Layered View of QoS Issues in IP-Based Mobile Wireless Networks

Haowei Bai
Honeywell Labs
3660 Technology Drive, Minneapolis, MN 55418
E-mail: haowei.bai@honeywell.com

Mohammed Atiquzzaman
School of Computer Science
University of Oklahoma, Norman, OK 73019-6151
E-mail: atiq@ou.edu

David Lilja
Department of Electrical and Computer Engineering
University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455
E-mail: lilja@ece.umn.edu

Abstract

With the convergence of wireless communication and IP-based networking technologies, future IP-based wireless networks are expected to support real-time multimedia. IP services over wireless networks (e.g., wireless access to Internet) enhance the mobility and flexibility of traditional IP network users. Wireless networks extend the current IP service infrastructure to a mix of transmission media, bandwidth, costs, coverage, and service agreements, requiring enhancements to the IP protocol layers in wireless networks. Furthermore, QoS provisioning is required at various layers of the IP protocol stack to guarantee different types of service requests, giving rise to issues related to cross-layer design methodology. This paper reviews issues and prevailing solutions to performance enhancements and QoS provisioning for IP services over mobile wireless networks from a layered and cross-layer design point of view.

1 Introduction

IP-based network technology had tremendous growth in recent years, and is becoming the backbone of the next generation data network. In the meanwhile, mobile wireless networks have gone through exponential growth in terms of the number of mobile telecommunication service subscribers and wireless LAN users. Mobile wireless networks have evolved from first generation networks carrying only voice to the current 3G (and beyond) networks based on the all-IP architecture. The wireless LAN technology is considered a complement toward wide-area 3G networks.

The concept of 3G networks is based on an all-IP architecture supporting voice, video and data, and is driven by the needs for more bandwidth, more network capacity, and new radio spectrum. This gives rise to the need for performance enhancements and QoS guarantees in wireless networks. Internet Engineering Task Force (IETF) has defined QoS as a service agreement (or a guarantee) to provide a set of measurable networking service attributes, including end-to-end delay, delay variation (jitter), and available bandwidth.

QoS issues have been widely studied for conventional IP networks, and almost all necessary elements now exist for providing QoS support in conventional IP networks. However, the conventional IP network architecture was originally designed for fixed nodes connected by wired links. Mobile wireless networks have a few fundamentally different characteristics from conventional wired networks (See [1] for a tutorial on wireless errors and their models). These include:

- low bandwidth wireless links,
- high link error rate of wireless links, and

- mobility of end hosts resulting in hand-offs between access points.

Due to these intrinsic differences between conventional wired networks and mobile wireless networks, network designers must address the problem of efficiently and effectively delivering broadband IP traffic, as well as satisfy QoS requirements, when the transmission medium (e.g., RF for wireless) is impaired by physical layer characteristics. Therefore, a good QoS model for wireless networks must be able to satisfy two requirements: *first*, to compensate for impairments of the wireless medium, and enhance the network throughput and the link capacity; and *second*, to perform resource negotiation, allocation and traffic control to provide QoS to end users. Although not directly relevant to QoS guarantees, the first requirement determines the efficiency and effectiveness of QoS protocols. QoS guarantees and service classifications would not be possible if a higher-layer QoS protocol, which is independent of lower layers, ran over a high-BER (Bit Error Rate) physical wireless link. In other words, a good QoS architecture is the coordination and cooperation among all layers of the IP protocol stack. Therefore, the goals of a QoS architecture for wireless networks is to compensate for impairments of the wireless medium, in addition to performing QoS management for wireless network resources.

What are essential issues to be addressed in order to achieve the above two goals, i.e., compensating wireless medium impairments and performing QoS management? *First*, a reliable low-delay physical link is necessary for TCP/IP performance and real-time traffic. Mobile hosts are easily affected by multipath propagation. Multipath propagation can cause fluctuations in the received signal's amplitude, phase and the angle of arrival, which yields transmission errors leading to packet losses. It degrades the performance of higher-layer QoS protocols by decreasing throughput and increasing the end-to-end delay. Solutions are therefore required to cope with multipath fading in wireless networks.

Secondly, the current TCP algorithm, originally designed for wireline networks, responds to all packet losses by decreasing the congestion window size and retransmitting lost packets. In mobile wireless networks, high packet losses due to link corruptions and hand-offs force TCP to unnecessarily reduce the congestion window size which degrades throughput and increases end-to-end delay. Consequently, several TCP enhancements have been proposed for use in wireless networks.

Thirdly, although a host which is connected to a wireless network does necessarily have to be mobile, increasing number of wireless hosts are becoming mobile. A network consisting of mobile wireless hosts, such as an Ad Hoc network, may have a dynamic topology, requiring frequent routing information updates to maintain network connectivity and packet forwarding path. Supporting QoS gives rise to the issue of QoS routing in mobile wireless environments, such as Ad Hoc networks.

Finally, most of the proposed QoS architectures and protocols are based on IETF's QoS parameters, i.e., bandwidth, delay and jitter. In practice, end user's QoS requirements may have to be translated and mapped to IETF's parameter set used by network protocols. Application software may be expected to be portable and reusable across a variety of commercially available lower-layer networking products. End users may have QoS requirements on local computing resources such as battery power. QoS-aware middleware can satisfy the above requirements.

The *objective* of this article is to provide a comprehensive survey from a layered view of design challenges and available solutions for QoS issues during the convergence of IP networks and mobile wireless networks. The discussion is based on a layered and cross-layer view of protocol stack for mobile wireless networks. In the rest of this paper, we review the QoS-aware middleware architecture in Sec. 2, TCP in wireless networks and its enhancements in Sec. 3, enhancing mobility performance at IP layer in Sec. 4, and enhancing the performance of physical layer in Sec. 5, followed by concluding remarks in Sec. 6. Enhancing IP services over mobile wireless networks is an emerging research topic, and many issues still remain open.

2 QoS-aware Middleware in Wireless Networks

The concept of *middleware* layer was originally developed for distributed systems. It is a layer of software between the operating system and the application program to provide an abstraction for the heterogeneity of operating system, networks, hardware, and even programming languages. Therefore, adding a middleware layer facilitates easy deployment of commercial off-the-shelf (COTS) hardware and software components which results in reduced system complexity and development costs.

2.1 Middleware technologies

The various middleware technologies are classified in terms of different programming abstraction as follows:

- *Transaction processing monitor (TP)* provides distributed client/server environment the capability of managing multi-database transactions.
- *Remote procedure call (RPC)* allows a procedure to be invoked across a network. It reduces the development complexity of applications spanning multiple operating systems and network protocols.
- *Message oriented middleware (MOM)* enables program-to-program data exchange by providing the abstraction of a message queue that can be accessed across a network.
- *Distributed object middleware* allows methods of a remote object to be invoked and shared distributively across heterogeneous networks. This makes object-oriented programming techniques to be available to distributed and networking application developers. Examples are Common Object Request Broker Architecture (CORBA) developed by the Object Management Group (OMG), Distributed Component Object Model (DCOM) and COM+ (the next generation DCOM) from Microsoft, and Remote Method Invocation (RMI) from Sun Microsystems.

2.2 QoS-aware middleware architecture in wireless networks

With the proliferation of real-time services in the Internet, providing end-to-end QoS has become a very important requirement for next generation data networks. Many network architectures and protocols have been developed and standardized to provide end-to-end QoS. However, these protocols are based on the assumption that the application's QoS requirements are transparent to the network layer. Based on this assumption, the end user's requirements are abstracted to a set of measurable parameters, i.e., throughput, delay, and jitter; in most of the existing protocols network bandwidth is considered as the only resource in end-to-end performance evaluations.

Unfortunately, in reality, end user's QoS requirements may not be as transparent as assumed. Some requirements have to be translated and mapped to predefined measurable parameters (e.g., jitter). Resources critical to end user's QoS requirements may include power, CPU processing/response time, and local Input/Output (I/O), etc. Furthermore, affordable QoS provisioning application-layer softwares are expected to be portable and reusable on top of different COTS lower-layer protocols and hardware. All of these issues we are facing in the reality indicate that an *integrated seamless* end-to-end QoS architecture is not yet available. An innovative QoS-aware middleware is expected to play an important role in protocol integration and QoS architecture commercialization processes.

A few QoS-aware middleware systems, based on wireline networks, have been proposed and developed recently. Authors in [2] present a Global Resource Management System (GRMS) for QoS resource negotiation and adaptation across heterogeneous computing nodes and communication networks. Authors in [3] compare several existing QoS-aware middleware systems in terms of QoS specification, QoS translation, supported applications, QoS enforcement, and QoS adaptation.

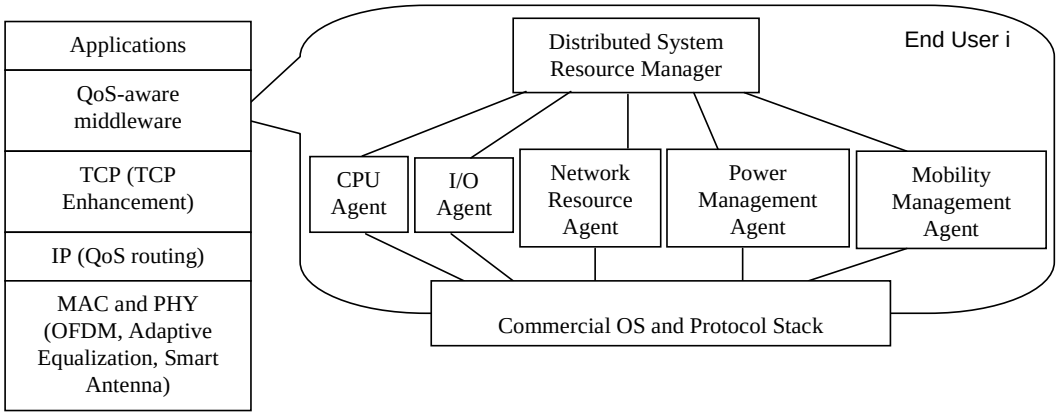

Figure 1: A QoS-aware protocol stack for wireless networks.

Wireless communication networks are expected to support real-time multimedia communication in the future. End user's QoS requirements have to be guaranteed in both wireline and wireless environments. The QoS-aware middleware therefore has to be extended to the wireless environment. Figure 1 shows a wireless network QoS-aware middleware architecture in an end host. This middleware runs on every end host in the network, and consists of the following major functional blocks. The notation *agent* used in the following description refers to the model of a resource management block consisting of *QoS negotiator*, *QoS allocator*, *QoS monitor*, and *QoS adapter*.

- *Distributed system resource manager* is a resource management agent for an end host. It is used to coordinate end-to-end resource negotiation and adaptation over underlying local resource agents. From a user's point of view, the distributed system resource manager accepts a service request, and then negotiates network resources with peer nodes and allocates local resources such as CPU and I/O.

- *CPU agent* is a local CPU resource management component. It is invoked by distributed system resource manager and performs QoS negotiation, QoS allocation, QoS monitoring, and QoS adaptation for each CPU resource request from applications.

- *I/O agent* is a local I/O resource management component. It is also invoked by system resource manager and performs QoS negotiation, QoS allocation, QoS monitoring, and QoS adaptation for each system I/O resource request from applications.

- *Network resource agent* is a network resource component. The network resource negotiation may be implemented by interfacing with the existing network resource reservation protocols, such as, RSVP signalling.

- *Power management agent* is a local power management component which may be used in mobile wireless hosts consuming battery energy. Without degrading system performance and application's QoS requirements, it dynamically scales the clock frequency and operating voltage of some computing components such as CPU through negotiation, monitoring and adaptation.

- *Mobility management agent* is responsible for continuing the QoS-negotiation or performing QoS-renegotiation without significantly interrupting the application. In a wireless network, a user may initiate resource reservations for an application, and then move to another location. The same QoS has to be delivered to the user at the new location. In order to achieve this, the QoS state and application execution state information of each end host has to be stored in the old location and retrieved in the new location.

Table 1: TCP throughput over IEEE 802.11 connections.

Connection	Data rate	TCP Throughput	Effective BW
IEEE 802.11	2 Mbps	0.98 Mbps	49%
IEEE 802.11b	11 Mbps	4.3 Mbps	39.1%

Wireless network QoS-ware middleware is an emerging research topic with many open questions. Little research work has been published in this area. Notable among them is a middleware called Mobiware (a testbed toolkit) [4], which is a software platform based on xbind and CORBA technology. It is designed to allow mobile multimedia applications to operate transparently during hand-offs and periods of persistent QoS fluctuation. Mobiware has been developed as a testbed toolkit. Readers are encouraged to read [4] for more details.

3 TCP Enhancements in Mobile Wireless Networks

QoS-aware middleware enables QoS management in wireless networks and facilitates easy deployment of underlying COTS operating systems and protocols. Application data, wrapped by QoS-aware middleware, are passed to the transport layer, e.g., TCP. TCP is supported by almost all existing network application programs. Currently, the vast majority of IP traffic is transmitted using TCP. The convergence of IP services with mobile wireless networks leads to various access methods to IP services, and the diversity of end-host computing devices. TCP will still be the dominant end-to-end reliable transmission control protocol in the evolution. However, TCP was initially designed for wired links and stationary hosts, where packet losses are mainly due to network congestion. TCP assumes that all packet losses are due to network congestion. When losses are detected, TCP drops its congestion window size, followed by retransmitting the lost packets, initiating congestion control or avoidance algorithms and backing off its retransmission timer. These actions result in a reduction of traffic load, thereby controlling the network congestion.

However, when wireless links are involved in the network connection, packet losses are mainly caused by link errors and hand-offs. TCP's unnecessary reduction of congestion window size decreases the network throughput, and increases the end-to-end delay. Table 1 [5] shows experimental results of TCP throughput over an IEEE 802.11 and an IEEE 802.11b wireless LAN. This shows the traditional TCP algorithm, if used in wireless environment, significantly degrades the network performance.

Several schemes have been proposed to improve the performance of TCP over wireless links. These can be classified into two approaches. In the first approach, the sender is *aware* of the existence of wireless links in the network, and attempts to either distinguish losses due to wireless links from those due to congestion, so the sender does not invoke congestion control algorithms when the packet loss is caused by wireless links [6], or quickly recover from packet losses. In the second approach, the TCP sender is *unaware* of the losses due to wireless links. The non-congestion related losses are hidden from the TCP at the fixed host (sender), and hence the TCP at the fixed host remains unmodified. In this section, we describe some proposed schemes based on wireless aware and unaware approaches. Please refer to [7] for a detailed analysis of transport layer design approach in mobile wireless networks.

3.1 Wireless aware TCP

In this approach, the fixed host (sender) is aware of the existence of wireless links and tries to either distinguish wireless link corruption losses from network congestion losses, or quickly recover from packet loss events. The following TCP extensions are based on this approach.

Table 2: Distinguishing congestion losses from corruption losses.

Algorithm	Method to Distinguish
TCP-Decoupling [13]	Sending TCP data packets and header packets in independent streams; congestion control is only applied to the header-packet stream.
TCP-Peach [14]	Sending dummy packets to probe the type of losses.
WTCP [15]	Measuring the inter-packet interval.
LEA [16]	Sender's receiving of either an acknowledgement packet, or an ICMP (Internet Control Message Protocol), or both.
ELN [17]	Explicitly setting the ELN bit in packet header whenever a non-congestion loss is detected.
Diff-C-TCP [18]	Optimally dimensioning ECN-capable RED gateway and notifying congestion losses with ECN.

- *Limited Transmit* [8]: This mechanism is effective in the cases of a large number of packet losses within a congestion window, or the congestion window size is small [9]. The Limited Transmit scheme extends Fast Retransmit and Fast Recovery algorithms [10] for TCP flows with small congestion windows that are not likely to generate three duplicate acknowledgements to trigger Fast Retransmit. Using Limited Transmit, if there are unsent packets in the sender's queue, the sender sends a new packet in response to the arriving of each of the first two duplicate acknowledgements. Authors in [8] have shown that over half of a busy server's retransmissions were due to the expiration of TCP retransmission timer. Furthermore, roughly 25 percent of these retransmissions could have been avoided using Limited Transmit.

- *Selective Acknowledgements* [11]: Using Selective Acknowledgements (SACKs) based algorithms, the sender can be precisely informed which packets need to be retransmitted in the first RTT (Round Trip Time) following the loss event. In this way, SACK allows TCP to recover from multiple segment losses in a window of data within one RTT of loss detection. Although Fast Retransmit, Fast Recovery and SACK are generally able to rapidly recover from multiple packet losses, they reduce the congestion window to avoid further congestion. The above behavior, which is based on the assumption that packet losses are indicators of congestion, results in the degradation of throughput in the presence of non-congestion related packet losses (such as wireless link errors). Therefore, when they are applied to wireless links, where most of packet losses are due to link errors instead of congestion, TCP is unable to determine the available bandwidth.

- *Distinguishing congestion losses from corruption losses*: This method makes the congestion window behave differently in the presence of congestion losses and corruption losses (due to link errors, hand-offs and fadings) by distinguishing the two types of losses. Many algorithms have been proposed using this method. They are concluded in Table 2. Authors in [12] provide an comparison for some of the algorithms.

3.2 Wireless unaware TCP

This approach is based on the intuition that since the problem is local, it should be solved locally, and TCP should be independent of the behavior of individual links. We present below some schemes based on this approach.

- *Snoop* [19] and *Delayed Duplicate Acknowledgements (DDA)* [20]: The Snoop algorithm assumes that the wireless link is the last hop in the TCP connection, and introduces a module, named the *snoop agent* at the base station. The agent caches TCP packets sent across the link that have not yet been acknowledged by the receiver. The agent retransmits the lost packet (if it has been cached) and suppresses the duplicate ACKs for TCP packets lost and retransmitted locally, thereby avoiding unnecessary fast retransmissions and congestion controls by the sender. However, it requires a base station to maintain the state information, and cache unacknowledged TCP packets, which results in the scalability issue.

  DDA attempts to imitate the behavior of Snoop by using link-layer retransmission. However, DDA tries to reduce the interference between TCP-layer retransmissions and link-layer retransmissions by delaying the third and subsequent duplicate packets for an interval of d. If the receiver receives out-of-order packets, it responds to the first two out-of-order packets by sending duplicate packets immediately.

- *Indirect-TCP (I-TCP)* [21]: The scheme breaks the connection between the fixed wired network and the wireless mobile host into two connections. One connection is between the fixed host and the base station; the other connection is between the base station and the wireless host. Data sent to the wireless host is first received by the base station. Upon receiving the data, the base station sends an acknowledgement to the fixed host and then the received data is forwarded to the wireless host. The base station and the wireless host does not need to use TCP for communication. Instead a specialized protocol that is optimized for mobile applications and for low speed and unreliable wireless medium can be used. This indirection helps shield the wired network from the uncertainties of the wireless network. However, I-TCP may violate the acknowledgement mechanism of the current TCP, because acknowledgements of data packets would possibly reach the original source before the data packets reach the wireless host.

- *M-TCP* [22]: This architecture was proposed for cellular networks to support high bandwidth, frequent hand-offs services. The architecture can be viewed as a three-level hierarchy. Mobile hosts which communicate with mobile stations in each cell are at the lowest level. Several mobile stations are controlled by a supervisor host at the second level. Supervisor hosts are connected to the high-speed wired network at the highest level and handles most of the routing and other protocol details for mobile users. M-TCP is used for the communication between mobile hosts and mobile stations. When the mobile station receives data from the sender, it forwards it to the wireless host but defers the ACK to the sender until it receives an ACK from the mobile host. If a mobile host undergoes a hand-off or a period of data losses, the mobile station sends the deferred ACK and advertises the window size of zero, which leads the sender to a *persist state*. During this period, all timers are frozen until the mobile host regains the connection. This algorithm provides a solution to the problem of frequent and periodic disconnection.

- *Freeze-TCP* [23]: The main design goal of Freeze-TCP is to handle hand-off disconnections. It is easy for a mobile host to monitor signal strengths, detect an impending handoff, and even predict a temporary disconnection. Therefore, the idea of Freeze-TCP is to modify the TCP algorithm at the mobile host so that the base station can be prevented from sending packets during hand-offs. If a handoff occurs, the mobile host sets a zero congestion window size to force the sender to enter the frozen mode and to prevent it from dropping its congestion window size.

3.3 Comparison of TCP enhancements

Table 3 compares the performance of major TCP enhancements, in terms of the following criteria:

- Is end-to-end semantics maintained?

Table 3: Comparison of different TCP enhancement schemes.

TCP Enhancement Schemes	End-to-end Semantics	Handle High BER	Handle Hand-off Disconnects	Distinguish Losses	Modify Current TCP
Limited Transmit	√	√			√
SACK	√	√			√
TCP-Decoupling	√	√		√	√
TCP-Peach	√	√		√	√
WTCP	√	√		√	√
LEA	√	√		√	√
ELN	√	√		√	√
Diff-C-TCP	√	√		√	√
Snoop	√	√			
DDA	√	√			
I-TCP		√	May run out of buffer		
M-TCP	√	√	√		
Freeze-TCP	√		√		√

- Is it able to handle high BER?
- Is it able to handle hand-off disconnections?
- Is it a loss-distinguishing scheme?
- Is it a modification of existing TCP?

As seen in Table 3, only I-TCP, M-TCP, and Freeze-TCP are able to handle hand-offs in mobile environment. Furthermore, only M-TCP is able to handle both high BER and hand-offs.

4 Mobility Management at IP Layer

A conventional wireless network has a predefined infrastructure and centralized administration. However, the availability of lightweight, portable computing devices, and wireless communication medium, has made mobile computing practical. Although the TCP enhancements described in Section 3 provide a reliable and efficient way to transmit data in wireless networks, forwarding IP packets and maintaining network connectivity in a mobile wireless environment is a key problem yet to be solved.

A network consisting of mobile wireless end hosts may have a dynamic topology, which requires frequent routing information updates. An Ad Hoc network is a good example of dynamic topology which is being actively studied. An *Ad Hoc* wireless network is formed by a group of mobile nodes interconnected by wireless links. Nodes communicate with each other either directly or through other nodes operating as routers. Several industry standards, such as, IEEE 802.11b, BlueTooth, HiperLAN2 have defined Ad Hoc as one of their network infrastructures. Potential applications of mobile Ad Hoc wireless networks include battlefield communications and coordinations, and sensor networks. In this section, we use Ad Hoc networks as an example to describe mobility enhancements at IP layer.

4.1 Characteristics of Ad Hoc networks and QoS requirements

A mobile Ad Hoc network is a wireless network, and hence is characterized by limited bandwidth, high link errors due to effects such as multipath fading, and limited security. A mobile Ad Hoc wireless network is characterized by:

- *Dynamic Topology*: Networks are self-creating; mobile nodes are free to join, leave and move. The topology may change randomly and at unpredictable time.
- *Multi-hop*: Due to the limitation of radio propagations, a mobile node may not be able to communicate with the other node directly. Similar to wireline networks, it needs other intermediate nodes to relay its messages. This gives rise to effective routing issues in Ad Hoc networks.
- *Power-constrained Communication*: Mobile nodes in an Ad Hoc network require batteries as their energy source. In order to provide reliable communication, power efficiency management schemes are very important for Ad Hoc network protocol designs.

Applications, such as real-time multimedia and battlefield coordinated network, require Ad Hoc networks to provide low-delay, small-jitter and bandwidth guaranteed communication services. In other words, an Ad Hoc network should provide QoS guarantees to end users. In order to implement an Ad Hoc network with dynamic infrastructure, mobile nodes, and multi-hop communications, the routing protocol becomes very important. In addition to detecting changes of network topology, maintaining connectivity, and performing packet routing, it has to optimize the utilization of network resources. In other words, QoS-enabled routing schemes are required for Ad Hoc networks.

4.2 Routing in Ad Hoc Networks

Many routing algorithms have been proposed in two areas: *Ad Hoc routing* and *QoS routing*. Most of Ad Hoc routing algorithms do not consider the requested QoS; most of proposed QoS routing algorithms were designed for wireline networks which have fixed infrastructures. QoS routing in Ad Hoc networks is a new research area and has become very attractive in the recent several years, and yet remains as an open issue. In this section, we describe Ad Hoc routing schemes, including those taking into account QoS.

4.2.1 Ad Hoc routing

Unlike traditional wireless networks with a predetermined infrastructure, Ad Hoc networks require the routing algorithm to be able to react efficiently to dynamic topology changes, i.e., the routing algorithm should be designed to frequently update the topology changing information and compute the new route. Among existing Ad Hoc routing schemes, there are two different approaches: topological routing and geographical routing.

Topological routing uses existing link information in the network to perform packet forwarding. They could be further divided into three groups:

- *Proactive algorithms* take advantage of classical routing strategies, such as, link-state routing (e.g., OLSR [24] and TBRPF [25]), and distance-vector routing (e.g., DSDV [26]). They update the routes continuously so that routes are already known when packets need to be forwarded. Proactive algorithms have lower latency since routes are maintained at all times. On the other hand, maintaining routes at all times may result in higher control overhead.
- *On-demand algorithms* compute routes only if needed. Therefore, it has higher latency and lower overhead than proactive algorithms. Frequent network topology changes result in significant

Table 4: Comparison of geographical routing protocols.

Protocol Name	Greedy	DREAM	LAR	Terminodes	Grid
Strategy	Greedy	Restricted directional flooding	Restricted directional flooding	Hierarchical	Hierarchical
Comm complexity	$O(\sqrt{n})$	$O(n)$	$O(n)$	$O(\sqrt{n})$	$O(\sqrt{n})$
Requires all-for-all location service	No	Yes	No	No	No
Robustness	Medium	High	High	Medium	Medium
Implementation complexity	Medium	Low	Low	High	High

amount of traffic, although on-demand algorithms compute routes only if needed. Examples of on-demand algorithms are DSR [27], TORA [28], and AODV [29].

- The third one is the *hybrid* of the proactive algorithm and on-demand algorithm, such as ZRP [30], in order to achieve high efficiency and scalability.

Geographical routing uses positioning information of the destination and the neighbors of forwarding nodes to determine packet forwarding routes. A typical geographical routing scheme performs packet forwarding in two steps: determine positions of all participating nodes, and then decide packet forwarding routes based on the positioning information. Geographical routing schemes therefore do not require the establishment or maintenance of routes. Nodes do not have to store routing tables, nor to exchange up-to-date routing information. This provides high level of scalability even if the network is highly dynamic.

Authors in [31] evaluate five geographical routing protocols which are summarized in Table 4, where *strategy* describes the fundamental strategy of the protocol, and *comm complexity* quantitatively measures the average number of hops required for a packet transmission. This assumes that the destination's position is known during packet forwarding. *Requires all-for-all location service* indicates whether or not a protocol needs a all-for-all location service. Location service is the process to identify the current position of a specific node. All-for-all location service means that in order to locate a node, all nodes in the network have to be able to perform location service, and each location service server has to maintain geographical information of all nodes in the network. *Robustness* indicates how a single intermediate node failure in the network affects the packet forwarding. *Implementation complexity* indicates the level of complexity to implement a protocol.

4.2.2 QoS-enabled Ad Hoc routing

The routing algorithms described in Section 4.2.1 were designed to support only best-effort traffic in Ad Hoc networks; they are not able to provide QoS to end mobile users. QoS routing in Ad Hoc networks is a way of selecting a packet forwarding path that better accommodates the requested QoS by end mobile users in multi-hop dynamic-topology networks. *Bandwidth-constrained* routing and *delay-constrained* routing are the most studied QoS-based Ad Hoc routing algorithms to date, though QoS metrics are not limited to only bandwidth and delay. Authors in [32] proposed a distributed QoS routing algorithm which determines a packet forwarding path with sufficient resources in a dynamic

multi-hop mobile environment. Their approach, called *ticket-based probing*, is a general QoS-based Ad Hoc routing approach, which is able to satisfy either a certain delay or bandwidth requirement. A ticket is the permission to search a path. Probes are routing messages. Flows with more critical requirements are issued more tickets. Probes are sent by the source and forwarded towards the destination to search for a low-cost path that satisfies the QoS requirement. Each probe carries at least one ticket. Probes that carries more than one tickets are split into multiple ones, each of which searches a different path. The maximum number of probes is bounded by the number of total issued tickets. By changing the QoS metrics and the corresponding ticket distribution methods, this framework is able to handle either a bandwidth-constrained requirement or a delay-constrained requirement.

Algorithms proposed by other researchers which can only handle a single QoS requirement are described in [32]. Readers interested to know more details in this topic are encouraged to read [32] and [33].

5 Enhancing PHY and LINK layers

Data and network control packets from sender's higher layer protocols (such as what we have presented in Sections 2, 3, and 4) are finally passed to the physical layer which forwards them to the receiver. Performance of mobile hosts are easily affected by impairments of signal transmission environments, e.g., fading due to multipath propagation. Multipath propagation can cause fluctuations in the received signal's amplitude, phase and the angle of arrival, which yields transmission errors leading to packet losses. It degrades the performance of higher-layer QoS protocols by increasing failure rate, decreasing throughput, and increasing the end-to-end delay, etc. Solutions are therefore required to mitigate multipath fading in wireless networks. A reliable PHY layer for mobile wireless networks is necessary for effective functioning of higher-layer protocols.

In a mobile wireless environment, when electromagnetic waves reflects off or diffracts around objects, a signal may travel between the transmitter and the receiver over multiple paths, which is referred to as *multipath* propagation. Multipath propagation can cause fluctuations in the received signal's amplitude, phase and angle of arrival, giving rise to *multipath fading*. Wireless network subscribers in small office/home office (SOHO) can experience severe multipath fading, which yields bit errors (packet losses in higher layers). This degrades the wireless network performance, such as throughput and end-to-end delay. Table 5 [34] shows various fading channel models classified by environments to which they apply. In the rest of this section, some anti-multipath approaches used to enhance wireless link capacity and error performance are presented.

Multipath fading can result in irreducible errors in system performance. Figure 2 [35] highlights three major performance categories in terms of bit error probability, P_B, versus signal-to-noise ratio, E_b/E_0. Among all curves, the topmost curve represents the worst performance, where the bit error probability can approach 0.5. No value of signal-to-noise ratio can help achieve better bit error probability.

Generally speaking, for such a system, two steps are taken to improve the performance. First, signal distortions as a result of multipath fading has to be reduced or removed. Once the distortion has been removed, the P_B versus E_b/N_0 curve should have moved from the upmost curve to the middle exponential one. Next step is to use some diversity schemes to strive approaching the leftmost Additive White Gaussian Noise (AWGN) performance in Figure 2. The term *diversity* refers to methods of providing a receiver with a collection of uncorrelated samples of the signal.

5.1 Combat signal distortions

General approaches used to mitigate signal distortion include:

Table 5: Models that can be used to characterize various wireless environments.

Environment	*Channel Type*
Mobile systems with no LOS path between transmitter and receiver antenna, propagation of reflected and refracted paths through troposphere and ionosphere, ship-to-ship radio links.	Rayleigh
Satellite links subject to strong ionospheric scintillation.	Nakagami-q (Hoyt) (spans range from one-sided Gaussian (q=0) to Rayleigh (q=1))
Propagation paths consisting of one strong direct LOS component and many random weaker components - microcellular urban and suburban land mobile, picocellular indoor and factory environments.	Nakagami-n (Rice) (spans range from Rayleigh (n=0) to no fading (n=∞))
Often best fit to land mobile, indoor mobile multipath propagation as well as ionospheric radio links.	Nakagami-m (spans range from one-sided Gaussian (m=$\frac{1}{2}$), Rayleigh (m=1) to no fading (m=∞))
Caused by terrain, buildings, trees - urban land mobile systems, land mobile satellite systems.	Log-Normal shadowing
Nakagami-m multipath fading superimposed on log-normal shadowing. Congested down town areas with slow-moving pedestrians and vehicles. Also in land mobile systems subject to vegetative and/or urban shadowing.	Composite gamma/log-normal
Convex combination of unshadowed multipath and a composite multipath/shadowed fading. Land mobile satellite systems.	Combined (time-shared) shadowed /unshadowed

- *Adaptive equalization* [35, 36]: This is a traditional way of using an adaptive filter to compensate for intersymbol interference (ISI) by gathering the dispersed symbol energy back together into its original time interval. The equalization process can be implemented in either time domain or frequency domain. This gives rise to the single carrier modulation with time domain equalization (SC-TDE) and the single carrier modulation with frequency domain equalization (SC-FDE) at the receiver. For channels with severe delay spread, the computation of SC-FDE is simpler than the corresponding SC-TDE. Figure 3 [36] shows a comparison of computation complexity between SC-TDE and SC-FDE. The complexity here is measured by the number of complex multiplication operations per transmitted data symbol.

- *Orthogonal Frequency-division Multiplexing (OFDM)* [35]: This method is used to avoid the use of equalization by lengthening the symbol duration. OFDM transmits multiple modulated subcarriers in parallel, each of which occupies only a very narrow bandwidth. The symbol rate of each sub-band is lower than that of the overall band. OFDM has been selected for IEEE 802.11a and European Telecommunication Standards Institute (ETSI) HiperLAN2.

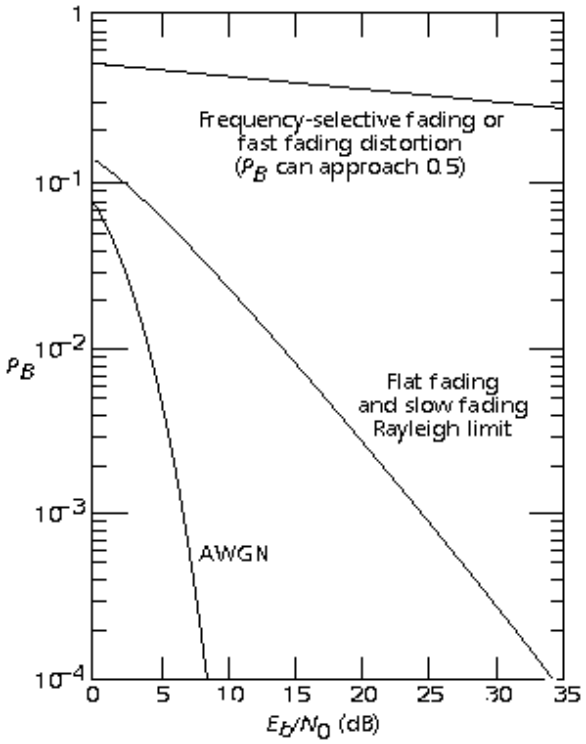

Figure 2: Three major error performance categories.

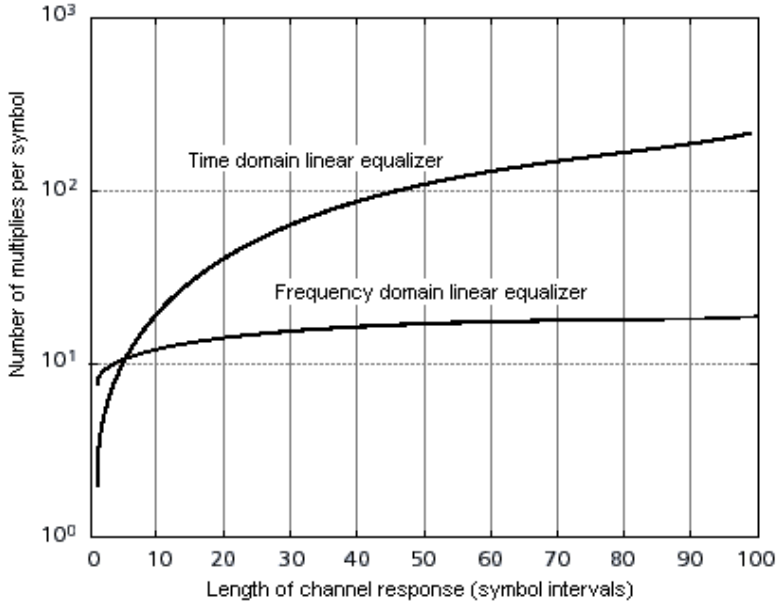

Figure 3: The computation complexity comparison between TDE and FDE.

- *Spread spectrum* [35]: Spread Spectrum (SS), a wideband technique, using either Frequency Hopping Spread Spectrum (FHSS) or Direct Sequence Spread Spectrum (DSSS), spreads a signal's power over a wide frequency spectrum in order to achieve good signal-to-noise performance. Interference from other wireless transmission and electrical noise, typically narrow in bandwidth, only interferes with a small portion of the SS signal. This unique nature makes the signal much less susceptible to any interference including ISI. Both DSSS and FHSS have been widely used in commercial products, such as wireless LAN, wireless home and building control, etc.

- *Ultra-wide bandwidth (UWB) technology* [37]: This relatively new term is used to describe an old technology which had been known since the early 1960's as carrier-free, baseband, or impulse transmission technique. A UWB system transmits and receives extremely short pulses whose duration is typically a few tens of picoseconds to a few nanoseconds, leading to an extremely wide spectrum. With very short pulses, the direct path comes and goes before the reflected path arrives, thereby avoiding multipath cancellation. Extensive experimental measurements have been performed in the dense multipath environment (indoor, modern office building). Results show that UWB signals do not suffer multipath fading, and therefore, very little fading margin is required to guarantee reliable communication [37]. UWB technology was approved on February 14, 2002 by Federal Communication Commission (FCC) under Part 15 of its regulation. The new rules for unlicensed UWB operation permits the applications related to imaging, vehicular radar, communications and measurement systems.

5.2 Diversity schemes

As described in Section 5, the anti-multipath approach should follow two steps: first, combat signal distortions; second, provide diversity. This section provides an overview of some diversity schemes. The *conventional antenna diversity* has been in commercial use in most of wireless communication systems for many years. However, in the presence of severe multipath interference, conventional diversity schemes is not able to improve the system performance. In such a case, *Smart antenna or adaptive antenna* is used to shape the antenna radiation pattern, enhancing the desired signals and eliminating the effect of interfering ones.

- *Conventional antenna diversity*: A diversity scheme is a method that transmits and receives signals from two or more uncorrelated antennas, resulting in independent fading. Therefore, it is likely that at least one antenna does not experience faded signals, while others are experiencing them. Typical schemes for providing uncorrelated antenna signals include space diversity (use multiple physically separated antennas), polarization diversity (use a dual antenna system with orthogonal polarizations), angle diversity (use multiple directional antennas receiving signals with different angle of arrivals), frequency diversity (transmit and receive signals at different carrier frequencies), and time diversity (transmit and receive data on multiple different time slots whose time separation are large enough) (please see [35, 36] for detailed description of each scheme).

- *Smart antenna*: A smart antenna is a combination of antenna array and innovative digital signal processing capability to optimize radiation and reception pattern adaptively in response to a signal environment. Smart antennas are categorized as either *adaptive array* or *switched beam*. A switched beam has multiple fixed beams with predefined patterns. Only one beam pattern among all candidates is chosen to be turned on at each time instant towards the desired signal as the mobile host moves throughout the coverage area. The beam pattern may change multiple times per symbol. An adaptive antenna system, which is the most advanced smart antenna solution to date, combines the adaptive signal processing algorithm to effectively track

Table 6: Smart antenna's features and benefits.

Features	*Benefits*
Signal Gain - inputs from multiple antennas are combined to optimize available power required to establish given level of coverage.	Better Range/Coverage - focusing the energy sent out into the cell increases base station range and coverage. Lower power requirements also enable a greater battery life and smaller/lighter handset size.
Interference Rejection - antenna pattern can be generated toward cochannel interference sources, improving the signal-to-interference ratio of the received signals.	Increased Capacity - precise control of signal nulls quality and mitigation of interference combine to frequency reuse reduce distance (or cluster size), improving capacity. Certain adaptive technologies (such as space division multiple access) support the reuse of frequencies within the same cell.
Spatial Diversity - composite information from the array is used to minimize fading and other undesirable effects of multipath propagation.	Multipath Rejection - can reduce the effective delay spread of the channel, allowing higher bit rates to be supported without the use of an equalizer.
Power Efficiency - combines the inputs to multiple elements to optimize available processing gain in the downlink (toward the user).	Reduced Expense - lower amplifier costs, power consumption, and higher reliability will result.

the mobile target, dynamically maximizing the desired signal and minimizing interference. Table 6 (taken from International Engineering Consortium smart antenna online tutorial) shows the features of and benefits derived from a smart antenna system.

6 Summary

We have presented a comprehensive review of the rapidly growing research area of enhancing IP services over mobile wireless networks. Our discussion is based on a multi-layer protocol stack for wireless networks. Issues and solutions in middleware layer, transport layer, network layer, MAC and physical layer have been discussed, and many open issues in this emerging research area have been highlighted.

A number of network architectures and protocols have been developed and standardized to provide end-to-end QoS in recent years. However, these protocols rely on the translation and mapping of application QoS requirements to QoS parameters defined by IETF. Affordable QoS provisioning application-layer softwares are expected to be portable and reusable on top of different COTS lower-layer protocols and hardware. All of these engineering requirements can be satisfied by wireless QoS-enabled middleware, due to its unique ability of abstracting the heterogeneity of operating system, networks, hardware, and even programming languages.

TCP was initially designed to perform well in networks with reliable wired links and stationary hosts, where packet losses are mainly due to network congestion. However, in a wireless environment, TCP's unnecessary reduction of congestion window size decreases throughput, and increases end-to-end delay, thereby degrading the quality of network services. Two categories of TCP enhancement

algorithms, i.e., TCP aware and TCP unaware, are presented in this article.

A mobile Ad Hoc wireless network frequently changes its topology, and packets are forwarded by intermediate nodes to reach their destinations. Therefore, routing algorithms are designed to frequently update the topology change and compute a new route. Both topological and geographical routing approaches are described in this article. Furthermore, QoS routing selects a packet forwarding path that better accommodates the requested QoS by end mobile users in multi-hop dynamic-topology networks. Bandwidth-constrained routing and delay-constrained routing are the most studied QoS-based Ad Hoc routing algorithms to date.

Information from a sender's higher layer protocols are passed to the physical layer which forwards them to the receiver. At the physical layer, multipath propagation can cause fluctuations in the received signal's amplitude, phase and angle of arrival, resulting in packet losses at higher layers due to bit errors in the physical layer. This degrades wireless network performance, such as throughput and end-to-end delay. The anti-multipath approach should follow two steps: first, combat signal distortions; second, provide diversity. Some conventional and advanced anti-multipath approaches have been discussed.

References

[1] H. Bai and M. Atiquzzaman, "Error modeling schemes for fading channels in wireless communications: A survey," *IEEE Communications Surveys and Tutorials*, vol. 5, no. 2, pp. 2–9, October 2003.

[2] J. Huang, Y. Wang, and F. Cao, "On developing distributed middleware services for QoS- and criticality-based resource negotiation and adaptation," *Journal of Real-time Systems*, vol. 16, no. 2-3, pp. 187–221, February 1999.

[3] K. Nahrstedt, D. Xu, D. Wichakakul, and B. Li, "QoS-aware middleware for ubiquitous and heterogeneous environments," *IEEE Communication Magazine*, vol. 39, no. 11, pp. 140–148, November 2001.

[4] O. Angin, A.T. Campbell, M.E Kounavis, and R.R.-F Liao, "The Mobiware toolkit: Programmable support for adaptive mobile networking," *IEEE Personal Communication Magazine*, vol. 5, no. 4, pp. 32–44, August 1998.

[5] G. Xylomenos, G. C. Polyzos, P. Mahonen, and M. Saaranen, "TCP performance issues over wireless links," *IEEE Communication Magazine*, vol. 39, no. 4, pp. 52–58, April 2001.

[6] H. Balakrishnan, S. Seshan, E. Amir, and R. H. Katz, "A comparison for improving TCP/IP performance over wireless links," *ACM SIGCOMM*, Palo Alto, CA, pp. 256–269, August 1996.

[7] H. Bai, S. Fu, and M. Atiquzzaman, "Transport layer design in mobile wireless networks," *Design and Analysis of Wireless Networks* (Yi Pan and Yang Xiao, eds.), Nova Science Publishers, 2005.

[8] M. Allman, H. Balakrishnan, and S. Floyd, "Enhancing TCP's loss recovery using limited transmit." RFC 3042, January 2001.

[9] H. Inamura, G. Montenegro, R. Ludwig, A. Gurtov, and F. Khafizov, "TCP over second (2.5G) and third (3G) generation wireless networks." RFC 3481, February 2003.

[10] W. Richard Stevens, "TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithm." RFC 2001, January 1997.

[11] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, "TCP selective acknowledgment options." RFC 2018, October 1996.

[12] S. Cen, P. C. Cosman, and G. M. Voelker, "End-to-end differentiation of congestion and wireless losses," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 703–717, October 2003.

[13] S. Y. Wang and H. T. Kung, "Use of TCP decoupling in improving TCP performance over wireless networks," *ACM Wireless Networks*, vol. 7, no. 3, pp. 221–236, May 2001.

[14] I. F. Akyildiz, G. Morabito, and S. Palazzo, "TCP-Peach: A new congestion control scheme for satellite IP networks," *IEEE/ACM Transactions on Networking*, vol. 9, no. 3, pp. 307–321, June 2001.

[15] P. Sinha, N. Venkitaraman, R. Sivakumar, and V. Bharghavan, "WTCP: A reliable transport protocol for wireless wide-area networks," *Proc. ACM/IEEE MOBICOM*, Seattle, WA, pp. 231–241, August 1999.

[16] S. Goel and D. Sanghi, "Improving performance of TCP over wireless links," *Proc. IEEE TENCON*, pp. 332–335, December 1998.

[17] H. Balakrishnan and R. Katz, "Explicit loss notification and wireless web performance.," *Proc. IEEE Globecom Internet Mini Conference*, Sydney, Australia, November 1998.

[18] H. Bai and M. Atiquzzaman, "Enhancing TCP throughput over lossy links using ECN-capable RED gateway.," *Proc. IEEE 58th Vehicular Technology Conference*, Orlando, FL, October 2003.

[19] H. Balakrishnan, S. Seshan, and R. Katz, "Improving reliable transport and handoff performance in cellular wireless networks," *ACM Wireless Networks*, vol. 1, no. 4, pp. 469–481, December 1995.

[20] N. H. Vaidya, M. Mehta, C. Perkins, and G. Montenegro, "Delayer duplicate acknowledgements: A TCP-Unaware approach to improve performance of TCP over wireless," Technical report 99-003, Computer Science, Texas A&M University, February 1999.

[21] A. Bakre and B. R. Badrinath, "I-TCP: Indirect TCP for mobile hosts," *15th International Conference on Distributed Computing Systems*, Vancouver, Canada, pp. 136–143, June 1995.

[22] K. Brown and S. Singh, "M-TCP: TCP for mobile cellular networks," *Computer Communication Review*, vol. 27, no. 5, pp. 19–43, October 1997.

[23] T. Goff, J. Moronski, D. S. Phatak, and V. Gupta, "Freeze-TCP: A true end-to-end enhancement mechanism for mobile environments," *Proc. IEEE INFOCOM*, Tel Aviv, Israel, pp. 1537–1545, March 2000.

[24] T. Clausen and P. Jacquet, "Optimized link state routing protocol." draft-ietf-manet-olsr-10.txt, May 2003.

[25] R. Ogier, M. Lewis, and F. Templin, "Topology broadcast based on reverse path forwarding (TBRPF)." draft-ietf-manet-tbrpf-08.txt, April 2003.

[26] C. Perkins and P. Bhagwat, "Highly dynamic destination sequenced distance-vector routing (DSDV) for mobile computers," *Computer Communications Review*, pp. 234–244, October 1994.

[27] D. B. Johnson, D. A. Maltz, and Yih-Chun Hu, "The dynamic source routing protocol for mobile ad hoc networks (DSR)." draft-ietf-manet-dsr-09.txt, April 2003.

[28] Vincent D. Park and M. Scott Corson, "A highly adaptive distributed routing algorithm for mobile wireless networks," *Proc. IEEE INFOCOM*, Kobe, Japan, pp. 1405–1413, April 1997.

[29] C. Perkins and E. Royer, "Ad-hoc on-demand distance vector routing," *Proc. The Second IEEE Workshop on Mobile Computing Systems and Applications*, New Orleans, LA, pp. 90–100, February 1999.

[30] Z. Haas and M. Pearlman, "The performance of query control schemes for the zone routing protocol," *ACM/IEEE Transactions on Networking*, vol. 9, no. 4, pp. 427–438, August 2001.

[31] M. Mauve, J. Widmer, and H. Hartenstein, "A survey on position-based routing in mobile ad hoc networks," *IEEE Network*, vol. 15, no. 6, pp. 30–39, November/December 2001.

[32] S. Chen and K. Nahrstedt, "Distributed quality-of-service routing in ad hoc networks," *IEEE JSAC*, vol. 17, no. 8, pp. 1488–1505, August 1999.

[33] S. Chakrabarti and A. Mishra, "QoS issues in ad hoc wireless networks," *IEEE Communication Magazine*, vol. 39, no. 2, pp. 142–148, February 2001.

[34] M. K. Simon and M. S. Alouini, "A unified approach to the performance analysis of digital communication over generalized fading channels," *Proc. IEEE*, vol. 86, no. 9, pp. 1860–1877, September 1998.

[35] B. Sklar, "Rayleigh fading channels in mobile digital communication systems part II: Mitigation," *IEEE Communication Magazine*, vol. 35, no. 7, pp. 102–109, July 1997.

[36] D. Falconer, S. L. Ariyavistakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Communication Magazine*, vol. 40, no. 4, pp. 58–66, April 2002.

[37] M. Z. Win and R. A. Scholtz, "On the robustness of ultra-wide bandwidth signals in dense multipath environments," *IEEE Communication Letters*, vol. 2, no. 2, pp. 10–12, February 1998.