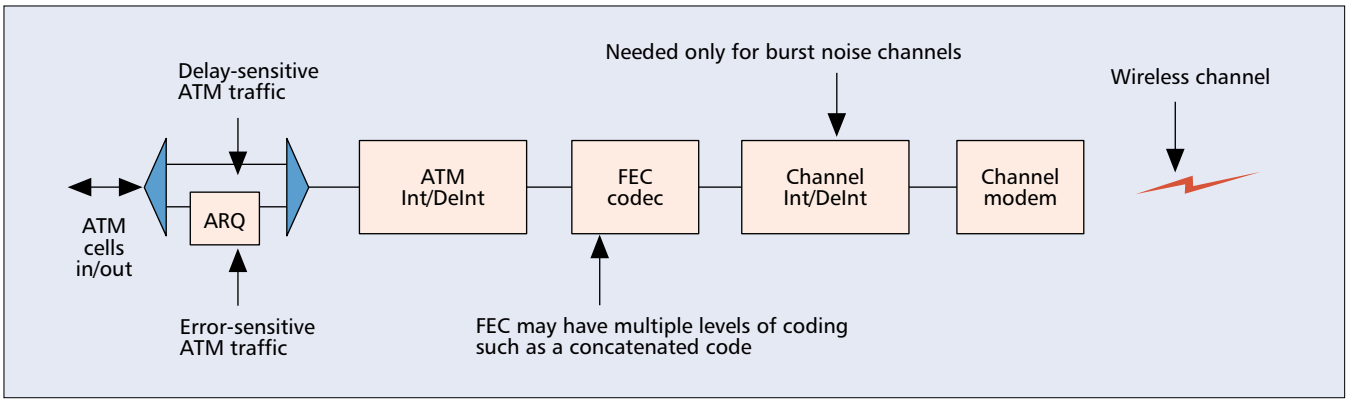

FIGURE 9. *Enhancement in the Physical Layer [68].*

characteristics of the wireless networks and implementation costs. In order to obtain the optimized networking performance, the enhancements had better be carried out in multiple protocol layers, including the physical layer, the MAC layer, the CAC layer, the data-link layer, the TCP layer, the Mobile IP panel, and the application layer.

REFERENCES

[1] M. Guizani *et al.*, "Wideband Wireless Access Technologies to Broadband Internet," *IEEE Commun. Mag.*, Apr. 2002, vol. 40, no. 4, pp. 34–35.

[2] J. H. Park, "Wireless Internet Access for Mobile Subscribers Based on the GPRS/UMTS Network," *IEEE Commun. Mag.*, Apr. 2002, vol. 40, no. 4, pp. 38–49.

[3] D. Vardalis, "On the Efficiency and Fairness of TCP over Wired/Wireless Networks," Master's thesis, State University of New York at Stony Brook.

[4] H. Balakrishnan *et al.*, "Improving TCP/IP Performance over Wireless Networks," *Proc. ACM MobiCom*, Nov. 1995.

[5] S. Hadjiefthymiades and L. Merakos, "Improving the Performance of the World Wide Web in Cellular CPN Environments," *Proc. 5th Int'l. Wksp. Mobile Multimedia Commun.*, Berlin, German, 1998.

[6] G. M. Voelker and B. N. Bershad, "Mobisaic: An Information System for a Mobile Wireless Computing Environment," *Proc. 1994 IEEE Wksp. Mobile Computing Systems and Applications*, Santa Cruz, California, 1994.

[7] B. C. Housel and D. B. Lindquist, "WebExpress: A System for Optimizing Web Browsing in a Wireless Environment," *Proc. ACM/IEEE MobiCom'96*, New York, Oct. 1996.

[8] S. Campadello *et al.*, "Performance Enhancing Proxies for Java RMI over Slow Wireless Links," *Proc. Second Int'l.Conf. Practical Application on Java*, 12-14 Apr. 2000, Manchester, UK, pp. 76–89.

[9] M. F. Kaashoek, T. Pinckney, and J. A. Tauber, "Dynamic Documents: Extensibility and Adaptability in the WWW," *Proc. 2nd Int'l. WWW Conf.*, Oct. 1994.

[10] T. B. Fleming, S. F. Midkiff, and N. J. Davis, "Improving the Performance of the WWW over Wireless Networks," http://www.cs.columbia.edu/~hgs/InternetTC/GlobalInternet97/Flem9711_Improving.pdf

[11] M. Kacin, "Optimizing Web Pages For Handheld Devices," design tips from software maker AvantGo, Inc. , http://www.intranetjournal.com/features/avantgo/designtips.shtml

[12] R. Krashinsky, "Efficient Web Browsing for Mobile Clients Using HTTP Compression," See: http://www.cag.lcs.mit.edu/~ronny/classes/httpcomp.pdf

[13] M. Kojo *et al.*, "An Efficient Transport Service for Slow Wireless Telephone Links," *IEEE JSAC*, vol. 15, no. 7, Sept. 1997.

[14] V. Tsaoussidis and I. Matta, "Open Issues on TCP for Mobile Computing," available from http://citeseer.nj.nec.com/tsaoussidis02open.html

[15] S. Dawkins *et al.*, "End-to-end Performance Implications of Links with Errors," RFC 3155, PILC Working Group, Internet Graft, Aug. 2001.

[16] H. Inamura *et al.*, "TCP over Second (2.5G) and Third (3G) Generation Wireless Networks," PILC Working Group, IETF Draft, draft-ietf-pilc-2.5g3g-08.txt, (work in progress), May 2002.

[17] T. R. Henderson and R. H. Katz, "Transport Protocols for Internet-Compatible Satellite Networks," *IEEE JSAC*, vol. 17, no. 2, 1999.

[18] H. Kruse, "Performance of Common Data Communications Protocols over Long Delay Links: An Experimental Examination," *3rd Int'l. Conf. Telecommun. Systems Modeling and Design*, 1995.

[19] M. Allman, "Improving TCP Performance over Satellite Channels," Master's thesis, College of Engineering and Technology, Ohio University, June 1997.

[20] I. F. Akyildiz, Giacomo Morabito, and Sergio Palazzo, "TCP-Peach: A New Congestion Control Scheme for Satellite IP Networks," *IEEE/ACM Trans. Net.* (TON), vol. 9, no. 3 (June 2001), pp. 307–21

[21] M. Gerla, W. Weng, and R. Lo Cigno, "BA-TCP: A Bandwidth Aware TCP for Satellite Networks," *Proc. IEEE ICCCN'99*, Boston, Massachusetts, Oct. 1999.

[22] A. Bakre and B. R. Badrinath, "I-TCP: Indirect TCP for Mobile Hosts," *Proc. 15th Int'l. Conf. Distributed Computing Systems (IDCS)*, May 1995.

[23] K. Brown and S. Singh, "M-TCP: TCP for Mobile Cellular Networks," *Comp. Commun. Review*, vol. 27, no. 5, Oct. 1997.

[24] S. Y. Wang and H. T. Kung,"Use of TCP Decoupling in Improving TCP Performance over Wireless Networks," *Mobile Networks and Applications*, 2000, pp. 1–20.

[25] N. Cardwell, S. Savage, and T. Anderson, "Modeling the Performance of Short TCP Connections," Technical Report, Computer Science Department, Washington University, Nov. 1998.

[26] K. Thompson, G. J. Miller, and R. Wilder, " Wide-Area Internet Traffic Patterns and Characteristics," *IEEE Network*, vol. 11, no. 6, Nov. 1997, pp. 10–23.

[27] Y. Zhang, L. Qiu, and S. Keshav, "Speeding Up Short Data Transfers: Theory, Architectural Support, and Simulation Results," *Proc. NOSSDAV 2000*, Chapel Hill, North Carolina, USA, June 2000.

[28] S. Goel and D. Sanghi, "Improving Performance of TCP over Wireless Links," *Proc. TENCON 98 – 1998 IEEE Region Ten Conf. Global Connectivity in Energy, Computer Communication and Control*, Dec. 17–19, 1998.

[29] P. Sinha *et al.*, "WTCP: A Reliable Transport Protocol for Wireless Wide-Area Networks," *Proc. ACM MOBICOM*, 1999.

[30] S. Biaz and N. H. Vaidya, "Discriminating Congestion Losses from Wireless Losses Using Inter-Arrival Times at the Receiver," *Proc. IEEE ASSET'99*, 1999, pp. 10–17.

[31] K. S. Phanse, "Analysis of TCP Performance over Asymmetric Wireless Links," available from http://fiddle.visc.vt.edu/courses/ecpe6504-wireless/projects_spring2000/pres_phanse.pdf

[32] H. Balakrishnan, V. Padmanabhan, and R. Katz, "The Effects of Asymmetry on TCP Performance," *ACM Mobicom'97*, Sept. 1997.

[33] L. Kalampoukas, A. Varma, and K. Ramakrishnan, "Improving TCP Throughput over Two-Way Asymmetric Links: Analysis and Solutions," *Sigmetrics'98*, June 1998.

[34] I. T. Ming-Chit, D. Jinsong, and W. Wang, "Improving TCP Performance over Asymmetric Networks," available from:

http://www.acm.org/sigcomm/ccr/archive/2000/july00/tam.pdf
[35] N. H. Vaidya *et al.*, "Delayed Duplicate Acknowledgments: A TCP-Unaware Approach to Improve Performance of TCP over Wireless," Technical Report 99-003, Computer Science Department, Texas A&M University, Feb. 1999.
[36] H. Balakrishnan, S. Seshan, and R. Katz, "Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks," *ACM Wireless Networks*, vol. 1, Dec. 1995.
[37] A. Kumar, "Comparative Performance Analysis of Versions of TCP in a Local Network with a Lossy Link," *IEEE/ACM Trans. Net*, vol. 6, 1998, pp. 485–98.
[38] J. Border *et al.*, "Performance Enhancing Proxies Intended to Mitigate Link-Related Degradations," RFC 3135, PILC Working Group, Internet Graft, June 2001.
[39] Christina Parsa and J. J. Garcia-Luna-Aceves, "Improving TCP Performance over Wireless Networks at the Link Layer," *Mobile Networks and Applications*, no. 5, 2000, pp. 57–71.
[40] A. DeSimone, M. C. Chuah, and O. C. Yue, "Throughput Performance of Transport-Layer Protocols over Wireless LANs," *Proc. IEEE Globecom 1993*, pp. 36–46.
[41] E. Ayanoglu *et al.*, "AIRMAIL: A Link-Layer Protocol for Wireless Networks," *ACM/Baltzer Wireless Networks Journal*, Feb. 1995, pp. 47–60.
[42] G. Xylomenos and G. C. Polyzos, "Link Layer Support for Quality of Service on Wireless Internet Links," *IEEE Pers. Commun.*, vol. 6, no. 5, 1999, pp. 52–60.
[43] T. V. Lakshman and U. Madhow, "The Performance of TCP/IP for Networks with High Delay Bandwidth Products and Random Loss," *IEEE/ACM Trans. Net.*, vol. 5, no. 3, June 1997.
[44] G. H. Forman and J. Zahorjan, "The Challenges of Mobile Computing," *IEEE Computer*, vol. 27, no. 4, 1994.
[45] C. Perkins, "Mobile IP," *IEEE Commun. Mag.*, May 1997, vol. 35, no. 5, pp. 84–99.
[46] CHOICE Network, designed by Microsoft Research, please visit: http://www.mschoice.com. Detail is stated in " The CHOICE Network: Broadband Wireless Internet Access in Public Places," *MSR-TR-2000*, Feb. 2000.
[47] A. Miu and P. Bahl, "Dynamic Host Configuration for Managing Mobility between Public and Private Networks," *Mobile Communications Group, Microsoft Research*, *3rd Usenix Int'l. Tech. Symp.*, San Francisco, California, USA, Mar. 2001.
[48] G. Karagiannis, " Mobility Support for Ubiquitous Internet Access," Tech. Rep., available from http://ing.ctit.utwente.nl/WU4/.
[49] T. Goff *et al.*, Freeze-TCP: A True End-To-End Enhancement Mechanism for Mobile Environments," *INFOCOM* (Israel), 2000.
[50] A. G. Valko, "Cellular IP: A New Approach to Internet Host Mobility," *ACM Comp. Commun. Review*, Jan. 1999.
[51] V. Chandrasekaran and A. Lo, "The Design and Implementation of a Session Manager for Mobile Applications," *Proc. 3rd ACTS Mobile Summit*, Rhodes, Greece, June 1998.
[52] I. Mahadevan and K. M. Sivalingam, "An Architecture for QoS Guarantees and Routing in Wireless/Mobile Networks," *Proc. 1st ACM Int'l. Wksp. Wireless Mobile Multimedia*, 1998, pp. 11–20.
[53] R. Caceres and L. Iftode, "Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments," *IEEE JSAC*, vol. 13, no. 5, June 1995.
[54] Lajos Hanzo, "Bandwidth-Efficient Wireless Multimedia Communications," *Proc. IEEE*, vol. 86, no. 7, July 1998, pp. 1342–75.
[55] D. Chalmers and M. Sloman, "A Survey of Quality of Service in Mobile Computing Environments," *IEEE Commun. Surveys & Tutorials*, 2nd Quarter 1999, vol. 2, no. 2, http://www.comsoc.org/livepubs/surveys/index.html
[56] M. Brunner *et al.*, "Requirements for QoS Signaling Protocols," NSIS Working Group, IETF Draft, draft-ietf-nsis-req-02.txt, (work in progress), May 2002.
[57] N. H. Vaidya, P. Bahl, and S. Gupta, "Distributed Fair Scheduling in a Wireless LAN," *Proc. ACM Int'l. Conf. Mobile Computing and Networking*, Mobile Communications Group, Microsoft Research, MobiCom 2000, Boston, Massachusetts, USA, Aug. 2000
[58] M. Margaritidis and G. C. Polyzos, " MobiWeb: Enabling Adaptive Continuous Media Applications over Wireless Links", available from http://www-cse.ucsd.edu/users/margarit/papers/3g00-Margaritis-Margaritidis-80.pdf
[59] V. Rexhepi, G. Karagiannis, and G. Heijenk, "A Framework for QoS and Mobility in the Internet Next Generation," *Proc. EUNICE 2000*, 6th EUNICE Open European Summer School, University of Twente, Enschede, the Netherlands, Sept. 13–15, 2000, also available from: http://ing.ctit.utwente.nl/WU4/
[60] H. Sanneck *et al.*, "Selective Packet Prioritization for Wireless Voice over IP," available from http://www.cs.unc.edu/~le/papers/Sann0109_Wireless-VoIP.pdf
[61] WAP Forum, Wireless Application Protocol Forum (www.wapforum.org), Apr. 2001
[62] K.-Y. Wang and S. K. Tripathi, "MobileEnd Transport Protocol: An Alternative to TCP/IP over Wireless Links," *IEEE INFOCOM'98*, San Francisco, California, Mar. 1998.
[63] P. Agrawal *et al.*, "SWAN: A Mobile Multimedia Wireless Network," *IEEE Pers. Commun.*, Apr. 1996, vol. 3, no. 2, pp. 18–33.
[64] C. E. Price *et al.*, "A Survey of Energy Efficient Network Protocols for Wireless and Mobile Networks," accepted for *ACM/Baltzer J. Wireless Networks*, Jan 2001.
[65] I. Katzela and M. Naghshineh, "Channel Assignment Schemes for Cellular Mobile Telecommunication Systems: A Comprehensive Survey," *IEEE Pers. Commun.*, June 1996, vol. 3, no. 3, pp. 10–31.
[66] F. Hu and N. K. Sharma, "A Novel Approach to Call Admission Control for Multimedia Traffic in Wireless ATM Networks," *IEEE ICC 2001*, Helsinki, Finland.
[67] A. Chandra, V. Gummalla, and John O. Limb, "Wireless Medium Access Control Protocols," *IEEE Commun. Surveys & Tutorials*, 2nd Quarter 2000, http://www.comsoc.org/livepubs/surveys/index.html
[68] J. B. Cain and N. McGregor, "A Recommended Error Control Architecture for ATM Networks with Wireless Links," *IEEE JSAC*, vol. 15, no. 1, Jan. 1997, pp. 16–28.
[69] C. Heegard *et al.*, "High-Performance Wireless Ethernet," *IEEE Commun. Mag.*, Nov. 2001, vol. 39, no. 11, pp. 64–73.
[70] Z. Haas and P. Agrawal, "Mobile-TCP: An Asymmetric Transport Protocol Design for Mobile Systems," *Int'l. Conf. Commun.*, Montreal, Quebec, Canada, June 1997, IEEE.

Additional Reading

[1] H. M. Chaskar, T. V. Lakshman, and U. Madhow, "TCP over Wireless with Link Level Error Control: Analysis and Design Methodology," *IEEE/ACM Trans. Net.*, vol. 7, no. 5, Oct. 1999, pp. 605–15.

Biographies

Fei Hu (fei.hu@ieee.org) is currently working as an assistant professor in the Department of Computer Engineering, Rochester Institute of Technology in Rochester, New York. He obtained his Ph.D. degree in 2002 in electrical and computer engineering at Clarkson University, Potsdam, New York. He received a Master's degree in telecommunication engineering from Shanghai Tiedao University of China in 1996. His research interests include high-speed computer networks, wireless and mobile computing, Internet, and ATM.

Neeraj K. Sharma (sharman@clarkson.edu) is currently a system engineer at Intel Corporation. From 1999 to 2000 he was an associate professor with the Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY. From 1993 to 1998 he was on the faculty of the Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Australia. He received his Ph.D. and MSEE degrees from the University of Akron, Akron, Ohio, and a BSEE degree from the University of South Alabama, Mobile, Alabama, all in electrical engineering. His research interests include fault-tolerant system design, and performance and reliability analysis of computer systems and networks.