TCP in Wireless Networks: Problems and Solutions

Sachin Gajjar and H. M. Gupta
Dept. of Electrical Engg., IIT Delhi
{sachin, hmgupta}@ee.iitd.ac.in

Abstract--**Recent years have witnessed the beginning of "The Wireless Revolution". There has been an explosive growth of wireless networks, ranging from Wireless Local Area Networks (WLANs) and Wireless Wide-Area Networks (WWANs) to Mobile Ad hoc Networks (MANETs). Such a fast growth in mobile computing and wireless networks gives a strong indication that mobile computing with wireless communication links will be an integral part of future internet works. The TCP has been widely used in today's Internet as transport layer protocol for providing reliable end-to-end data delivery. As in the Internet it is desirable that TCP must provide reliable data transfer services for communication within wireless networks and also between wireless networks and wired Internet. Thus it is very important that TCP performs well over all kinds of wireless networks so that wired Internet can be extended to wireless world. Unfortunately, standard TCP is not suitable for mobile hosts and their wireless links. This paper discusses the reasons associated with this unsuitability. Further it presents a survey of the research done so far to improve the performance of TCP over wireless networks. Finally a comparison of the proposed protocols is done.**

Index Terms-- **AIRMAIL, ATCP, BDP, base station, CWL, FR, I-TCP, mobility, mobile host, M-TCP, standard TCP, SNOOP.**

I. INTRODUCTION

The TCP has been widely used in today's Internet. It is a transport layer protocol which provides reliable end-to-end data delivery. This reliability is mainly achieved by retransmitting lost packets. TCP employs a retransmission timer that handles the retransmission time, which is the waiting time for an acknowledgement of a packet. When TCP sends a packet it creates a retransmission timer for that particular packet. Under this condition two situations may occur [1],

1. In case an acknowledgement is received for this packet before the timer goes off the timer is ignored.
2. If timer goes off before acknowledgement arrives, the segment is retransmitted and timer is reset.

The round trip time (RTT) is calculated dynamically. TCP sends a packet, starts a timer, and waits for an acknowledgement. It measures time between sending of packet and receiving an acknowledgement. Each packet has a RTT. The value of the RTT used in the calculation of the retransmission time of the next segment is the updated value of the RTT according to the Jacobsen's formula [1]:

$$\mathrm{RTT} = \alpha\,(\mathrm{previousRTT}) + (1-\alpha)\,(\mathrm{currentRTT}) \qquad (1)$$

The value α is usually 90 percent. Thus the new RTT is 90 percent of the value of the previous RTT plus 10 percent of the value of current RTT. Wired networks are inherently reliable so there is an implicit assumption made by TCP that any packet loss is due to congestion. To reduce congestion TCP invokes its congestion control mechanisms whenever any packet loss is detected. Unfortunately, there is a significant difference between wired networks and wireless networks in terms of bandwidth, propagation delay and link reliability. The main difference in wireless networks is that packet losses are no longer mainly due to network congestion. They may be due to other wireless network related reasons like high bit error rate in wireless channels, handoffs between two cells, transmission medium contention and route breakages. Thus even though TCP performs well in wired networks it suffers from severe performance degradation in case of wireless networks. It misinterprets such non congestion related losses as sign of congestion and invokes congestion control and avoidance mechanisms [4]. This results in decreasing of the throughput of the network.

Wireless networks can be broadly classified into two categories: 1. One-Hop 2. Multi-hop. For the first category only one hop wireless link is needed for communication between mobile host and stationary host in wired networks. In the second case there is no fixed infrastructure such as base stations or access points. The main problems imposed in using standard TCP for one-hop and multi-hop wireless networks are different and so are the solutions.

II. PROBLEMS IN USING STANDARD TCP IN ONE-HOP WIRELESS NETWORKS

One-Hop Wireless Networks, typically include Wireless LAN and Wireless Cellular Networks. There are certain inherent adverse features of One-Hop Wireless Networks which degrade the performance of TCP. These mainly include transmission channel errors, node's mobility, communication link's asymmetry and varying packet size [3]-[5].

1) Transmission channel errors

There is high Bit Error Rate (BER) in wireless channels. This is mainly due to multipath fading and shadowing which may corrupt the packets in transmission, causing the losses of TCP data packets or ACKs. If the TCP sender does not receive the ACK within the retransmission timeout (RTO), it immediately reduces its congestion window size to one segment, exponentially backs off its RTO and retransmits the lost packets. Eventually the channel errors may cause the congestion window size at the sender to remain small, resulting in low TCP throughput.

2) Node's Mobility

There are frequent handoffs in cellular networks due to node's mobility. Generally, handoffs result into temporary disconnections of the nodes, which results in packet losses and transmission delay. TCP performance deteriorates if it treats such losses as congestion and unnecessarily invokes congestion control mechanisms. Typically size of the cells is decreasing with the advance in cellular networks technologies resulting into large number of handoffs.

3) Communication link's asymmetry

The base station and the mobile terminal in nature are asymmetric. The mobile terminal has limited power, processing capability and buffer space. Also there is a vast difference in the characteristics of wired links and wireless links. Wired link is reliable, has large bandwidth while wireless link is error-prone and has limited and highly variable bandwidth

4) Varying packet size

Packet size over the wireless links is typically much smaller than that over the wired links. Hence each packet on the wired networks must be fragmented when transmitted over the wireless link. Finding the optimal packet size on the wireless link is one of the problems in using standard TCP for wireless networks.

III. SOLUTIONS PROPOSED FOR USING TCP IN ONE-HOP WIRELESS NETWORK

The solutions proposed for the improvement of the TCP performance for one-hop wireless networks can be categorized as: split-connection solution, proxy-based solutions, link-layer solutions, and end-to-end solutions [4], [5].

A. Split-connection solutions

In Split-connection solutions the TCP connection is split between the mobile host (MH) and the fixed host (FH) into two separate connections: a wired connection between the fixed host and the base station (BS), and a wireless connection between the BS and the MH. Thus, there is no need of changing the existing software on the FH in the wired connections. Also the wireless connection can be optimized by using a mobile protocol specialized to give better performance. Many split connection protocols are proposed, two of which are discussed.

a) Indirect TCP

Indirect-TCP or I-TCP protocol proposed by Bakre and Badrinath [7] suggests splitting the TCP connection between MH to fixed host FH on the wired network into two connections viz. the connection between the MH and its BS through the wireless medium and the other between the BS and the FH in the wired network. The arrangement is shown in the Fig. 1. Whenever a packet is sent to MH it is first received by BS, which sends an acknowledgement to FH and the packet is then forwarded to MH. When MH moves to a different cell while communicating with an FH, the entire connection information maintained at the current BS is transferred to the new BS and the new BS now forwards the packets without the awareness of the FH. TCP connection of wireless-link can be tailored to handle wireless channel shortcomings like the channel errors, handoffs and MHs weak features like limited battery and processing power. The two separate links help in identifying the reason for the packet losses i.e. congestion or anything else and thus take correct actions.

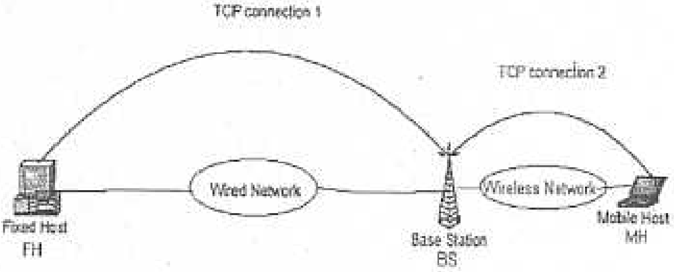

Fig. 1. Splitting of the TCP connection in I-TCP

b) M-TCP

The protocol proposed by Brown and Singh [8], uses a three-layer hierarchical architecture shown in Fig. 2. At the bottom layer, there are MHs and the BS in each cell. Several BSs are controlled by the supervisor host (SH) at the second layer. These hosts handle the routing and maintain the connections for the MHs. The SHs are connected to the wired network at the top level of hierarchy.

The operation of M-TCP follows. Assume that the MH has acknowledged packets up to sequence number Y, the SH sends an ACK for packets up to the Y-1 to the TCP sender only when it receives ACKs from the MH. If SH does not receive ACKs beyond Y for time being, SH assumes this to be a temporary wireless link breakage. Hence, it sends an ACK for the last byte Y with a zero window size. On receiving this, the sender enters into persist mode; it freezes all its transmission states like the RTO and congestion window. When the wireless link is re-established, MH notifies the SH by sending a hello packet. In turn, the SH, notifies the sender about the re-established connection, allowing the sender to resume its transmission from the frozen state. Thus the adverse effects of disconnections on TCP performance are eliminated as no congestion control is invoked.

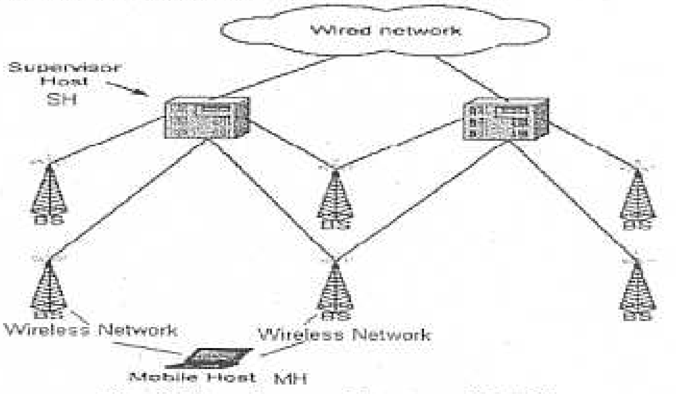

Fig. 2. Three-layer architecture of M-TCP

B. Proxy-based solutions

Such solutions put an implicit or explicit intelligent agent at the BS for detecting packet losses over wireless links and taking appropriate actions. These actions may be suppression and/or local retransmission of ACK which ensures that TCP sender responds correctly.

a) SNOOP

This protocol proposed by Balakrishnan et al. [6] modifies the network-layer at the BS to improve TCP performance. A SNOOP module is introduced at the BS which monitors every packet passing through connections in both the directions. This module maintains a cache of TCP packets sent from the fixed host that have not been acknowledged by the MH. Arrival of duplicate ACK or local timeout by RTT detects a packet loss. Lost packet is retransmitted by the SNOOP module if it is in the cache. Thus BS hides packet loss from FH, which avoids it to unnecessarily invoke congestion control mechanism. Further selective retransmissions from the BS can be added to improve performance. For the sender to decide whether the packet loss happened on the wireless or on the wired link due to congestion, track of the packets lost at the BS is kept. BS then generates negative acknowledgements for packets back to MH.

SNOOP is also designed to handle handoffs. Several BSs near a MH form a multicast group. They buffer some latest packets sent from the FH. Before handoff, MH sends control messages to determine that a BS with strongest signal should be the one forwarding the packets to MH and other BSs buffer packets. Thus handoff latency and packet losses both get reduced.

C. Link-layer solutions

This approach tries to solve the problem at the link layer and thus hides the characteristics of the wireless link from the transport layer. Thus a reliable link layer is built by adopting some link error recovery mechanisms.

a) AIRMAIL

A reliable link layer called AIRMAIL (Asymmetric Reliable Mobile Access In Link-layer) which can avoid unnecessary TCP congestion control invocation due to channel errors was proposed by E. Ayanoglu, et. al [9]. In it the famous link error recovery techniques, the foreword error correction (FEC) and automatic repeat request (ARQ) are employed. AIRMAIL uses asymmetric ARQ error control and window-based flow control techniques. Timers are at the BS during transmission as well as reception and hence all timers related operations are performed in the BS. The protocol requires the BS to send periodic status messages, while the MH combines several ACKs into a single to conserve power. The mobile ACK, unlike the BS, is event driven to reduce the processing load at the MH. There are three possible levels of FEC viz. bit-level achieved in hardware at the physical layer, byte-level done by per-packet cyclic redundancy check (CRC), and packet level done by allocating some packets for correction which are used to recover lost packets without retransmission. The protocol involves changes only at the link layer and takes into account the asymmetry between the BS and MH.

D. End-to-end solutions

a) Fast retransmission

The scheme proposed by Caceres and Iftode [10], uses the fast retransmission scheme employed in the current TCP implementation, to provide smooth handoffs on networks that lose packets during handoffs. In the changed TCP, MH signals the completion of handoff to the FH. TCP on FH then invokes the retransmission scheme. This forces the FH to reduce the congestion window to half and retransmit one packet immediately. This causes the connection to resume communication quickly as compared to regular TCP.

b) Selective acknowledgement

SACK-enabled sender retransmits in one RTT multiple lost packets in one transmission window and hence avoids continuous timeouts. . The drawback in this mechanism is it does not differentiate the reason for packet losses and assumes all losses are due to congestion. It is seen that SACK is useful over the error-prone wireless link where losses occur in bursts.

IV. PROBLEMS IN USING STANDARD TCP IN MULTI-HOP WIRELESS NETWORKS

Multi-Hop Wireless Networks typically include the WLANs like the Wi-Fi technology and WWANs like 2.5G/3G/4G cellular networks. In such type of networks there is no fixed infrastructure such as BSs or access points. Each node can move independently and function as router to discover and maintain route and forward packets to other nodes. The main problems include transmission channel error, hidden terminal problem, exposed terminal problem, node mobility and multi-path routing [4], [6].

1) Transmission channel errors

These are similar to those found in one-hop wireless networks. In fact they are more serious in this case since TCP connection now consist of multi-hop wireless links, where as in one-hop networks only the last hop is wireless. This results into dramatically shrinking of the congestion window at the sender due to channel errors in several wireless hops, further reducing the throughput of the network significantly.

2) Hidden station and Exposed station problems [2]

IEEE 802.11 MAC based protocol has been incorporated for wireless networks. In this protocol the neighboring nodes contend for the shared wireless channel before transmitting the packet. There are three problems viz. the hidden station problem, the exposed station problem and the unfairness.

Consider Fig.3, where four wireless stations are shown. The radio range is such that A and B are within each other's range and can interfere with one another. C can interfere with B and D, but not with A. Now suppose A is transmitting to B, as shown in Fig. 3(a). If at this instance of time C senses the medium, it will not hear A as A is out of its range, and falsely concludes that it can transmit. If C starts transmitting, it will interfere at B, corrupting the packet from A. The problem of station not being able to detect a contender for the medium because the contender is too far away is called hidden station problem.

On the other hand consider the situation where B is transmitting to A, as shown in Fig. 3(b). If C senses the medium, it will recognize a live transmission and falsely conclude that it will not be able to send to D, while in reality such a transmission would cause bad reception only in the area between B and C, where neither of the intended receivers is located. This problem is called exposed station problem.

The problem is that before starting a transmission, a station has to know whether or not there is transmission activity around the receiver. CSMA checks this activity around the station sensing the carrier. With wired network all signals propagate to all stations, as a result only one transmission can take place at a time anywhere in the system. In short-range radio waves based system, multiple transmissions can occur at the same time if all of them have different destinations which are out of range of each other.

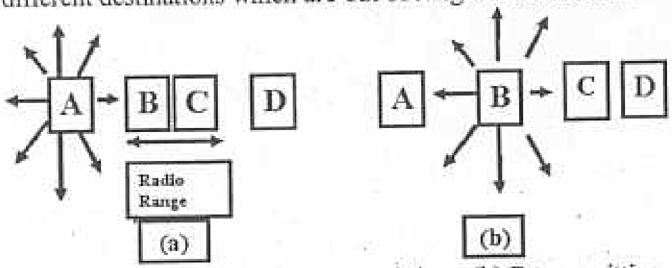

Fig. 3.A wireless LAN (a) A transmitting. (b) B transmitting.

3) Node's Mobility

As the nodes move there is link breakage and route failure between two neighboring nodes, when one of the node moves out of the transmission range of the other. This in turn causes packet losses. TCP understands this to be due to congestion and starts the congestion control algorithms.

4) Multi-path Routing

The routes in the MANETs are of short durations due to frequent link breakages. In order to reduce transmission delay due to route re-computation, some of the routing protocols like TORA maintain multiple routes between transmitter-receiver pair and use multi-path routing for packet transmission. Consecutively packets coming from different routes may not come at the receiver in order. TCP would misinterpret such out-of-order packet arrivals as a sign of congestion. The receiver will then generate duplicate ACKs which further cause the sender to invoke congestion control algorithms.

V. SOLUTIONS PROPOSED FOR USING TCP IN MULTI-HOP WIRELESS NETWORK

The solutions proposed for the improvement of the TCP performance for multi-hop wireless networks can be categorized as: TCP with feedback solutions, TCP without feedback solutions and TCP with lower layer enhancements.

A. TCP with feedback solutions

In this scheme feedback information is used to signal non-congestion-related causes. The feedback from the network layer helps TCP sender to distinguish between true network congestion and other problems like channel errors, link contention and route failures.

a) ATCP

Ad hoc TCP by Liu and Singh [11] uses the feedback from the network layer. The concept of the approach is to insert a thin layer called ATCP between IP and TCP, which ensures correct behavior in case of route failures and high BER. Sender decides an appropriate state by learning the network state information through explicit congestion notification (ECN) and ICMP, "Destination Unreachable" messages. On receiving the message, the sender enters the persist state. TCP at the sender is frozen and no packets are sent until a new route is found, so the sender doe not invoke congestion control unnecessarily. On receiving ECN, congestion control is invoked without waiting for time out to occur. If packet loss occurs and the ECN flag is not set, ATCP assumes the loss is due to bit errors and retransmits the lost packet. In case of multi-path routing, on receiving duplicate ACKs, TCP sender does not invoke congestion control, as multi-path routing shuffles the order in which segments are received. Thus the protocol works well when multi-path routing is applied.

B. TCP without Feedback Solutions

a) Adaptive Congestion Window Limit setting

Chen et al. [12] proposed an Adaptive Congestion Window Limit (CWL, measured in number of packets) setting approach to dynamically adjust TCP's CWL based on the current round-trip hop-count (RTHC) of the path which can be obtained from routing protocols like DSR. More precisely CWL should always be less than RHTC of the path.

In order to fully utilize the network capacity, TCP flow should set its CWL to the bandwidth-delay product (BDP) of the current path where BDP is equal to bandwidth of forward path x packet transmission delay in round trip. CWL < BDP to avoid network congestion. Now suppose for the ad hoc networks, S=size of data packet, b_{min} = bandwidth along the forward and return path. It can be seen that the delay at any hop along the path is less than the delay at the bottleneck link, i.e., S/b_{min}. As the size of TCP ACK is normally smaller than that of the data packet, by the definition of BDP, BDP <= RTHC x S. Thus the CWL, bounded by the path's BDP, should never exceed the RTHC of the path.

b) Lower Layer Enhancement Solutions

In this scheme the lower layer like the routing layer is tailored with for TCP's congestion control algorithms.

a) Routing Layer Enhancement

The mechanisms used are: Symmetric Route Pinning (SRP), Route Failure Prediction (RFP), and Proactive Route Errors (PRE), are used to minimize the probability of route failures, predict route failures in advance and to minimize the latency in conveying route failure information to source, respectively. In the first mechanism, the ACK path of a TCP connection is always kept the same as the data path to decrease the probability of route failure due to asymmetric path. The second mechanism enables the node to predict the occurrence of link failure base on the progression of signal strengths of packet receptions from the concerned neighbor. With the PRE all sources that have used the link in the past certain period are informed of link failure whenever it is detected. This reduces the latency involved in the route failure notification which reduces the number of packet losses and triggers early alternate route computations.

VI. COMPARISION OF CATEGORIES OF PROPOSED PROTOCOLS

Table I summarizes the problems handled by the protocols discussed in this paper, and their features. SACK protocol is not mentioned in it since it was basically designed for wired network, and hence it only satisfies the end-to-end semantics of TCP and the compatibility feature in the above table. From the table it can be concluded that the link layer protocols maintain the end-to-end semantic of TCP. They increase the reliability of the wireless link by decreasing the BER. They also fit naturally into the layered structure of network protocols. But these protocols do not take care of long or frequent disconnections.

End-to-end protocols suffer from major drawback that they require modification to TCP at the FH code and thus do not satisfy compatibility. But they do satisfy end-to-end semantics of TCP.

The major advantage of the split-connection protocols is that they provide backward compatibility with the existing wired network protocols and can handle disconnections efficiently. They shield the mobility and wireless issues from the FH. But they might not provide the end-to-end semantics of TCP.

In the case of solutions for multi-hop network it is found that ATCP is good at handling high BER, packet loss due to route breakage and identify true network congestion. Adaptive CWL makes effective use of channel bandwidth. Routing layer enhancement decrease the latency involved in route failure information delivery.

TABLE I

COMPARISION OF PROPOSED PROTOCOLS

	I-TCP	M-TCP	SNOOP	AIRMAIL	FR	ATCP	CWL	Routing layer enhancement
High BER	Y	N	Y	Y	N	Y	Y	Y
Bursty error	N	N	Y	Y	N	N	N	N
Handoff	Y	Y	Y	Y	Y	Y	Y	Y
Long disconnections	Y	Y	N	N	N	Y	Y	Y
Frequent Disconnections	Y	Y	N	N	N	Y	Y	Y
Bandwidth	N	Y	N	Y	N	N	Y	N
Cell size	N	N	N	N	Y	N	N	N
Power scarcity	N	Y	N	Y	N	N	N	N
Serial timeouts	Y	Y	N	N	N	N	N	N
Variable packetsize	N	N	N	N	N	N	N	N
End-to-end TCP semantics	N	Y	Y	Y	Y	Y	Y	Y
Standard TCP Compatibility	Y	Y	Y	Y	N	Y	Y	Y

VII. CONCLUSION

One of the most challenging and interesting recent trends in computer networks is the integration of wireless communications into the wired networks. With the increasing importance of host mobility, and the popularity of TCP/IP in wired networks, there is a requirement of reliable wireless TCP/IP protocol to be used in wireless networks. The wireless TCP/IP protocol should be designed by keeping in mind the problems of mobility and wireless connections. This paper presented a survey of different protocols that improve the performance of standard TCP in wireless networks. It can be concluded that the feature that need to be satisfied in a standard wireless TCP are:

Avoiding false triggering of congestion control algorithms and serial timeout problem in the FH, efficiently deal with handoff, frequent and long disconnections of the MH; consider the limited bandwidth and power scarcity of MH, variable packet size depending on bandwidth available for MH, provide end-to-end semantics of TCP and compatibility, i.e. no requirement in change of software of MH. Apart from all this issues TCP fairness is also an important issue because the bandwidth over the wireless link being very limited it is crucial for every flow to fairly share the bandwidth. It can be noted that compared with one-hop networks, multi-hop networks require more efforts to handle as things are much more complicated in their case. Finally, although many encouraging improvements have been proposed, none of them can work well in all scenarios and meet the challenges. Thus there is still much work needed including cross layer optimization.

VIII. REFERENCES

[1] Behrouz A. Forouzan, "Data Communications and Networking," 3rd edition, TMH Pub., Chapter 22, pp. 621-623, 2004.

[2] Andrew S. Tanenbaum, "Computer Networks," 3rd edition, PHI Pub., Chapter 4, pp. 264- 265,1998.

[3] Kaveh Pahlavan, Prashant Krishnamurthy, "Principles of wireless networks-A unified approach," Pearson Edu., Chapter 1, pp. 12-20, 2004.

[4] Xiang Chen, Hongqiang Zhai, Jianfeng Wang and Yuguang Fang, "A Survey on Improving TCP Performance over Wireless Networks". http://winet.ece.ufl.edu/~zhq/book05chen1.pdf.

[5] Hala Elaarag, "Improving TCP performance over mobile networks, "ACM Computing Surveys, Vol. 34, pp.357- 374, Sept. 2002.

[6] H. Balakrishnan, S. Seshan and R. H. Katz, "Improving reliable transport and handoff performance in cellular wireless networks," ACM Wireless Networks, pp. 469-481, Dec. 1995.

[7] A. Bakre and B. R. Badrinath, "I-TCP: indirect TCP for mobile hosts," Proc. 15th International Conf. On Distributed Computing systems (ICDCS), pp. 136-143, May 1995.

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

[9] E. Ayanoglu, S. Paul, T. F. LaPorta, K. K. Sabnani and R. D. Gitlin, "AIRMAIL: a link-layer protocol for wireless networks," ACM Wireless Networks, pp. 47-60, Feb.1995.

[10] R. Caceres and L. Iftode, "Improving the performance of reliable transport protocols in mobile computing environments," IEEE JSAC Vol.19 No.7, pp. 850-857, July 2001.

[11] J. Liu, S. Singh, "ATCP: TCP for mobile ad Hoc networks," IEEE JSAC. Vol.19 No.7, pp. 1300-1315 July 2001.

[12] K. Chen, Y. Xue, K. Nahrstedt, "On Setting TCP's Congestion Window Limit in Mobile Ad Hoc Networks," IEEE ICC'03, pp. 251-258, May, 2003.

Sachin Gajjar is currently doing his M.Tech. in Computer Technology from Indian Institute of Technology, Delhi. His areas of interest are computer communication and mobile computing.

H. M. Gupta is Professor with the Electrical Engineering Department, Indian Institute of Technology, Delhi. His research interests are in the areas of Communication Systems, computer communication, Photonic Systems and Multimedia Information Systems.