

HMP: Hotspot Mitigation Protocol for Mobile Ad hoc Networks

Seoung-Bum Lee and Andrew T. Campbell

COMET Group, Department of Electrical Engineering,
Columbia University, New York, NY 10027
{sbl, campbell}@comet.columbia.edu

***Abstract.** “Hotspots” represent transient but highly congested regions in wireless ad hoc networks that result in increased packet loss, end-to-end delay, and out-of-order packets delivery. We present a simple, effective, and scalable Hotspot Mitigation Protocol (HMP) where mobile nodes independently monitor local buffer occupancy, packet loss, and MAC contention and delay conditions, and take local actions in response to the emergence of hotspots, such as, suppressing new route requests and rate controlling TCP flows. HMP balances resource consumption among neighboring nodes, and improves end-to-end throughput, delay, and packet loss. Our results indicate that HMP can also improve the network connectivity preventing premature network partitions. We present analysis of hotspots, and detail the design of HMP. We evaluate the protocol’s ability to effectively mitigate hotspots in mobile ad hoc networks that are based on best effort on-demand routing protocols, such as, AODV and DSR.*

1. Introduction

Hotspots are often created in regions of mobile ad hoc networks (MANETs) where flows converge and intersect with each other. We define hotspots as nodes that experience flash congestion conditions or excessive contention over longer time-scales (i.e., order of seconds). Under such conditions nodes typically consume more resources (e.g., energy) and attempt to receive, process, and forward packets but the performance of the packet forwarding and signaling functions is considerably diminished and limited during the duration of hotspots. This is the result of excessive contention of the shared media wireless access, and due to flash loading at hotspot nodes, and importantly, at neighboring nodes that are in the region of hotspots. Hotspots are often transient in nature because the mobility of nodes in the network continuously creates, removes, and to some degree, migrates hotspots because node mobility changes the network topology and causes flows to be rerouted. Hotspots are characterized by excessive contention, congestion, and resource exhaustion in these networks. In other words, hotspots appear when excessive contention exists, prompting congestion when insufficient resources are available to handle the increased traffic load.

Hotspots are intrinsic to many on-demand MANET routing protocols because most on-demand routing protocols [4][5][8] utilize shortest path (or hop count) as their primary route creation metric. Most on-demand routing protocols allow an

intermediate node to reply to a route query from cached route information, causing traffic loads to concentrate at certain nodes. We observe from our analysis of hotspots presented in this paper that although many on-demand routing protocols prove to be effective in routing packets in these networks they also have a propensity to create hotspots. Other researchers have also made such observations [1][3][7]. We also observe that hotspot nodes consume a disproportionate amount of resources (e.g., energy).

In this paper, we present a simple, effective, and scalable *Hotspot Mitigation Protocol (HMP)*, which seamlessly operates with existing ad hoc routing protocols, such as AODV [8] and DSR [4]. HMP balances resource consumption among neighboring nodes and improves end-to-end throughput, delay, and packet loss. Our results indicate that HMP can also improve network connectivity preventing premature network partitions. Ideally, establishing routes through non-congested areas of the network and rerouting active flows away from congested areas to non-congested areas would be the best approach to hotspot mitigation. However, this would require extensive collaborations between nodes to establish load-aware routes and sophisticated algorithms to update time-varying loading conditions. Such an approach is unscalable and not practical in mobile ad hoc networks.

HMP represents a fully distributed, localized, and scalable protocol where nodes independently monitor local conditions, and take local actions:

- *to declare* a node to be a hotspot if a combination of MAC contention/delays, packet loss, buffer occupancy, and energy reserves exceed certain predefined system thresholds;
- *to suppress* new route requests at hotspots to ensure that routed traffic does not compound the hotspot's congestion problems; and
- *to throttle* traffic locally at hotspots to force TCP flows to slow down.

HMP also seeks to decrease the energy consumption of nodes via use of these mechanisms.

This paper is structured as follows. In Section 2, we first analyze the behavior of hotspots using existing on-demand MANET routing protocols. Observations from this evaluation show that hotspots are evident even under relatively lightly loaded conditions, motivating the need for HMP. Related work is discussed in Section 3, followed by the design of the protocol in Section 4. We present a detailed analysis of HMP in Section 5 using AODV and DSR. In Section 6, we present some concluding remarks.

2. Hotspots

2.1 Hotspot Observations

Figure 1 illustrates some typical hotspot conditions found in mobile ad hoc networks. Hotspots are generally created where traffic loads converge to a node or small cluster of nodes. Flows traversing multiple wireless hops from various locations intersect with each other and create transient hotspot conditions. We observe that hotspot nodes and nodes in the vicinity of the hotspots (i.e., in hotspot regions) are prone to consume more resources than others. Left unchecked such unbalanced resource consumption is

detrimental to mobile ad hoc networks because overtaxed nodes would prematurely exhaust their energy reserves before other nodes. As a consequence the network connectivity can be unnecessarily impacted. In addition, we observe that hotspot nodes are often responsible for generating a large amount of routing overhead. In general, as the traffic load increases more hotspots appear and conditions in hotspot regions become aggravated.

In what follows, we make a number of observations about hotspots using ns-2 [9] and AODV [1]. Note that the observations we make in this section are common to other on-demand protocols such as DSR [4]. Our simulation consists of 100 mobile nodes in a 1200m by 1200m network with moderate mobility conditions (i.e., pause time of 80 seconds using the random waypoint mobility model with maximum speed of 10 m/sec). Thirty CBR/UDP and 10 TCP flows are used to produce an offered load of approximately 480 Kbps. We detect hotspots through a combination of MAC-delay measurements of unicast packets, packet loss, buffer occupancy, and by optionally considering the remaining energy resources at a node. While the thresholds for these hotspot metrics are configurable, we consider a node to be a hotspot in our current implementation (which is based on IEEE 802.11), when the node consecutively measures *i*) MAC delays that exceed a predefined value, *ii*) packet loss during the RTS-CTS-DATA-ACK cycle, and *iii*) buffer overflow; we discuss these metrics and their configuration in Section 5.1 on hotspot detection.

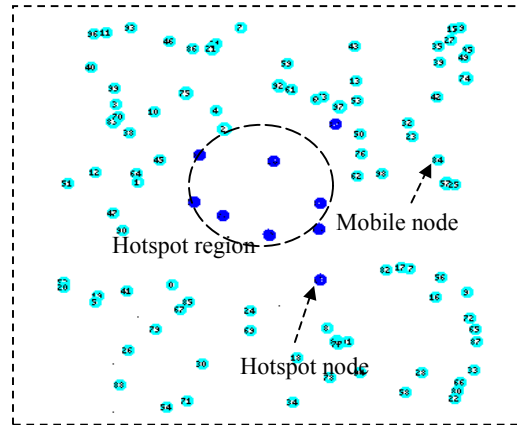


Figure 1. Illustrative Snapshot of Hotspots

Note that hotspots are often transient because of the mobility of nodes changes the topology and continuously varies the traffic load in the network causing hotspots to migrate. We observe in our simulations that nodes are rarely in a permanent hotspot state. As a rule of thumb once a node is declared a hotspot, it is marked as a hotspot for the next 5 seconds. Thus, under simulation, nodes could be declared a hotspot a number of times (e.g., 20 times) during the lifetime of the simulation run. Using this time-scale, we observe an average of 816 congestion hotspot incidents during the simulation runs (i.e., 300 seconds) described above where the offered load is 480 Kbps. Note, that 816 hotspots instances corresponds to 4080 seconds of hotspot

conditions, or, an average of 40.8 seconds of hotspot conditions per node. Results are from 5 simulation runs.

2.2 Traffic Load

Figure 2 shows the packet delivery ratio (PDR), number of hotspots, and offered load for the simulation. The packet delivery ratio is defined as the total number of packets received out of the total number of packets sent. The offered load is varied from 50 Kbps to 963 Kbps for moderate mobility involving 4831 link changes and 39830 route changes. The y-axis represents the packet delivery ratio and x-axis the offered load. In Figure 3, we also show the corresponding number of hotspot instances.

As expected, the number of hotspots increases with offered load, while the packet delivery ratio decreases with increasing load. When the offered load is light, only few hotspots are detected where the network encounters few problems in routing packets. For example, when the traffic load is 72.2 Kbps, approximately 98% of packets are delivered correctly, and only 22 hotspot instances are detected during the simulation. This means that mobile nodes in the network encounter 110 seconds of congested conditions that in turn represents an average of 1.1 seconds/node of congestion. Note that link/route errors can occasionally be interpreted as congestion conditions because packet loss due to congestion is indiscernible from packet loss due to route failure.

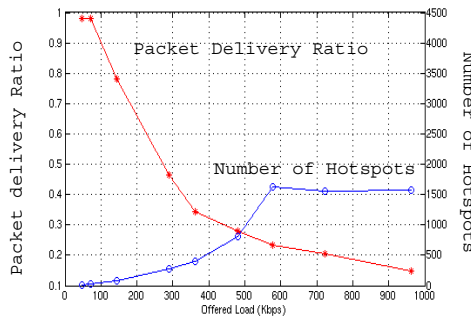


Figure 2. Packet Delivery Ratio and Number of Hotspots

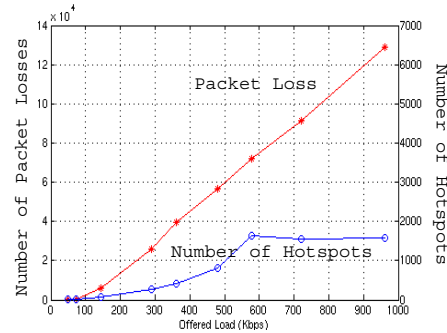


Figure 3. Packet Loss and the Number of Hotspots

In contrast, when the offered load is increases to 963 Kbps then only 15 % of the data packets are correctly delivered with 1566 hotspots instances observed. The difference is more than 70-fold when compared to an offered load of 72.2 Kbps. One interesting observation shown in Figure 3 is that number of hotspots levels-off when the offered load exceeds 580 Kbps. We identified that the reason for this anomaly is mainly due to the failure of congestion detection. All types of packets continuously fail to complete the collision avoidance cycle of IEEE 802.11, and as a consequence, they are considered to be route errors while our hotspot detection mechanism, which relies on the measurement of the RTS-CTS-DATA-ACK cycle, fails to capture the congestion implications. The corresponding packet loss count observed during the simulation clearly supports this.

2.3 Overhead

Figure 4 illustrates the total number of packets transported when offered load is 290 Kbps. The x-axis represents the node IDs and y-axis the number of packets handled by each node. Figure 4 also shows the number of data packets handled or forwarded by each node. One interesting observation is that most of the packets handled in the system are routing-related packets and only a small portion of the total transit traffic are data packets. For example, mobile node 2 handles 20103 packets in total during the simulation but only 1076 are data packets while 19027 are routing packets. Such observations are consistently observed in the network with the result that the ratio of signaling to data packets grows with the offered load.

The increase in the offered load aggravates congested conditions and as a consequence more packet loss is observed. Consecutive packet loss is often treated as route failures by ADOV triggering route recovery procedures that entail additional route requests, route errors, and route reply packet exchanges. It is observed that the routing overhead and number of hotspots increases with the offered load but begins to decrease beyond a certain load (e.g., 700 Kbps in this simulation set) due to substantial packet loss, as discussed earlier (i.e., route request packets continuously fails to be forwarded and rarely reach destination nodes, route replies are rarely generated, with the result that routes are seldom successfully established).

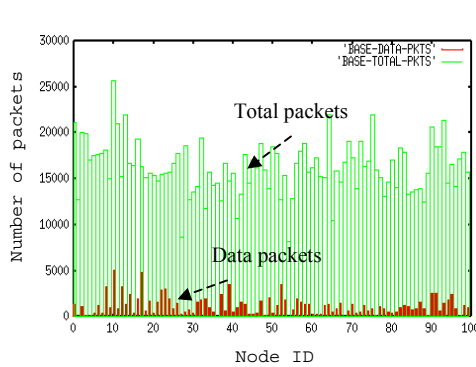


Figure 4. Packets Handled by Nodes

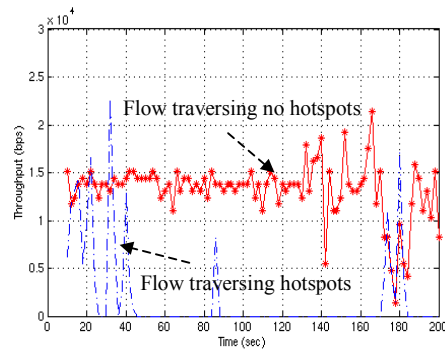


Figure 5. Throughput Traces of Two Monitored Flows

2.4 Hotspot Regions

Figure 5 shows the throughput traces of two similar flows under the simulation configuration discussed previously. We selected a flow traversing multiple hotspots and a flow encountering no hotspots (from our simulation results) and compare their throughput performance. The trace intuitively demonstrates how hotspots impact flow performance. Among 100 mobile nodes, 11 nodes are identified as ‘severe’ hotspots where they experience congestion for more than 110 seconds out of 200 seconds (monitoring period). We identified 59 nodes as immediate neighbors of the 11 severe hotspots. We observed the packet loss of these 70 nodes (i.e., 11 severe hotspots and their 59 neighbors) that resided in hotspot regions, and compared their performance to

other nodes in the rest of the network. We observed that nodes residing close to hotspot nodes also experience degradation in performance. For example, when the offered load is 290 Kbps, hotspot regions are responsible for 94.9 % of total packet loss while rest of network contributed only 5.1 % to the total packet loss. Moreover, nodes in hotspot regions have an average congestion time of 94 seconds while the rest of the network nodes only experience 36 seconds of congestion time. Based on these observations we argue that there is a need to study, design, and evaluate mechanisms that can seamlessly interwork with existing routing protocols to mitigate the impact of hotspots in MANETs.

3. Related Work

MANET [2] routing protocols can be simply classified into best effort routing protocols that have no built in mechanisms to provide better than best effort service [1] [3], QoS-based routing protocols [10][11][12][13], and multipath routing protocols [14][15][16]. While HMP is not a routing protocol it is designed to interwork with the existing best effort routing protocols (e.g., on-demand and proactive protocols) to provide hotspot mitigation support.

Currently, none of the existing on-demand best effort routing protocols [4][8][18][19] take hotspots into account in their routing decisions. As shown in the last section this allows hotspots to quickly emerge and build up in the network under normal operating conditions. There is a clear need to propose new mechanisms that can interwork with, or be directly incorporated into, these best effort routing protocols, hence enhancing the network's performance. HMP is designed as a separate mechanism and is therefore capable of being used in combination with any of the existing best effort routing schemes.

HMP incorporates measures of congestion and contention as well as resource shortages (e.g., energy) into its definition of hotspots. We believe that this is a more realistic definition for wireless mobile networks than one that only considers the buffer occupancy statistics at intermediate nodes. Using buffer occupancy as an indication of congestion has been widely used by a number of Internet congestion control/ hotspot management schemes. HMP manages these hotspots locally (i.e., at the point of interest) in a fully distributed fashion as opposed to traditional end-to-end approach for managing congestion.

The simple goal of HMP is to disperse new flows away from being routed through hotspots and congestion-prone areas, avoiding the further build up of traffic load at hotspots or in hotspot regions. HMP distinguishes itself from the various QoS routing approaches, which in practice are complex to implement, in that HMP does not attempt to provide QoS support nor QoS routes. However, the deployment of QoS routing and multipath routing algorithms would also minimize the likelihood of hotspots, but not eradicate them. QoS routing algorithms require accurate link state (e.g., available bandwidth, packet loss rate, estimated delay, etc.) but due to the time-varying capacity of wireless links, limited resources and mobility, maintaining accurate routing information is very difficult if not impossible in mobile ad hoc networks. Finding a feasible route with just two independent path constraints is an NP-complete problem [17]. Moreover, finding a QoS satisfying path is merely the first part of the problem because it is more challenging to maintain QoS routes when the

network topology changes [11]. Because QoS routing relies on this distributed but global review of resources in the network the likelihood of stale state and traffic fluctuations beyond the anticipated load also calls for localized reactive mechanisms such as HMP to help alleviate transient hotspots. We therefore consider that HMP would also be useful in QoS routed networks.

Alternate path routing and multipath routing protocols can outperform single path routing protocols. A common feature of these protocols is that they utilize backup or alternate routes when primary routes fail. Some multipath routing protocols are designed to distribute traffic among multiple paths and reassemble the traffic at the destination nodes. However, reassembling traffic at destination node in this manner can be problematic because it leads to out-of-sequence delivery and extra re-sequencing delays [12]. Moreover, maintaining additional path information requires additional routing and computational overhead. Alternate paths should be comprised of disjoint-paths [15] in order to be effective. Such alternate paths often do not exist, particularly in single channel mobile ad hoc networks (e.g., based on IEEE 802.11).

In summary, HMP is designed as a localized node mechanism that takes local actions to prevent the build up of hotspots, which we believe will be very likely in MANETs. While HMP is targeted to interwork with the existing best effort routing protocols it could also provide efficient support for hotspot mitigation in MANET networks based on QoS routing and multipath routing. This is the subject of future work.

4. Hotspot Mitigation Protocol

4.1 Protocol Operations

The simple goal of HMP is to redirect new “routes” away from hotspots. HMP disperses new flows away from being routed through hotspots and congestion-prone areas, avoiding the further build up of traffic load in hotspot regions. HMP effectively mitigates hotspot conditions and reduces congestion-related problems. Mitigating hotspot in this manner also helps to balance resource consumption among neighboring nodes, and can extend the lifetime of certain overtaxed nodes.

HMP utilizes MAC-delay measurements, buffer occupancy information, neighbor status information and other resource monitoring mechanisms (i.e., buffer, energy) to detect hotspots. HMP does not limit the scope of monitoring and detection mechanisms, however. Operators are free to introduce additional mechanisms and algorithms according to their needs. In fact, we envision that a HMP network would embody diverse mechanisms operating concurrently. HMP utilizes monitored and measured information to respond to conditions by executing the most appropriate algorithms to alleviate the condition at hand. The measured conditions are explicitly expressed by a multimetric parameter called STATUS, which consists of two components: *symptom* and *severity*. Symptom describes the dominant condition a node is experiencing while severity expresses the degree of the symptom. For example, a node may declare its status as $Y_{\text{CONGESTION}}$ while another node may declare its status as R_{ENERGY} . This status is analogous to traffic lights, where green (denoted by G) indicates a good condition, yellow (Y) represents a marginal condition, and red (R)

represents a critical condition. Therefore, $Y_{\text{CONGESTION}}$ indicates marginal congestion and R_{ENERGY} indicates critically low energy reserves. Users/operators are free to introduce more granularity if needed. HMP piggybacks this status information in the IP option field and neighboring nodes operating in promiscuous mode learn the status of transmitters by eavesdropping their packets. The eavesdropped information is used to create and update a *Neighborhood Status Table (NST)*. This cached information is locally maintained and updated at each node.

An NST caches a list of immediate neighbors and their status. It is primarily used to manipulate new-route-creation decisions at nodes. In other words, a node refers to its NST to ensure that it is not aggravating the conditions of neighboring nodes by creating additional routes through them. We assume a finite number of neighboring nodes surrounding any node, which in effect defines the size of the NST at a node.

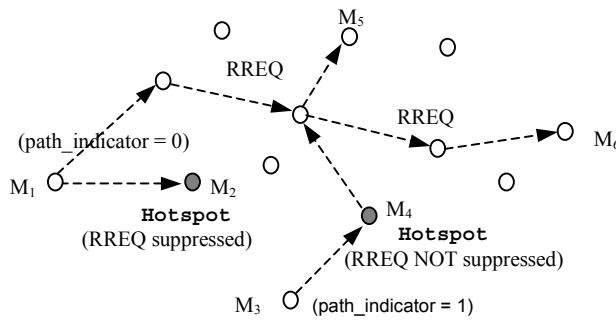


Figure 6. Hotspot Mitigation Protocol Illustration

The naïve suppression of new route creation may prevent the use of the only possible path between two hosts and may yield poor connectivity in the network, or even cause network partitions. To avoid this, a new-route-suppression mechanism is used, if and only if, there exists a sufficient number of non-hotspot neighbors within its transmission range. HMP also makes sure that preceding nodes en-route also have enough non-hotspot neighbors. The notion of ‘enough neighbors’ is defined by the `enough_nh_neighbor` parameter (i.e., currently set at 6 in our implementation). The value of this parameter has a direct impact on the network connectivity, as discussed in Section 5.5. If `enough_nh_neighbor` is too small, (e.g., 2), then HMP manifests low connectivity among mobile nodes and often fails to provide useful routes. HMP also ensures that it is not inadvertently denying the only possible path between two end hosts by utilizing an indicator called the `path_indicator`, which is carried in the IP option field of Route Request (RREQ) messages. A node that has only a few neighbors sets this indicator (`path_indicator = 1`) and upstream nodes that receive the RREQ (with IP option that includes `path_indicator`) check this indicator and avoid suppressing new routes if it is set. This is illustrated in Figure 6 where hotspot M_4 forwards RREQ toward M_5 because the source node M_3 has set its path indicator whereas hotspot M_2 suppresses RREQ from M_1 because its `path_indicator` is not set in the IP option field of the RREQ.

4.2 Congestion Levels

The main objective of congestion avoidance algorithms is preventing the further build up of traffic at hotspots. HMP distinguishes two levels of congestion (i.e., levels 1 and 2) and adopts two corresponding algorithms to support this view. The first algorithm is activated when HMP determines the current status of a node is in a moderately congested condition (i.e., level 1), denoted by $Y_{\text{CONGESTION-1}}$. This algorithm simply suppresses the creation of additional routes at hotspots by discarding new route request packets. As mentioned previously, HMP ensures not to deny the ‘only route’ between two hosts.

The second algorithm is more aggressive and executes when nodes encounter substantial congestion (i.e., level 2), denoted by $Y_{\text{CONGESTION-2}}$. This algorithm is executed when a node experiences severe hotspot conditions without any non-hotspot neighbors. This algorithm not only suppresses new route creation but also throttles best effort TCP flows traversing the node in an attempt to reduce the load using rate control mechanisms discussed in [6]. TCP flows are bandwidth hungry and unless controlled can easily occupy all remaining wireless medium bandwidth. Throttling TCP rates locally in this manner does not necessarily hurt TCP sessions but can effectively relieve congestion bottlenecks. Users and operators are free to introduce other schemes to relieve congestion conditions. One simple policy is dropping TCP packets at bottleneck nodes.

HMP attacks the congestion at the point of congestion (POC) as opposed to a traditional end-to-end approach. Although congestion is an end-to-end issue where it is detected and controlled (e.g., in the case of TCP), traditional remedies for end-to-end congestion control are not effective in mobile ad hoc networks. In fact, such traditional control mechanisms may limit the utilization of the wireless medium that is constrained by hotspots. We argue that we can avoid such shortcomings if we tackle the problem at the point of congestion rather than responding on end-to-end basis.

4.3 Energy Conservation

Mobile ad hoc networks are essentially energy-limited networks and are likely to be comprised of heterogeneous nodes with diverse energy constraints. Some mobile devices will have large energy reserves in comparison to others. There exist various energy-aware power-conserving protocols for mobile ad hoc networks [20]. The common objective these protocols lie in conserving energy as much as possible to prolong the lifetime of the network or extend the lifetime of individual nodes.

Although energy conservation is not a primary function of HMP, the protocol provides a simple mechanism to conserve energy through its status declaration mechanism. A node with limited energy reserves can declare itself a hotspot by setting its status to Y_{ENERGY} or R_{ENERGY} when its energy reserves are marginally or critically low, respectively. The triggering thresholds are $P_{\text{YELLOW-THRESH}}$ and $P_{\text{RED-THRESH}}$. In our current implementation, $P_{\text{YELLOW-THRESH}}$ is set to 50% of node’s initial (or maximum) energy reserves and $P_{\text{RED-THRESH}}$ is fixed at 1.00 joule. The latter value represents the amount of energy needed for a node to sustain a CBR flow for approximately 300 packets in most of our simulation sets. However, we note that operators and users are free to set these values according to their own needs, based on the characteristics of the targeted network. A node with energy concerns is acknowledged by neighboring

nodes and new route creation through the node is avoided if possible. On the other hand, a node with critical energy (i.e., R_{ENERGY} status) immediately relinquishes its role as a router and functions strictly as an end host in order to conserve energy (maximize its lifetime) unless it is identified as the only intermediate node between two communicating end hosts.

5. Performance Evaluation

In what follows, we evaluate HMP through simulation and discuss the performance improvements that the protocol offers. Simulation metrics such as packet delivery ratio, packet loss, throughput, end-to-end delay, per-hop delays, and energy consumption are used in our evaluation. We also discuss the impact of various parameters on the performance of HMP.

In the initial part of the evaluation we use the AODV [1] [8] routing protocol with HMP, and in the latter part, DSR [4] with HMP. We implemented HMP using the ns-2 simulator and its wireless extension. The HMP implementation includes monitoring modules, measurement mechanisms, an NST module, and the HMP algorithms discussed in Section 4. The simulated network size is 1200 meters by 1200 meters where 100 mobile nodes create 10 TCP and 30 CBR/UDP flows that arbitrarily last for 60 to 280 seconds. Moderate mobility is assumed with a pause time of 80 second using the random way point mobility model [1] [4] unless specified otherwise. All data packets are of fixed size of 128 bytes, each simulation run lasts for 300 seconds, and each data point represents an average of 5 simulation runs with the identical traffic model but different mobility scenarios. Each mobile node has a transmission range of 250 meters and shares a 2 Mbps radio channel with its neighboring nodes. The simulations also include a two-ray ground reflection model, finite energy module, and IEEE 802.11 MAC protocol. Throughout the evaluation section we use the terms ‘HMP system’ and ‘baseline system’ to refer to wireless ad-hoc networks with and without the HMP mechanisms, respectively.

5.1 Hotspot Detection

Accurate and timely hotspot detection is one of most crucial aspect of HMP. To determine hotspots, the protocol relies on MAC-delay measurements, packet loss detection in RTS-CTS-DATA-ACK exchanges, buffer occupancy, and residual node energy. Among the measurements we have observed that the MAC-delay measurement is the most useful since a hotspot always manifest increased delays in the RTS-CTS-DATA-ACK cycle. Surprisingly, relying solely on the buffer occupancy is rather inaccurate. We often witnessed that hotspot conditions are created without any buffer occupancy. Such events are due to excessive contention among neighboring nodes. Therefore, in order to minimize the margin of error in hotspot detection, we utilize both buffer information and MAC-delay measurements together with some other additional system parameters discussed later.

Figure 7 shows a typical trace of the MAC-delay measurement of a node. The x-axis represents the simulation time and y-axis represents the MAC-delay measurements of a randomly selected mobile node. As shown in the figure, MAC-delay measurements continuously fluctuate throughout the simulation. Spikes in the

delay trace typically represent congested conditions, while zero delay measurements are observed when the node is not participating in the RTS-CTS-DATA-ACK activity. Note that detection of a hotspot is dependent on two key parameters: (i) MAC-delay threshold (i.e., denoted by `cong_thresh`), which determines when a packet is considered a delayed packet; and (ii) `num_thresh`, which determines when a node is considered a hotspot. Specifically, a node is considered a hotspot when the measured MAC-delay measurements exceed a predetermined threshold (i.e., `cong_thresh`) for more than `num_thresh` times consecutively. These two parameters have an impact on how many hotspots are detected by HMP. When `cong_thresh` and `num_thresh` are configured as large values, HMP is too conservative and only detects a small number of hotspots rendering the protocol to be less effective against moderate congestion. In contrast, when `cong_thresh` and `num_thresh` are configured with small values, HMP is aggressive and detects too many hotspots too hastily. Therefore, the appropriate choice of these parameters is important for HMP to function properly. The use of the `num_thresh` parameter also prevents HMP from premature detection of a hotspot when experiencing a momentary increase in the MAC-delay measurement. It was observed that the MAC-delay measurements intermittently “spike” without any noticeable congestion conditions (e.g., during rerouting). To avoid reacting to such transient behavior and to increase the accuracy of hotspot detection, HMP marks a node as a hotspot, if and only if, the MAC-delay measurements are violated (i.e., exceeds `cong_thresh`) more than `num_thresh` times ‘consecutively’. Currently, `cong_thresh` is set to 20 msec and `num_thresh` is set to 4. In Section 5.5, we evaluate a number of different configurations of the protocol based on these parameters and study the sensitivity of the parameter settings to HMP’s ability to efficiently and accurately detect and mitigate hotspots in MANETs.

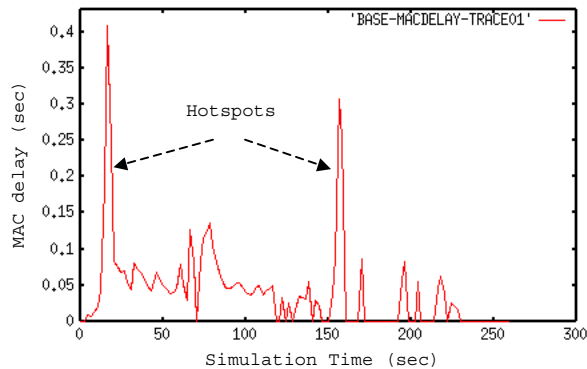


Figure 7. Trace of MAC-delay Measurements

5.2 Throughput Analysis

We first observe how the HMP system performs in comparison to the baseline system in terms of the packet delivery ratio (PDR). Figure 8 shows a comparison of the packet delivery ratio against increasing load for two different HMP system configurations (discussed below) and the baseline system. The two HMP systems are simply called HMP-P and HMP-R where HMP-R is more aggressive than HMP-P in its route

suppression mechanism. HMP-P stands for HMP-POC where HMP mechanisms are executed only at points of congestion (POC). On the other hand, HMP-R represents HMP-Regional signifying the regional execution of hotspot mitigation algorithms. In other words, when a hotspot is detected HMP-P executes hotspot mitigation algorithms at the point of hotspots whereas HMP-R executes its mechanisms across a hotspot region. A node belongs to a hotspot region if it is a hotspot or it is an immediate neighbor of a hotspot. We note that both `enough_nh_neighbor` and `path_indicator` are always considered in all hotspot mitigation decisions.

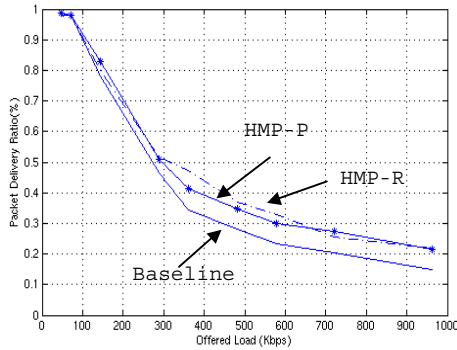


Figure 8 Comparison of PDR against network load

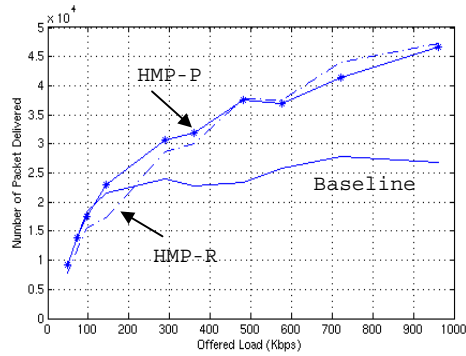


Figure 9. Number of Data Packets Delivered

As observed in Figure 8, HMP-P and HMP-R have little impact on lightly loaded networks, (e.g., below 100 Kbps). This is because the baseline system already achieves more than 90 % PDR and HMP has little room to make any improvements. However, as the offered load increases, and congestion builds up, HMP begins to provide improvements, as shown in the figure. Both HMP-P and HMP-R provide substantial improvements in the PDR. Specifically, HMP-P and HMP-R provide up to a 43% and 46% increase in the packet delivery ratio when compared to the baseline system performance. From Figure 8, we also observe the behavior of HMP-R is more aggressive than that of HMP-P. When the offered load is moderately high, HMP-R often outperforms HMP-P and the baseline systems but becomes less effective when the offered load is light, (e.g., below 250 Kbps). The performance of HMP-R varies with different loads, as shown in Figure 8. We conclude that HMP-R is too aggressive for lightly loaded networks rendering it only useful in heavily loaded networks.

Further analysis of the HMP-P, HMP-R and baseline systems can be seen by inspecting the number of delivered packets, as shown in Figure 9. One interesting observation is that number of packets delivered by the baseline system levels-off around 2.3×10^4 delivered packets but in the HMP-P and HMP-R systems the number of delivered packets continuously increases with increasing offered load. There are two major reasons for this improvement. First, HMP creates routes through non-congested nodes whenever possible allowing networks to utilize more distributed routes in the network even if these routes are not the shortest path. Creating routes at non-hotspot nodes helps traversing flows to encounter fewer problems, and as consequence, more packets are delivered. Secondly, HMP generates less routing overhead through suppression executed at hotspots. Many hotspot nodes rebroadcast

route request packets and these route request packets often flood large areas of the network or even the entire network. However, many of these rebroadcast route request packets are lost before reaching destination nodes. We observed that a considerable amount of route request packets are just wasted in the network without successful route creation in heavily loaded networks.

In the HMP systems, routing packets (i.e., route request) are pre-filtered at hotspot nodes/regions. This not only prevents new routes being created through hotspots but also helps reduce the number of ‘wasted’ new route requested packets (that rely on broadcast/flooding), which are likely to be lost. This opens up room for more data packets and as consequence more packets are delivered in HMP systems in comparison to the baseline system. Moreover, as congestion become more severe more nodes encounter packet loss and often interpret this packet loss as route errors, triggering route recovery routines. As a consequence, additional routing overhead is added to an already congested network. In HMP systems, congested nodes avoid participating in new route creation to mitigate congested conditions, and consequently less routing packets are observed in the network.

We observe that HMP-R outperforms HMP-P when the offered load is heavy. However, HMP-R is too aggressive for lightly loaded networks. We observe that the PDR of HMP-R is less than that of HMP-P and no better than that of the baseline system when the offered load is less than 150 Kbps. However, both HMP systems outperform the baseline system. Hereafter, we refer to HMP-P when we discuss HMP unless specified otherwise.

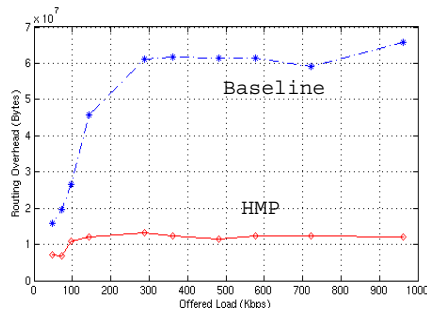


Figure 10. Comparison of Routing Overhead

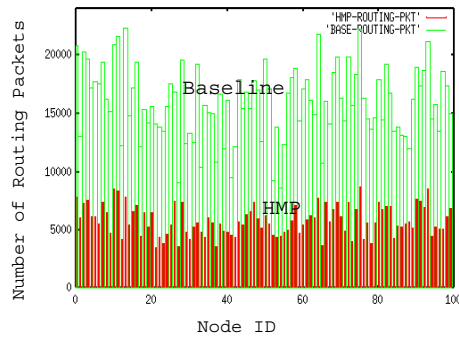


Figure 11. Routing Overhead Compared

5.3 Routing Overhead Analysis

Figure 10 shows the routing overhead of the HMP and baseline systems accumulated over 300 seconds of simulation. The advantage of the HMP system over the baseline system in terms of the routing overhead is shown in the figure. It is observed that the HMP system provides up to a 75.7 % reduction in the routing overhead over the baseline system because of better route selection and routing packet suppression in hotspot regions. For example, when the offered load is 722 Kbps, the baseline system generates approximately 59×10^6 bytes of routing overhead while the HMP system only generates 13.3×10^6 bytes of routing overhead.

Figure 11 compares the total number of routing packets transported by the HMP and baseline systems when the offered load is 577 Kbps. As expected there is a substantial difference between the two systems. It is observed that the HMP system always carries less routing load in comparison to the baseline system. This implies that HMP is not over-suppressing routes because if connectivity were limited, the number of route request packets would quickly increase and be reflected in the routing overhead. Therefore, it is safe to say that HMP provides sufficient connectivity in all the simulated scenarios. The HMP system outperforms the baseline system in terms of the PDR, number of packet delivered, and routing overhead. These improvements are mainly due to effective hotspot mitigation through implicit route dispersal and suppression of new route request packets. HMP is prudent in route suppression decisions while ensuring sufficient connectivity when the configuration of the system (i.e., `cong_thresh`, `num_thresh`, `enough_nh_neighbor` and `path_indicator`) is enforced.

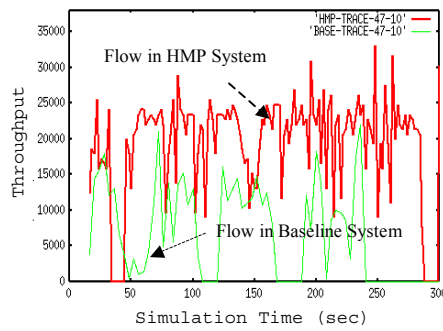


Figure 12. Throughput Trace Comparisons

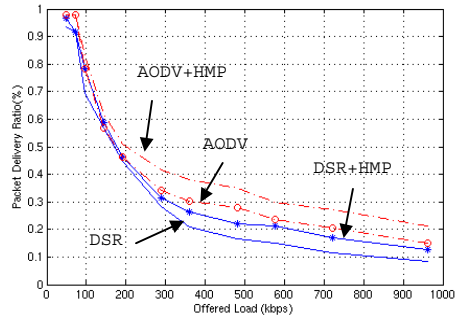


Figure 13. Impact of HMP on DSR and AODV

Next, we compare throughput traces of a flow in the two systems. Figure 12 shows a monitored flow between node 47 and node 10. The monitored flow in the HMP system shows substantial improvements over the baseline system. More importantly, it is observed that the monitored flow traverses different routes in the two systems. Specifically, flow 47-10 traverses nodes 16, 18, 43, 51, 78 and 83 in the baseline system and traverses 16, 21, 38, 65, 78 and 83 in HMP system, during the monitored period of 50 seconds. Nodes 18 and 43 are identified as hotspots and consequently flow 47-10 avoided these two hotspots when using HMP. Such characteristics are consistently observed throughout the simulation, and as a consequence, the HMP system provides better throughput performance.

The previous evaluation of HMP considered AODV routing only. In what follows, we describe how HMP performs with DSR [4]. First, we observe the PDR traces of the baseline DSR system in comparison to the HMP+DSR system. Figure 13 shows the PDR trace for increasing offered load with moderate mobility for these systems. The figure also includes the PDR trace for the baseline AODV system and the AODV+HMP system (taken from Section 5.2) for comparison purposes.

As expected, the DSR+HMP system provides improvements over the baseline DSR system. From Figure 13, it is observed that all the systems demonstrate similar performances under lightly loaded conditions but they begin to diverge as the offered load increases. One interesting observation is that DSR and AODV display different

performances against increasing offered load. They show similar results only in lightly loaded conditions. As congestion intensifies AODV begins to outperform DSR. This observation coincides with the results reported in [1]. Moreover, the HMP is seen to be more effective with AODV than DSR. The main reason for this is related to the amount of routing load reduction.

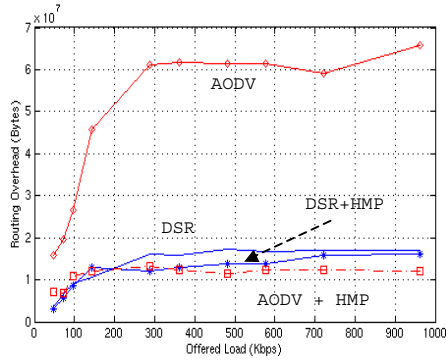


Figure 14. Routing Overhead Compared

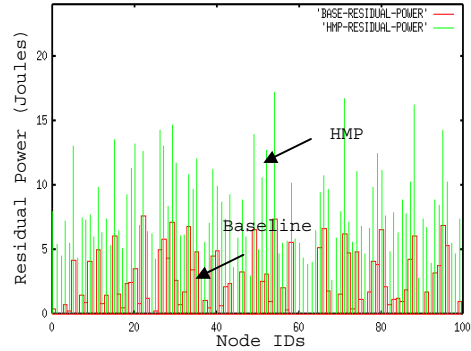


Figure 15. Residual Energy Compared

It is observed from Figure 14 that HMP provides substantial reductions in routing overheads when operating with AODV but demonstrates different results with DSR. The difference in routing overhead reduction is directly reflected in the PDR traces (or when we compare the number of packets delivered). HMP provides improvements with AODV mainly through the reduction in the routing overhead, and route diversions away from hotspots. In contrast, in the DSR system the dominant reason for the improvement is mainly due to route diversion from hotspots. It is also observed that DSR's aggressive use of route-cache limits its performance. Under harsh conditions (i.e., increased mobility, increased load), it is observed that DSR maintains stale routes, generating a large amount of route-error messages. This observation is also reported in [1]. HMP successfully routes traffic through non-hotspot nodes but DSR's route-optimization scheme [4] utilizes cached routes, which often introduce new hotspots. Figure 14 reflects this observation. The bottom line is that HMP can provide improvements for both AODV and DSR.

5.4 Energy Analysis

We note that energy consumption is concentrated in hotspot regions and nodes. Figure 15 shows measurements of residual energy for nodes at the end of the simulation run. We assign a uniform energy of 25 joules to each node and conducted simulation for 100 seconds with AODV. The x-axis represents node IDs and y-axis represents the residual energy in joules. Bars represent the residual energy measurements of baseline system and the superimposed impulses represent the corresponding measurements of the HMP system. As shown in the plot, the energy conservation provided by HMP for nodes in hotspot regions is significant. The baseline system exhibits 21 energy-depleted nodes (i.e., remaining energy is less than 0.01 joule such that it can no longer

participate in communications) while there is not even one depleted node in the HMP system at $t = 100$ sec. Note that HMP improves packet delivery ratio, delay measurements, and reduces routing overheads, while providing energy conservation in hotspot regions.

5.5 Parameterization Sensitivity

In what follows, we describe four different HMP system configurations to study the responsiveness of the protocol to detect and mitigate hotspots. Four key parameters govern the HMP system control mechanisms; these are, `cong_thresh`, `num_thresh`, `enough_nh_neighbor` and `path_indicator`. For example, if the `cong_thresh` value is too small HMP may become too aggressive and declare too many hotspots. A small increase in the MAC-delay threshold measurement (or jitter) may falsely be recognized as congestion with many nodes being claimed as hotspots. In contrast, if the `cong_thresh` value is too large HMP may not identify any hotspots in the network and relegate itself to the baseline system. The second parameter `num_thresh` is used to prevent HMP from reacting to transient behavior. A momentary increase in the MAC-delay measurement and buffer occupancy are not necessarily a product of congestion or excessive contention. Delay may be observed for a very short period due to the rerouting of flows or a small burst of route query packets. Reacting to such transitory phenomenon is not beneficial because real hotspots cannot be distinguished from transient events. The third parameter is the `path_indicator`, which indicate that insufficient conditions exist for new route suppression. Nodes receiving packets with this indicator set know that at least one preceding node explicitly requested ‘no new-route-suppression’. This is a valuable HMP feature because it provides a safeguard against potential over-suppression of new route creation that may result in limited connectivity. The fourth parameter is the `enough_nh_neighbor` that prevents the HMP algorithm from being too aggressive.

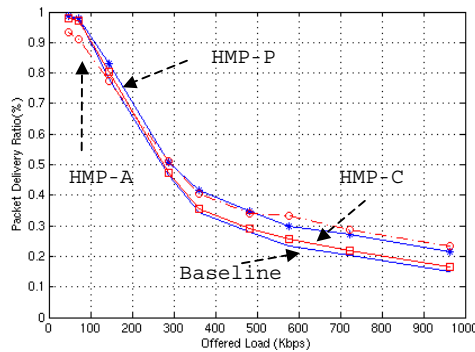


Figure 16. PDR Trace of HMP-A, HMP-C, HMP-P and Baseline System Compared

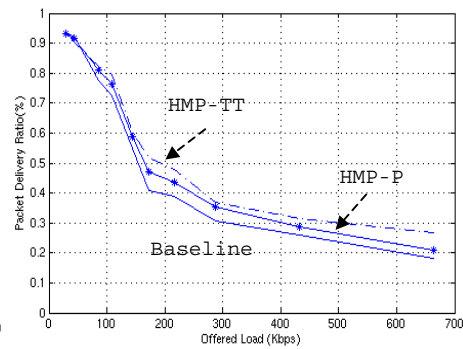


Figure 17. PDR of HMP-P, HMP-TT and Baseline system Compared

Figure 16 shows the PDR traces of the four different HMP systems under discussion. HMP-P and HMP-R are described in Section 5.2 while HMP-C and HMP-A represent HMP-Conservative and HMP-Aggressive HMP system configurations,

respectively. HMP-A is literally an aggressive version of HMP-P that quickly determines hotspots (i.e., `num_thresh = 3`, `cong_thresh = 10 msec`, `enough_nh_neighbor = 4`) and without utilization of the `path_indicator`. HMP-A is equally effective as HMP-P when the network is heavily loaded but results in limited connectivity when lightly loaded. As a consequence of its aggressiveness, HMP-A supports only 91 % PDR even when the offered load is only 72 Kbps. At the slightest indication of congestion, HMP-A suppresses new route creations and limited connectivity among mobile nodes.

HMP-C is a conservative version of HMP-P that utilizes the `path_indicator` and configures `num_thresh = 8`, `cong_thresh = 100 msec`, `enough_nh_neighbor = 8`. As shown in Figure 16, HMP-C closely resembles the baseline system with slight improvements. In other words, if the HMP protocol is too aggressive it is deemed to limit connectivity. On the other hand, if HMP is too conservative, it rarely detects hotspots and degrades to the baseline system performance.

As mentioned in Section 2.2 HMP relies on TCP throttling when severe congestion persists. As observed in Figure 17, HMP-TT (for HMP-TCP-Throttle) can selectively throttle TCP flows to relieve hotspots. TCP throttling is meaningful in the presence of congestion since TCP flows are typically transported as a best effort service where traffic rates are often transparent to higher layers (i.e., applications). In contrast, CBR/UDP traffic often requires better than best effort service. However, details on these issues are not immediately related to HMP and considered outside the scope of our initial research.

6. Conclusion

In this paper, we have presented a protocol that works with existing best effort routing protocols to mitigate hotspots in mobile ad hoc networks. We have demonstrated through simulation that hotspots exist in mobile ad hoc networks and can limit the performance of these networks. HMP tackles the congestion problem at the point of congestion as opposed to traditional end-to-end approaches. We argue that traditional remedies such as end-to-end congestion control are often not effective in mobile ad hoc networks and can limit the utilization of the wireless network in the face of hotspots. We are currently working on a testbed implementation of HMP and studying a more comprehensive set of simulation scenarios with larger numbers of nodes, a wider variety of offered traffic, and other routing schemes, (e.g., proactive schemes).

Acknowledgements

This work is supported in part by the Army Research Office (ARO) under Award DAAD19-99-1-0287 and with support from COMET Group industrial sponsors. The authors would like to thank Jiyoung Cho for her comments on this paper.

References

1. Samir R. Das, Charles E. Perkins, and Elizabeth M. Royer. "Performance Comparison of Two On-demand Routing Protocols for Ad Hoc Networks." Proceedings of the IEEE Conference on Computer Communications (INFOCOM), Tel Aviv, Israel, March 2000
2. MANET Working Group, <http://www.ietf.org/html.charters/manet-charter.html>
3. S.B Lee, G.S. Ahn, and A.T. Campbell, "Improving UDP and TCP Performance in Mobile Ad Hoc Networks with INSIGNIA", June 2001, IEEE Communication Magazine.
4. David B. Johnson, David A. Maltz, and Josh Broch. "DSR: The Dynamic Source Routing Protocol for Multi-Hop Wireless Ad Hoc Networks", in Ad Hoc Networking, edited by Charles E. Perkins, Chapter 5, pp. 139-172, Addison-Wesley, 2001
5. V. Park and S. Corson "A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks, Proc. IEEE Infocom 1997, Kobe, Japan
6. G.S. Ahn, A.T. Campbell, A. Veres and L-H Sun, "SWAN: Service Differentiation in Stateless Wireless Ad Hoc Networks", Proc. IEEE Infocom 2002, New York, New York, June 2002
7. S.J. Lee and M. Gerla "Dynamic Load-Aware Routing in Ad hoc Networks" Proceedings of IEEE ICC 2001, Helsinki, Finland, June 2001
8. C. Perkins and E. Royer. "Ad hoc On-Demand Distance Vector Routing." Proc. of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, February 1999, pp. 90-100
9. The NS-2 Simulator, <http://www.isi.edu/nsnam/>
10. Chunhung Richard Lin, "On-demand QoS Routing in Multihop Mobile Networks", Proc. IEEE Infocom 2001, Anchorage, April 22-26, 2001
11. S. Chen and K. Nahrstedt, "Distributed Quality of Service Routing in Ad Hoc Networks", IEEE Journal on Selected Areas in Communications", vol. 17, No. 8, Aug 1999.
12. C. R. Lin and C-C Liu, "On-Demand QoS Routing for Mobile Ad Hoc Networks", IEEE International Conference on Networks (ICON'00), September 5-8, 2000, Singapore
13. W.H. Liao, Y.C. Tseng, S.L. Wang, and J.P. Sheu, "A Multi-path QoS Routing Protocol in a Wireless Mobile Ad Hoc Network," Telecommunication Systems Vol. 19, No. 3-4, pp. 329-347, 2002
14. S. Guo and O.W. Yang, "Performance of Backup Source Routing in mobile ad hoc networks", in Proc. 2002 IEEE Wireless Networking Conference
15. A. Nasipuri and S. Das, "On-demand multipath routing for mobile ad hoc networks," in Proc. IEEE ICCCN '99, Oct. 1999
16. M.R. Pearlman, Z.J. Haas, P. Scholander, and S.S. Tabrizi, "Alternate Path Routing in Mobile Ad Hoc Networks," IEEE MILCOM'2000, Los Angeles, CA, October 22-25, 2000
17. M. Garey and D. Johnson, Computer and Intractability: A Guide to Theory of NP-Completeness: W.H. Freeman, 1979
18. D. Johnson, D. Maltz, Y-C Hu and J. Jetcheva, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks", Internet Draft, draft-ietf-manet-dsr-07.txt, work in progress, Feb 2002.
19. C. Perkins, E. Royer and S. Das, "Ad Hoc On-demand Distance Vector Routing", Internet Draft, draft-ietf-manet-aodv-12.txt, work in progress Nov 2002.
20. Christine E. Price, Krishna M. Sivalingam, Prathima Agarwal and Jyh-Cheng Chen, "A Survey of Energy Efficient Network Protocols for Wireless and Mobile Networks", in ACM/Baltzer Journal on Wireless Networks, vol. 7, No. 4, pp. 343 - 358, 2001