Wireless Networks 12, 5–21, 2006

Energy Efficiency and Capacity for TCP Traffic in Multi-Hop Wireless Networks

SORAV BANSAL, RAJEEV SHOREY and RAJEEV GUPTA
IBM India Research Laboratory, Block 1, Indian Institute of Technology, Hauz Khas, New Delhi 110016, India

ARCHAN MISRA
IBM T J Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA

Abstract. We study the performance metrics associated with TCP-regulated traffic in multi-hop, wireless networks that use a common physical channel (e.g., IEEE 802.11). In contrast to earlier analyses, we focus simultaneously on two key operating metrics—the energy efficiency and the transport-layer (TCP) throughput. Using analysis and simulations, we show how these metrics are strongly influenced by the radio transmission range of individual nodes. Due to tradeoffs between the individual packet transmission energy and the likelihood of retransmissions, the total energy consumption is a convex function of the number of hops (and hence, of the transmission range). On the other hand, the throughput of a single TCP session decreases with a decrease in the transmission range. The overall achievable TCP throughput in an ad-hoc network thus involves a tradeoff between the reduced throughput of an individual flow and the greater degree of spatial reuse possible. As a consequence of this tradeoff, the overall network capacity turns out to be a concave function of the transmission range. We analyze how parameters such as the node density and the radio transmission range affect the overall network capacity under different operating conditions. Our analysis shows that capacity metrics at the TCP layer behave quite differently from the capacity results previously presented in literature. We then extend the work and examine the sensitivity of the TCP-layer capacity to the speed of the nodes and the number of TCP connections in an ad hoc network. By incorporating the notion of a minimal acceptable QoS metric (loss) for an individual session, we show why the *QoS-compliant capacity* is a more accurate metric for studying the capacity of TCP traffic in an ad hoc network. Finally, we study the dependence of capacity on the source application (Telnet or FTP traffic) and on the choice of the ad hoc routing protocol (AODV, DSR or DSDV).

1. Introduction

Analyses of the transmission capacity of multi-hop, ad hoc wireless networks typically relate bounds on the *maximal achievable* throughput to spatial reuse constraints and MAC-layer effects. In networks where all nodes uses the same physical channel (such as IEEE 802.11 [7] based ad hoc LANs), a packet transmission by a node effectively precludes any simultaneous transmissions by neighboring nodes (within its interference range). In such networks, the total network capacity, defined as *the cumulative number of bits received by all destination nodes from all traffic flows*, is clearly dependent on the transmission range of each node. For a similar reason, this maximum achievable throughput is also a function of the node density, which implicitly determines the average number of one-hop neighbors (who are subject to constraints on concurrent transmissions). In recent work, Gupta and Kumar [6] showed how an increase in N, the total number of nodes, causes the throughput of an individual node to degrade as $O(\frac{1}{\sqrt{N \log N}})$ when the nodes are randomly distributed. Li et al. [11] studied the behavior of the IEEE 802.11 MAC layer and showed how the end-to-end throughput for an individual node degrades as $O(\frac{1}{\sqrt{N}})$ for random traffic patterns, and remains constant if the sessions exhibit appropriate *localization* properties.

In this paper, we examine how the characteristics of the transport-layer protocol (TCP) affect the achievable transmission capacity in such ad-hoc networks. The analysis presented in this paper differs from prior work in the following key aspects:

(i) In contrast to the use of greedy traffic sources used in earlier studies on the the maximum achievable capacity, we consider *TCP regulated* flows. The two metrics can be very different, since TCP flow control may prohibit certain packet transmissions, even if they satisfy the underlying MAC-layer constraints. It is well-known that the achievable throughput of a TCP connection is a function of both its round-trip time (RTT) and the packet loss rate—we shall show how both those parameters are affected by the underlying radio transmission range.

(ii) Besides the total number of bits received by all destination end-points per unit time (achievable throughput), we also concentrate on another metric of interest: the *energy efficiency*, defined as the average total transmission energy required to *reliably* transmit a single packet (or byte) to its destination. Our metric includes the energy spent in potential retransmissions needed to overcome possible errors in the traffic path.

(iii) We examine how variations in the mobility rates impact the throughput[1] achieved at the TCP layer. Since the cumulative TCP throughput is also a function of the traffic load (number of TCP sessions), we take care to ensure that the offered load is feasible (in the sense that the resulting network performance does not violate any associated Quality of Service (QoS) constraints).

(iv) To ascertain the sensitivity of our TCP-centric analyses and results, we also study the behavior for two different traffic sources, representing two extremes of TCP-based applications. We consider both persistent or greedy (e.g., FTP) traffic, as well as non-persistent or intermittent (e.g., Telnet, HTTP) traffic.

(v) We also study the sensitivity of our results to the choice of a specific ad-hoc routing protocol (AODV, DSR or DSDV) and demonstrate that our analytical conclusions are essentially independent of protocol-specific features.

For the analysis in this paper, we assume that all nodes are identical in the sense that they all use the same transmission range R; we study the properties of TCP traffic as R is varied. Our focus is on treating R as a design parameter, and evaluating how changes in R affect the overall network performance in different operating conditions. We shall also study how changes to N, the number of ad-hoc nodes, affect the network performance for different values of R. We also assume that the maximum capacity of the physical channel is *independent of the transmission range* and is denoted by C; for our studies with IEEE 802.11 LANs, we have used $C = 2$ Mbps.

We first demonstrate how the *energy-efficiency* metric is a function of the transmission range. In a variety of multi-hop wireless networking scenarios, the energy efficiency is indeed the most critical metric, since it directly affects the network lifetime. Energy-aware ad hoc routing algorithms typically choose a path that results in the minimum total transmission energy for a single packet; Banerjee and Misra [1] shows why a more accurate objective should be the *minimum total effective transmission energy*, which focuses on reliable packet reception and includes the energy spent in one or more retransmissions.

We then study how the radio transmission range affects the *achievable throughput* of a TCP session in such wireless networks. It is well known that the throughput of a TCP session (whose capacity is determined by the error rate and not by network buffering constraints) varies as $O(\frac{1}{RTT*\sqrt{p}})$ [4,12] if the path error rate p is small, and as $O(\frac{1}{RTT*p})$ [13] if p is moderately high. We study how the range parameter, R, indirectly affects both p and RTT and hence, bounds the TCP session throughput. Additionally, we also consider the TCP throughput achieved over a chain of nodes using the 802.11 MAC layer, and observe how this throughput varies from the ideal maximum presented in [11].

Both studies mentioned above are compared with practical results obtained via simulations performed using IEEE 802.11. We subsequently use the analytical results to derive the *total network transmission capacity* with TCP traffic for such ad hoc networks and its relation to the number of nodes N and the transmission range R. Since the capacity definition for TCP traffic is not immediately apparent, we define the network's TCP-centric capacity as the *total (cumulative) goodput achieved by all TCP sessions*. We consider *two* different scenarios. In the first scenario, we assume that the number of TCP sessions, as well as the number of nodes are fixed. We then vary the total area A of the wireless network (implicitly varying the node density) and then observe how the cumulative goodput varies with changes in the transmission range of individual nodes. In the *second* scenario, we assume that the network is dispersed over a fixed area A and that the number of TCP sessions is proportional to the total number of network nodes. Our analysis shows, that in contrast to earlier studies based on maximal link-layer throughput, the throughput of the individual TCP is $O(\frac{1}{N^{\frac{3}{2}}})$ and the total network goodput is $O(N^{\frac{1}{4}})$ for moderate link error rates. We also use simulation studies with 802.11-based multi-hop wireless networks to quantitatively explore the validity of our analysis. After establishing the accuracy of our analysis for *static* networks, we extend the framework to consider the impact of node mobility on the total transmission capacity. Simulation-based studies are also used to investigate the behavior of network capacity subject to reasonable QoS constraints and to demonstrate that our results hold for a variety of routing protocols and application environments.

The paper is organized as follows. In Section 2 we discuss the related work. Transmission energy efficiency and transmission range of nodes are discussed in detail in Section 3. We consider the impact of the transmission range on the throughput achieved by an idealized TCP session in Section 4. Having studied both the energy-efficiency and the individual TCP session behavior with varying R, we focus on the total capacity of the ad hoc network in Section 5. In Section 6, we extend the work to study the transmission capacity for different levels of node node mobility and with varying number of TCP connections. We argue why the incorporation of QoS constraints is essential for determining a useful TCP-centric notion of capacity in an ad hoc network. We also study the dependence of this 'QoS-compliant' capacity on the source application (Telnet or FTP), and, on the choice of the ad hoc routing protocol (AODV, DSR and DSDV). We conclude in Section 7 with a brief description of our future work in this area.

2. Related work

It is widely recognized that low network capacity (or the aggregate throughput achieved by different sessions) is a major constraint in the effective deployment of multi-hop wireless

[1] When referring to TCP traffic, we shall use the terms 'throughput' and 'goodput' interchangeably—both refer to the number of unique packets correctly delivered to the eventual destination node, and do not consider retransmitted packets.

networks. In networks where nodes use the same physical channel, the transmission range of individual nodes is a key determinant of this capacity, since it effectively determines the extent of spatial reuse possible. When seesion end-points are chosen at random and the transmission range is fixed, Gupta and Kumar [6] demonstrated that the capacity of each individual session would degrade as $O(\frac{1}{\sqrt{N \log N}})$ with an increase in N (the number of nodes) and presented the design of an idealized MAC which would achieve this bound. Gupta and Kumar [6] also showed that, even if nodes were placed optimally, the maximum average per-session throughput would degrade as $O(\frac{1}{\sqrt{N}})$ as long as the session end-points were chosen at random. Shepard [17] considered the design of an optimal MAC layer to maximize the total utilization of the shared channel over all the nodes in a multi-hop network. Li et al. [11] considered how the IEEE 802.11 MAC algorithm performed relative to the bounds enumerate in [6], and also showed that if the traffic patterns showed appropriate stochastic locality (more accurately, if the probability of the session distance decayed faster than D^{-2}), then the ideal throughput per session would remain a constant. These studies, however, consider idealized sources that are completely greedy and are constrained purely by the MAC layer. In particular, they do not consider the use of TCP traffic and the impact of transmission errors in the link layer on the maximal link utilization by such TCP sources.

Studies on energy-efficient communication for wireless networks typically focus on the routing problem alone: they are concerned solely with maximizing some measure of the total transmission energy or minimizing some function of the battery drainage. For example, [18] adapts Dijkstra's minimum cost path selection algorithm to find minimum total energy paths, by setting the link cost to the associated transmission energy. Such energy-efficient routing protocols assume that, when the physical distance of a hop is smaller, the wireless nodes are able to appropriately reduce their transmission power. Similarly, newer routing algorithms (e.g, [5]) replace a long-distance hop with a series of short-distance ones, thereby minimizing the total power usage. Battery-aware routing protocols [19,22] often consider the residual energy level of the node's battery as a metric, and hence attempt to form routes using potentially less-drained nodes. *Since modification of the transmission range implies modification of the session throughput, such power-conscious routing algorithms implicitly affect the networkcapacity.* Current studies do not however analyze how the selection of such energy-efficient paths impact other metrics such as session throughput.

The performance of TCP congestion avoidance under varying loss rates and RTT has been extensively analyzed in literature (e.g. [4,12,13]), especially for point-to-point links. For moderate to low loss rates, the TCP throughput varies inversely as the square-root of the loss probability. The interaction of TCP performance with the contention-based MAC scheduling in multi-access media is less clearly understood.

3. Energy efficiency and transmission range

We consider a scenario where the transmitter radios are capable of dynamically altering their transmission power; accordingly, the transmission energy of a node is a non-decreasing function of the transmission distance. We first focus solely on the packet transmission cost, and then show how the energy budget may change substantially if we additionally consider the computing cost.

The power attenuation with distance D in wireless environments is is usually proportional to $D^K : K \geq 2$. Under the assumption of omni-directional antennas, it follows that the transmission power needed to communicate over a radial distance R is proportional to R^K. Moreover, a transmission range of R implies a coverage area (within which concurrent transmissions are not allowed) $\propto R^2$. Accordingly, an energy-efficient transmission scheme will ensure that the transmission energy over a single hop (or link), $E(R)$, of distance R is:

$$E(R) \propto R^K \tag{1}$$

Given the above relationship between the transmission energy and the total transmission distance, it is easy to see that the total energy associated with a *single* transmission event actually decreases if a hop is sub-divided into multiple smaller ones: clearly, if $D_1 + D_2 = D$, then $D_1^K + D_2^K < D^K$ if K > 2. Energy-efficient routing protocols thus usually seek to transmit a packet between a source S and a destination D using multiple short-distance hops, as opposed to a smaller number of long-distance hops. Indeed, minimum total-energy routing algorithms, such as [18], result in the formation of routes with a large number of short-range hops. This intuition is, however, misleading: the formulation neglects the fact that an increase in the hop-count leads to an increase in the packet error rate over the entire path, and thereby increases the likelihood of retransmissions and thus decreasing the session throughput. Accordingly, Banerjee and Misra [1] proposes the use of the *effective transmission energy* (which includes the energy spent in retransmissions) as the appropriate metric.

Analysis in [1] shows that, in the absence of reliable link layers (or what is called the end-to-end retransmission or EER model), the actual *effective* energy per reliably transmitted packet over a $H - hop$ path (with nodes indexed as $(1, \ldots, H + 1)$) is given by:

$$E_{\text{total}}^{EER} \propto \frac{\sum_{i=1}^{H} D_{i,i+1}^K}{\prod_{i=1}^{H}(1 - p_{i,i+1})}, \tag{2}$$

where $p_{i,i+1}$ indicates the packet error rate of the ith hop (between nodes i and $i+1$). On the other hand, if the number of permitted retransmissions on each link is unbounded (hence, each link ensures accurate delivery to the next hop), the total effective energy per packet (in the so called hop-by-hop or

HHR model) is given by:

$$E_{\text{total}}^{HHR} \propto \sum_{i=1}^{H} \frac{D_{i,i+1}^{K}}{1 - p_{i,i+1}}. \quad (3)$$

Analysis of the expression for the EER mode shows that, even if all the links have identical error rates, there is an optimal value for the number of hops associated with a specific transmission path. If the number of hops is smaller, the energy budget is dominated by the larger transmission energies needed to transmit over larger distances; if the number of hops is larger, it is the overhead associated with retransmissions that negates the energy gains associated with smaller individual hops. In contrast, if each link is allowed potentially unlimited number of retransmission attempts, the total effective transmission energy always decreases with increasing H.

3.1. Transmission energy efficiency and transmission range

Before proceeding further, it is necessary to extend the analysis of *effective transmission energy* mentioned above. To apply our insights quantitatively to technologies, such as IEEE 802.11, we need to analyze the case where each link has an *upper bound on the maximum number of retransmission attempts*. This bound is a practical necessity to avoid abnormally large latencies and buffer overflows at the link layer. We assume that each link layer is permitted a total of *max* transmissions; clearly, such a restriction resurrects the possibility of end-to-end retransmissions in the case of forwarding failure at an intermediate link. Also, for analytical ease, we assume that all links have the same packet error rate p and the same transmission energy E. We relegate the complete mathematical analysis to the Appendix, mentioning only the relevant results here.

Result 1. If each link has a transmission packet error rate p, then the conditional expected number of distinct transmissions, *given the successful forwarding over the link*, is given by:

$$T_{\text{good}} = \frac{1}{1-p} - \frac{\max * p^{\max}}{1 - p^{\max}},$$

and the expected number of distinct transmissions, *given the failure of the link forwarding process* is given by:

$$T_{\text{bad}} = \max.$$

Result 2. In case of an end-to-end failure in reliable packet delivery (one of the H intermediate links failed to reliably forward the packet), the total number of expected distinct transmissions is given by:

$$T_{\text{bad}}^{\text{total}} = T_{\text{bad}} + T_{\text{good}} * (1-q) * \left\{ \frac{1 - H * (1-q)^{H-1} + (H-1) * (1-q)^{H}}{q * \{1 - (1-q)^{H}\}} \right\}, \quad (4)$$

where $q = p^{\max}$. Similarly, if the packet was indeed successfully forwarded to the destination node, the total number of expected distinct transmissions is:

$$T_{\text{good}}^{\text{total}} = H * T_{\text{good}} \quad (5)$$

By combining the above two results with the fact that the probability of successful packet end-to-end delivery is given by $(1-q)^H$ (where $q = p^{\max}$), we can finally derive the following result:

Result 3. The total effective number of distinct packet transmissions needed for reliable packet delivery is given by:

$$T = T_{\text{bad}}^{\text{total}} * \frac{P_{fail}}{1 - P_{fail}} + T_{\text{good}}^{\text{total}}, \quad (6)$$

where $P_{fail} = 1 - (1-q)^H$.

Since T is really a function of H, p and max, we represent this result generically as $T(H, p, \max)$. We defer the quantitative comparisons of our analytical expression with simulation results to the next sub-section and, instead, focus on the expected qualitative behavior. Clearly, in the limited-retransmission case, there is an optimal value for H, the number of hops: if H becomes too large, then the probability of an end-to-end error becomes non-negligible and the consequent effects of end-to-end retransmissions begin to dominate the energy budget. In fact, the approximate value of this optimal value can be obtained by realizing that, from the standpoint of energy consumption alone, a link with a packet error rate of p and a transmission bound of *max* is essentially equivalent to a link with no retransmissions but a link packet error rate of $p^{\max}$. (This is not completely accurate when we consider the effects on protocols at higher layers; for example, link-layer retransmissions are likely to result in greater variation in the forwarding latency and hence, the possibility of spurious TCP-layer timeouts.) Accordingly, using the analysis in Banerjee and Misra [1], the optimal value of H is, to a good approximation, given by $\frac{-1}{\log(1-p^{\max})}$.

For a generalized ad hoc network, it is now easy to see the connection between the transmission radius and effective energy. If we assume that the average distance between the end-points of a session is $\bar{L}$, then a transmission range of R implies that the average number of hops, H is given by $\lceil \frac{\bar{L}}{R} \rceil$, or to a good approximation by $\frac{\bar{L}}{R}$. Accordingly, with a link layer bound of *max* on the number of retransmissions, equation (6) shows that the effective energy efficiency of the ad hoc network is given by (ignoring proportionality constants):

$$E_{\text{total}} = R^K * T\left(\frac{\bar{L}}{R}, p, \max\right) \quad (7)$$

Clearly, as long as the decrease in R in the expression (7) dominates over the corresponding increase in $T(\cdot)$, the energy consumption per byte decreases. Beyond the optimal value for R, the decrease in the energy spent in any single transmission activity is negated by the larger increase in $T(\cdot)$. From an energy efficiency perspective, there is an optimal value to the

radius of acceptable reception quality R in an ad hoc network; decreasing the transmission range below this optimal value does not lead to greater energy savings.

3.1.1. Applicability to the 802.11 environment

We applied this analytical model to the 802.11-specific environment, using the 802.11 implementation in the ns-2 [21] simulator. For our simulations, the distance between the source and destination was kept at 750 meters, while the transmission range was varied between (30, 700) meters; H was thus varied from 2 to 24. The energy associated with each transmission was assumed to be (ignoring proportionality constants) given by $E \propto R^2$; the simulations were run for both uncorrelated (i.e., i.i.d) and correlated error models. For the results plotted here, we set the transmission power for a distance of 250 meters to 0.03346 W, and then computed the corresponding power for other transmission distances by appropriate scaling.

The effective transmission energy per packet was computed by determining the total transmission energy spent in transferring a 10 MB-sized file using a TCP flow from the source to the destination. Since the number of packets transferred reliably by TCP is the same for all simulations, the total communication energy consumption is a direct indicator of the transmission energy efficiency. The number of hops H was varied by simply inserting the corresponding number of intermediate nodes between the source and destination. The total energy consumption is clearly a function of the maximum number of retransmissions supported at each link (the T $\times$ Thresh parameter in ns-2). We present results here for T $\times$ Thresh equal to 1 and 4; the corresponding value of max (see equation (7)) was thus 2 and 5 respectively. We simulate the energy efficiency for TCP file transfer using two standard models for the link error: a) the two-state Markov-modulated channel model with correlated errors and b) the independent identically distributed (i.i.d) model with independent and identically distributed bit error rates.

Figure 1 plots the simulated total transmission energy consumption, under the i.i.d model, as H varies between 2 and 23 for two different values of p, 0.1 and 0.2, and T $\times$ Thresh equal to 1. The figure also includes the energy efficiency values (with appropriate scaling) predicted by equation (7). We can see that the theoretical model, while an accurate reflector of the overall trend, underestimates the energy consumption, especially for larger values of H. This is to be expected, since our analytical formulation does not include the energy spent in the 802.11 signaling (such as RTS/CTS/ACK packets), as well as the energy wastage in potential MAC layer collisions (which can be expected to occur more often for higher values of H). It is easy to see that, when the link layer permits only one retransmission, the optimal value of H (from simulations) is larger than 23 for $p = 0.1$; even when the error rate is fairly large ($p = 0.2$), the optimal number of hops is approximately 15. The number of TCP level retransmissions for the two cases have also been plotted in figure 2; as expected, the number of source-initiated retransmissions needed increases with increasing H.

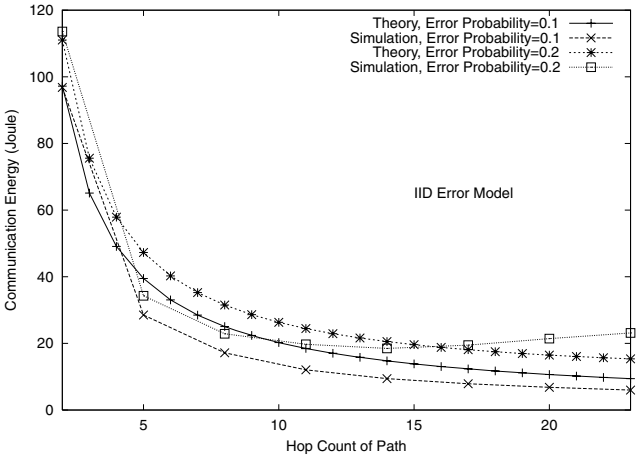

Figure 1. Effective transmission energy vs. number of hops (T $\times$ Thresh = 1).

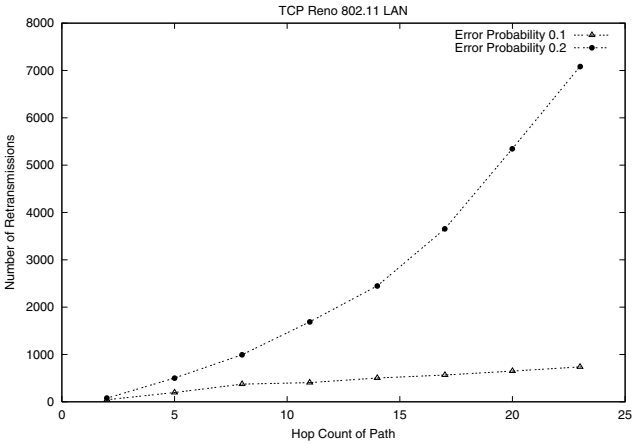

Figure 2. Number of TCP retransmissions vs. number of hops (T $\times$ Thresh = 1).

To further study the impact of link-layer retransmissions, we also studied the total energy consumption with T $\times$ Thresh equal to 4 and three link error rates:

(a) The two-state Markov model where the average sojourn times in the Good and Bad states were 1.0 and 0.3 ms respectively.

(b) The two-state Markov model where the average sojourn times in the Good and Bad state were identical and equal to 1.0 ms.

(c) The i.i.d model with p set to 0.5 (a very high value).

Figure 3 plots these simulation results for the total transmission energy with T $\times$ Thresh $=$ 4; it is again seen that under all these operating conditions, the transmission energy consumption decreases as long as H is increased over any realistic range.

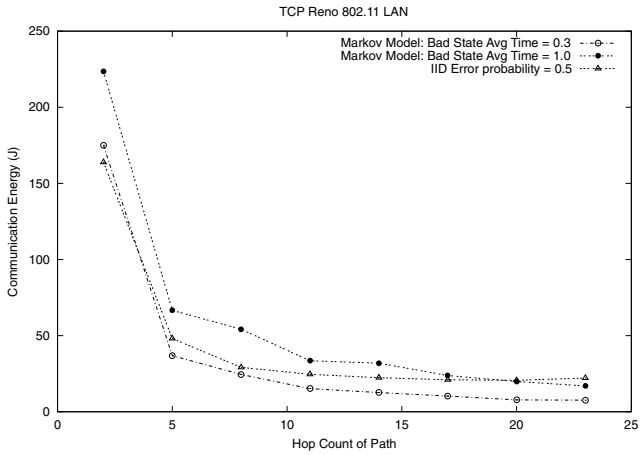

Figure 3. Effective transmission energy vs. number of hops (T × Thresh = 4).

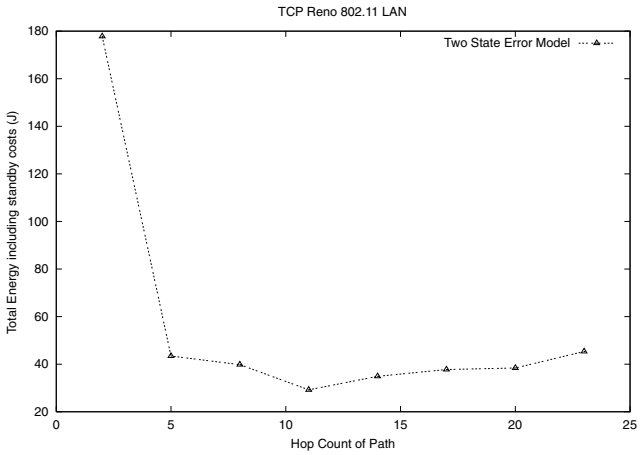

Figure 4. Total energy vs. number of hops (T × Thresh = 4).

3.2. Total energy efficiency

The discussion and results of the previous section show that a larger number of hops, or equivalently a smaller transmission range, typically always increases the energy efficiency. This argument is, however, misleading, since this formulation ignores the *computing energy*– any node engaged in packet transmissions also expends ambient energy in addition to that consumed by the radio interface. In particular, we shall see in the next section that an increase in H typically leads to a corresponding drop in the TCP goodput, even if the physical distance between the source and destination nodes is unchanged. Hence, while the transmission energy efficiency may indeed increase with H, the resultant loss in throughput implies that the transfer of a fixed number of bytes will take a longer time. Since the total *computing energy* can be assumed to be proportional to the total activity duration, it should be clear that this cost will only increase with H.[2]

To formally explore this concept, we repeated the energy-related simulations, taking care to measure the total time taken by TCP to reliably transfer the entire 10 MB file. If we then assume then P_a is the ambient or standby power spent by each node during the lifetime of the session, the computing energy expenditure over all the H nodes is equal to $P_a * H *$ simulation duration. Denoting $E(transmission)$ to be the total transmission energy spent in transferring a 10 MB-sized file using a TCP flow from the source to the destination, the total energy consumption is now given by:

$$E_{\text{total}} = E(transmission) + H * P_a * \text{simulation duration}.$$

Figure 4 plots the variation in this total energy with changing H for the experiments using the two-state error model with Good and Bad sojourn times of 1.0 msec and 0.3 msec respectively (T × Thresh = 4). Similarly, figure 5 plots the total energy consumption versus the number of hops for the two-state error model with Good and Bad sojourn times of 1.0 msec and 0.3 msec respectively (T × Thresh = 1), and the i.i.d error model with $p = 0.1$ (T × Thresh = 1). These results correspond to a choice of $P_a = 0.004$ W.

It is easy to see that, when the total energy is considered, the energy consumption is minimized for realistically small values of H. For example, if we consider only the transmission energy, the optimal value of H was certainly greater than 23 for the i.i.d channel with an error rate of 0.1. However, when the total energy consumption is considered, it is clear that increasing the number of hops beyond ∼10–12 hops will prove to be disadvantageous. (In our simulated environment,

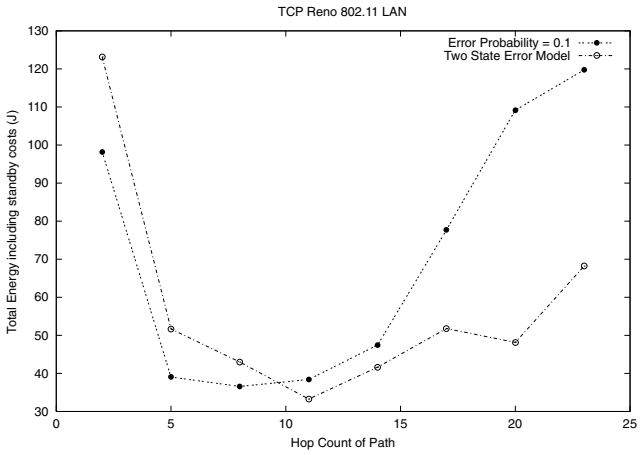

Figure 5. Total energy vs. number of hops (T × Thresh = 1).

[2] To keep the analytical framework simple, we have ignored the energy spent by nodes in packet reception, although earlier studies [20] have documented that packet reception in current wireless cards is almost as expensive as actual packet transmission. For one thing, the packet reception energy is really dependent on the receiver hardware implementation and can be expected to reduce as more efficient receiver circuits are designed in the future. Moreover, including the reception energy does not alter our qualitative conclusions, since it really serves to increase the energy cost associated with an increase in the number of hops H. The inclusion of the computing energy is itself adequate for illustrating the point that a significant piece of the energy budget actually increases with increasing H.

an optimal hop count of 12 corresponds to an optimal transmission range, R_e, of ∼65 meters.) Our studies thus clearly show that any adjustments to the transmission range to improve the network capacity (which we shall define appropriately in the Section 5) must also consider the potential effect on the energy efficiency of the resulting network. If the transmission range is decreased beyond an optimal value R_e such that the average number of hops traversed by a session increases beyond ∼10–15, then any increase in network capacity comes only at the expense of higher energy consumption.

4. Maximum throughput of a single TCP session

After analyzing the energy-related metrics of an ad hoc network, we now consider the impact of the transmission range on the throughput achieved by an idealized TCP session. In this section, we assume the absence of any cross-traffic from other sessions; the path for the session of interest is thus simply a node-chain. Analysis in [11] showed that, for such a chain topology (where the nodes could interfere with their one and two-hop neighbors), the maximum ideal capacity is $\frac{C}{4}$; with 802.11 MAC-based scheduling, the maximum obtained throughput is usually around $\frac{C}{7}$. To achieve such an ideal throughput, the MAC layer must be the only bottleneck; in contrast to these analyses, we consider a persistent flow subject to the dynamics of TCP flow control. The throughput of a persistent TCP flow depends on the range of the magnitude of the error rates and the buffer capacity available at intermediate nodes.

If the TCP losses occur primarily due to link errors, and if buffer overflow is a fairly rare event, then the throughput of a TCP connection as a function of p and RTT is given by the well-known square-root formula:

$$\rho(RTT, p) \sim \frac{\kappa * MSS * 8}{RTT * \sqrt{p}}, \tag{8}$$

where RTT equals the round-trip delay, p equals the effective error rate, MSS indicates the packet size (in bytes) and where κ is an implementation-specific constant. (For example, κ is $\sim \sqrt{2}$ for TCP without delayed acknowledgments and ∼1 with delayed acknowledgments.) The above equation holds as long as p does not become much larger than ∼15–20% for most TCP versions; larger values of p lead to undesirable transients such as retransmission timeouts and a sharper drop in the TCP throughput.

On the other hand, if TCP losses occur primarily due to buffer overflows, the dynamics of the connection becomes much harder to analyze in the presence of multiple hops. In such a situation, the RTT is dominated by the various queuing delays; however, in general, the throughput of the TCP flow decreases with an increase in the RTT.

For practical ad hoc topologies, the propagation delays are usually small–consequently, the *RTT is dominated by the queuing and transmission delays.* Assuming that nodes are homogeneous, the RTT is thus directly proportional to H, the number of hops, since each additional hop introduces queuing and transmission delays. If the error probability of each link is a constant p, the end-to-end error probability is given exactly by $1-(1-p)^H$; if $H * p << 1$, the end-to-end packet error rate is then approximately $H * p$. Accordingly, for ad hoc networks operating under relatively small end-to-end packet error rates (say, less than ∼10%), the maximal throughput of a TCP connection should behave as the following function of H:

$$\rho \propto \frac{1}{H * \sqrt{H}} \propto \frac{1}{H^{\frac{3}{2}}}. \tag{9}$$

However, if the error rates are so low that the TCP flow almost never halves its window in response to a link loss, it should be clear that the throughput becomes independent of the link error probabilities. In such a case, since $RTT \propto H$, the TCP throughput will vary as:

$$\rho \propto \frac{1}{H} \tag{10}$$

For a fixed mean distance $\bar{L}$ (in absolute units) between the end-points of an ad hoc session, the average number of hops, H, as a function of the transmission range R is given by $H = \frac{\bar{L}}{R}$. Accordingly, the maximum throughput of a persistent TCP flow will vary $\propto R^{\frac{3}{2}}$ if the flow is link-loss controlled, and $\propto R$ if the flow is buffer-loss controlled. Of course, the above equations hold good only when ρ is less than the theoretical goodput of the chain topology. For example, in a linear topology with ideal MAC scheduling and interference radius equal to the acceptable reception radius, the dynamics of TCP flow control act as the primary flow capacity constraint as long as $\rho \leq \frac{C}{3}$. If the inequality does not hold, then the session throughput is constrained, not by TCP dynamics, *but by the interference at the MAC layer* among simultaneous transmissions by neighboring nodes.

4.1. Applicability to the 802.11 environment

To study the variation of TCP session throughput with the number of hops in the 802.11 environment, we performed simulations with our chain topology. As before, the distance between the session end-points was kept constant– the number of intermediate hops was varied by varying the transmission range. Moreover, we plotted $\log(\rho)$ against $\log(H)$; in this case, the slope of the resultant curve determines the exponent in the relationship between ρ and H.

Figure 6 plots the TCP throughput (in terms of packets/sec for an MSS of 512 bytes) against H on a logarithmic scale when the link error rate is very small (0.001) and T × Thresh = 1; in this case, the resultant end-to-end loss rate is negligible and TCP is primarily buffer controlled. The slope of the curve is ~ -1, indicating fairly good agreement with our analysis. On the other hand, figure 7 plots the TCP throughput (again in units of packet/sec for 512 byte packets) against H for $p = 0.1$ and T × Thresh = 1. In this case, the resultant error

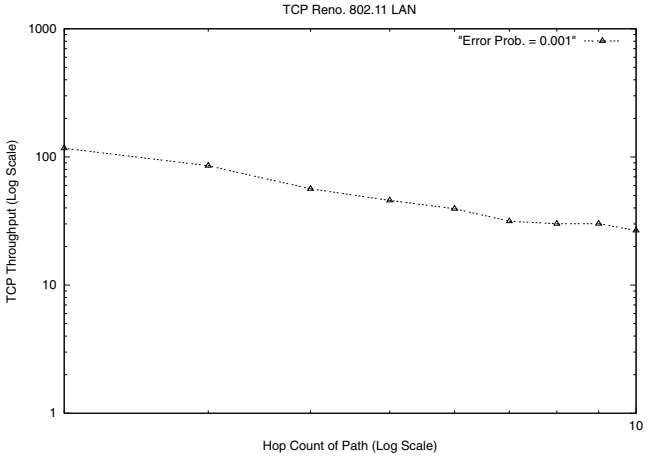

Figure 6. Throughput vs. number of hops (T × Thresh = 1).

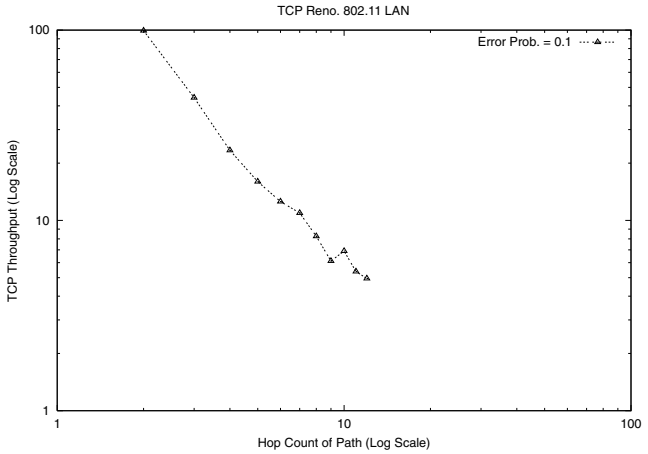

Figure 7. Throughput vs. number of hops (T × Thresh = 1)

rate is moderately high; the slope of the curve is around -1.7 in this case, which indicates fairly close agreement with our theoretical analysis.

The results on the TCP throughput in such multi-hop networks are important from the capacity analysis standpoint. The results show that for TCP-controlled traffic, decreasing the transmission range actually penalizes the maximum session throughput, since the consequent increase in the number of hops increases both the RTT and the end-to-end loss rate. As we shall see in the next section, this phenomenon impacts the amount of TCP traffic that such a multi-hop, wireless network may be expected to carry.

5. TCP-based ad-hoc network capacity

Having studied both the energy-efficiency and the individual TCP session behavior with varying R, we now focus on the total capacity of the ad hoc network. Most literature defines the network capacity Cap as the total "one-hop throughput" or the "bit-distance product"—fundamentally speaking, this is a weighted sum of all the session throughputs, with the weight of each session equal to the distance (or the number of hops) over which it passes. We recall from Section 1 that the network's TCP-centric capacity is defined as the total (cumulative) goodput achieved by all TCP sessions in an ad hoc network. It is to be noted that the number of TCP sessions in an ad hoc network cannot be arbitrary. An unbounded increase in the number of TCP connections can ultimately lead to a drop in the system capacity since it degrades the performance metrics of an individual session. It is with this in mind, we study the QoS-compliant network throughput in Section 6.4.

From a theoretical perspective, if the transmission (and interference) range of the ad hoc nodes are R, then a node transmitting packets at the channel capacity C effectively prohibits any transmission activity for all nodes within the coverage area, which is $\propto R^2$. Accordingly, if the area of the ad hoc network is A, and the transmission and interference radii are both R, the maximal ideal (MAC-constrained) capacity of the ad hoc network is $\frac{C*A}{\pi*R^2}$. In a more generic context, where reception and interference radii are not necessarily identical, the maximal network capacity Cap is $\propto \frac{A}{R^2}$. In general, we would thus expect the maximal ideal throughput to increase quadratically with a reduction in the transmission radius.

Since a greedy TCP flow (where $cwnd$ is the only constraint for packet generation at the transport layer) cannot avail of the maximal capacity, the concept of maximal TCP throughput and network capacity becomes trickier. It is also apparent that attempting to attain $\sim$100% link utilization by pumping up the number of parallel TCP sessions is also not feasible, especially in wireless networks where the buffer capacity on individual nodes is fairly limited. We thus study the expected throughput behavior for two different, but interesting, operational scenarios.

5.1. The fixed session, variable area framework

In the fixed session, variable area framework, the number of simultaneous TCP sessions and the total number of network nodes is assumed to be a constant. We study changes to the total TCP throughput when the area of the ad-hoc network (or equivalently, the node density) is varied. We recall from the previous section on capacity of a single TCP session that for a fixed mean distance $\bar{L}$ between the end-points of an ad hoc session, the average number of hops, H, as a function of the transmission range R is given by $H = \frac{\bar{L}}{R}$. From the standpoint of the MAC layer, the number of permissible concurrent transmissions decreases with increasing range R; on an average,

$$Cap \propto \frac{A}{R^2} \tag{11}$$

Now, for a fixed number of TCP sessions, the total throughput is proportional to the throughput of an individual TCP session (as long as the MAC layer bounds are not

violated), i.e.,

$$Cap \propto \frac{1}{\left(\frac{\bar{L}}{R}\right)^{\frac{3}{2}}} \tag{12}$$

If R is very small, the average degree of connectivity of the graph is fairly small. The resultant sub-optimal paths imply that each packet has to travel a large number of hops (H) to reach to the destination. Accordingly, the TCP session throughput decreases with decreasing R, if R is below a certain value. Therefore the sum of the throughputs (over the fixed number of sessions) becomes smaller. On the other hand, if R is larger than a certain value, then the resultant MAC-layer channel interference and collisions limit the capacity of the TCP sessions. In this range of R, the TCP sessions are prevented from better exploiting the network by the larger delays caused due to collisions and backoffs at the MAC layer. We can thus expect an optimal value of R, denoted by R^*. *To values of R larger than R^*, the network is MAC-layer constrained, with the channel interference dominating the throughput; to the left of this value (smaller R), the network is TCP-layer constrained (equation (12)), with the TCP sessions unable to pump enough packets into the network.*

Accordingly, it follows that for R smaller than this optimal value, the network capacity will degrade in proportion to the TCP throughput degradation ($\propto R^{\frac{3}{2}}$ from equation (9)), if p lies within a sensible operating range. To the right of this optimal value, the resultant throughput is determined by the competing effects of higher TCP-layer throughput (lower loss rates due to smaller H) and greater MAC contention. Thus, from equations (12) and (11), we would expect the 'capacity' in this range to vary as the product of two conflicting components:

$$Cap \propto \frac{A}{R^2}\frac{1}{\left(\frac{\bar{L}}{R}\right)^{\frac{3}{2}}} \propto R^{\frac{-1}{2}}. \tag{13}$$

Figures 8–10 show results for capacity as transmission range R is varied. In these simulations, 50 nodes were randomly distributed in a square grid area. 25 TCP connections were chosen randomly and every node was either a TCP source or a TCP destination, but not both. All our simulations with random topologies use DSR for computing the session paths; in the absence of mobility, the choice of paths (and consequent network performance) is expected to be independent of the choice of a specific ad hoc routing protocol.

In figure 8 we plot the capacity versus R for an error-free channel model and a square grid of 500 m $\times$ 500 m. We see that the optimal value of R (from a capacity standpoint) is $\sim$35–40 meters.

In figure 9, we have plotted TCP goodput versus the transmission range for various link Packet Error Rates (PER) (for a constant 500 m $\times$ 500 m grid) under the IID error model. We see that as PER increases, the TCP goodput decreases and the the optimal transmission range (i.e., the range corresponding

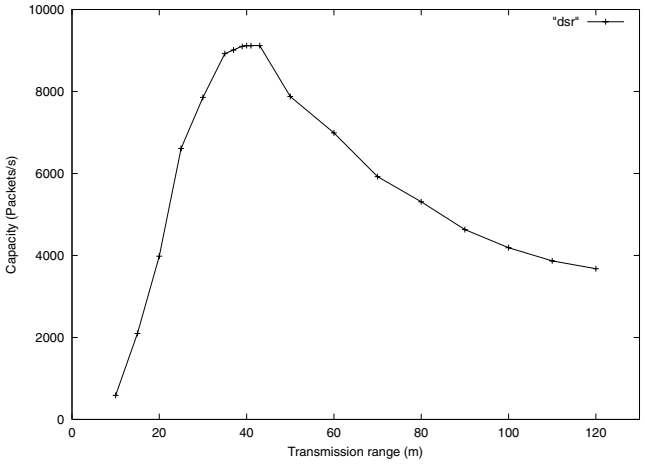

Figure 8. Total network throughput (TCP traffic) vs. transmission range (A = 500 m $\times$ 500 m).

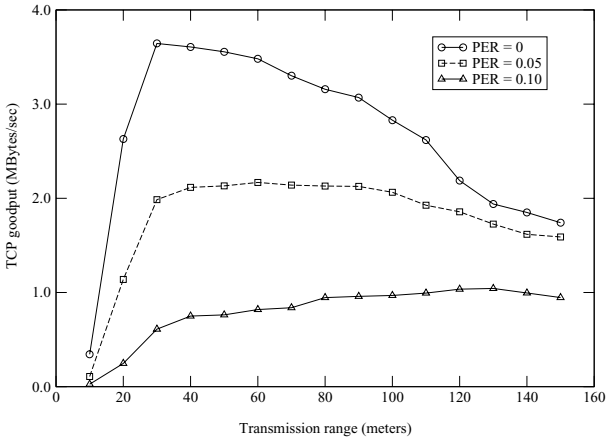

Figure 9. Total network throughput (TCP traffic) vs. transmission range for varying PER (A = 500 m $\times$ 500 m).

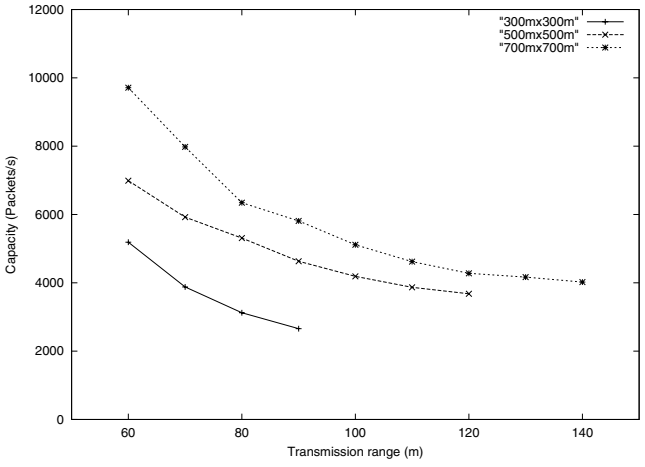

Figure 10. Total network throughput (TCP Traffic) vs. transmission range for varying density.

to maximum TCP goodput) increases. This can be explained by observing that a larger packet error rate implies a faster degradation in TCP throughput with the number of hops in a path. Thus, a value of R that is optimal for smaller p will prove sub-optimal for larger p. As R is increased, the average value of H, and hence $H * p$, the end-to-end error rate, decreases leading to more aggressive behavior. Of course, the resultant increase in the value of R^* cannot be very large, since a larger R also implies greater delays and interference at the MAC layer.

It is also interesting to see what happens if the total area A of the wireless, multi-hop network is increased without varying the total number of nodes N or the transmission range R. If η is the node density, then clearly $A = \frac{N}{\eta}$. Further, for networks where the source and destination are chosen at random, the average distance of a session, $\bar{L}$ is clearly $\propto A^{\frac{1}{2}}$. If the transmission radius R is chosen to be greater than the optimal value, then equation 13 shows that the total 'TCP capacity' is given by:

$$Cap \propto \frac{N^{\frac{1}{4}}}{R^{\frac{1}{2}}\eta^{\frac{1}{4}}} \tag{14}$$

Thus, in the fixed session, variable area and constant range framework, the capacity of the system is inverse in proportion to $\eta^{\frac{1}{4}}$, or proportional to $A^{\frac{1}{4}}$. In networks where the radio ranges cannot be adjusted, one must thus guard against packing too many nodes into too small an area.

In figure 10, we plot the system capacity versus the transmission range for varying node densities by changing the area (300 m × 300 m, 500 m × 500 m, 700 m × 700 m). The simulation is done for an error-free channel (i.e, PER = 0). It is seen from the plot that for a fixed transmission range, the capacity decreases with an increase in the density.

5.2. The variable sessions, fixed area framework

In contrast to the assumptions of the previous section, now consider an operational mode where the coverage area, A, of the ad hoc network is fixed. Further, the number of simultaneously active TCP sessions in the network, denoted by TA, is directly proportional to N, the total number of ad hoc nodes. Thus. mathematically

$$TA = \gamma * n, \tag{15}$$

where γ indicates the probability that any given node is engaged in a TCP-based transfer at any instant.

This formulation is a useful model for understanding network dynamics under certain very practical situations. Consider, for example, the problem of covering a geographic area with a certain number of sensor (say thermal sensor) nodes. Each node is autonomously programmed to periodically activate itself, monitor the temperature and communicate it to a central authority. Thus, if the communication process happens for 15 minutes every hour, we have a model where the number of active sessions is $\frac{1}{4}$th of the total number of nodes N. The network designer would clearly be interested in evaluating how his choice of the nodal density (how closely to place the wireless nodes), denoted by η, affects the achievable network capacity.

To study the dependence of total capacity on η, we make the *fundamental assumption* that a larger η leads to a smaller transmission range R. In well-designed networks, the choice of R is actually based on the need to keep the average degree of each node, defined as the number of one-hop neighbors, moderately high; in fact, classical results [10] state that the optimal number of one-hop neighbors is ∼6. As η increases, a node is able to find one-hop neighbors within a smaller radial distance, and consequently, can lower its transmission radius.

Then, since each TCP session, by our previous section, has $\rho \propto$ either $\left(\frac{R}{\bar{L}}\right)^{\frac{3}{2}}$ (for moderate values of p) or $\propto \left(\frac{R}{\bar{L}}\right)$ (for low values of p), it follows that the total capacity utilized by the ad hoc network is then:

$$\begin{aligned} Cap &\propto \gamma * \eta * A * \left(\frac{R}{\bar{L}}\right)^{\frac{3}{2}} \text{ for moderate } p \\ &\propto \gamma * \eta * A * \left(\frac{R}{\bar{L}}\right) \text{ for very low } p \end{aligned} \tag{16}$$

We consider a fixed area A and progressively increase ad hoc node density η. Since the transmission radius needed to maintain a constant nodal degree decreases as the square-root of the number of nodes, it is easy to see that node density and the transmission radius are related as

$$R \propto \frac{1}{\sqrt{\eta}}$$

By substituting this into equation (16), we finally get the 'capacity' of the TCP-based ad hoc network as

$$Cap \propto \frac{\gamma * \eta * A}{\eta^{\frac{3}{4}} * \bar{L}^{1.5}} \propto \frac{\gamma * A * \eta^{\frac{1}{4}}}{\bar{L}^{\frac{3}{2}}}, \tag{17}$$

or,

$$Cap \propto \frac{\gamma * \eta * A}{\eta^{\frac{1}{2}} * \bar{L}^{0.5}} \propto \frac{\gamma * A * \eta^{\frac{1}{2}}}{\bar{L}^{\frac{1}{2}}}, \tag{18}$$

where equation (17) holds for moderately low values of link error rates, and equation (18) holds for very low values of link error rates.

To illustrate the validity of our conclusions, we ran simulations where the area was kept constant and the number of nodes was progressively increased. Figure 11 plots the TCP throughput against the logarithm of the node density, for an operating environment where the link packet error rate (i.i.d.) was only 0.001 and T × Thresh = 1. The slope of the graph in this case is ∼ 0.6, showing the applicability of equation (18) to this case (since the effective end-to-end error rate was very low).

It is interesting to contrast these results with those on the idealized link capacity in [11], which showed that, under similar operating conditions, the idealized link-layer network capacity would increase as $O(\sqrt{\eta})$. Clearly, the bursty nature

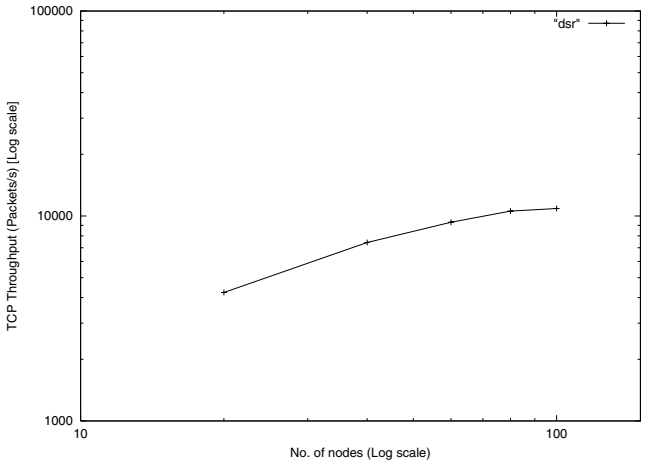

Figure 11. Cumulative TCP throughput vs. node density (log scale) for optimum range.

of TCP traffic (which prevents us from indiscriminately increasing the total number of sessions), and the dependence of TCP session throughput on the link error rate can prevent TCP-based data traffic from achieving this ideal value.

6. Sensitivity of capacity to network parameters, routing protocols and applications

The previous sections of the paper dealt with energy efficiency and transport-layer throughput in a *static*, multi-hop network with N nodes. We had also assumed a greedy model (FTP transferring a large file) for source traffic, with TCP's congestion window acting as the only constraint on the injection of new packets into the network. In this section, we firstly study the TCP-centric transmission capacity in ad hoc networks for different rates of node mobility and for varying number of TCP connections. Studying the cumulative network throughput as a function of solely the number of TCP flows can be misleading; while the total throughput may increase, individual flows may see an unacceptable degradation in their quality of service. Accordingly, we incorporate bounds on the maximum acceptable loss rate in our computation on the maximum achievable cumulative TCP throughput. We then study the dependence of capacity on the source application (Telnet or FTP traffic). As we shall see, applications, such as Telnet, with very intermittent traffic, result in very low offered loads. In such environments, the network is almost never MAC-layer constrained, but is largely *source constrained.* Accordingly, the cumulative throughput behavior in this case is very different from scenarios involving greedy (FTP) sources. Finally, we investigate the manner in which this computed capacity varies with the choice of a specific ad-hoc routing protocol. In principle, it is clear that different routing protocols (e.g., DSR [9], DSDV [15], AODV [14], etc) result in the selection of different paths, and will, consequently, result in different values for the total throughput. Through our studies, we primarily examine the *sensitivity* of our throughput results to variations in the ad-hoc routing protocol; for example, does the optimal value of the transmission radius (which results in peak network capacity) vary appreciably across different routing protocols?

6.1. Simulation parameters

The performance studies in this section are also carried out using simulations performed on the *ns-2* simulator [21]. While we have experimented with a variety of node densities and layouts, we report all results using a representative 50 node ad hoc network. The nodes are distributed randomly and move about in an area of 500 m $\times$ 500 m. For our studies, we set the interference range to be twice the transmission range. A fixed number of TCP connections are run for a duration of 500 seconds and the capacity is calculated by summing the TCP goodputs over all the connections. Results are averaged over a minimum of 10 separate runs. While TCP Reno is used as the transport layer, the data sources ("the application") are chosen to be either persistent (FTP) or intermittent (Telnet). Unless otherwise specified, results are reported using DSR as the ad hoc routing protocol. Node mobility is modeled using the Random Waypoint model [2], with the *pause time* of all nodes set to 0 in all simulations.

6.2. Capacity with varying node mobility

To begin with, we study the effects of mobility on two different classes of application – persistent and non-persistent. While the persistent traffic (FTP) is greedy and attempts to inject packets whenever permitted by TCP's congestion window, the non-persistent traffic (Telnet) produces only sporadic bursts of packets. Hence, as will be seen later, while the effects of interference are clearly visible in the case of FTP, the MAC-layer interference is not so critical in applications such as Telnet.

The capacity of a network with 40 FTP connections with different mobilities has been plotted in figure 12. In figure 13, we plot capacity versus transmission range with varying mobility for Telnet traffic. The speed of a node is uniformly distributed between 0 m/s and a maximum value (shown in the figures).

In figure 12, we see that the capacity of the network decreases with increasing node speed. Clearly, the overhead of route re-establishment, and the fraction of packets dropped due to routing failures, increases with with increasing node mobility. Furthermore, the optimal transmission radius R^* (corresponding to maximum capacity) shifts to the right (i.e., R^* is higher) with an increase in the node speed. In other words, we need a higher transmission range to counteract the high mobility in the network. Note that, as in the case of a static topology used in figure 8, the shape of the capacity versus transmission range plot is bell-shaped for mobile

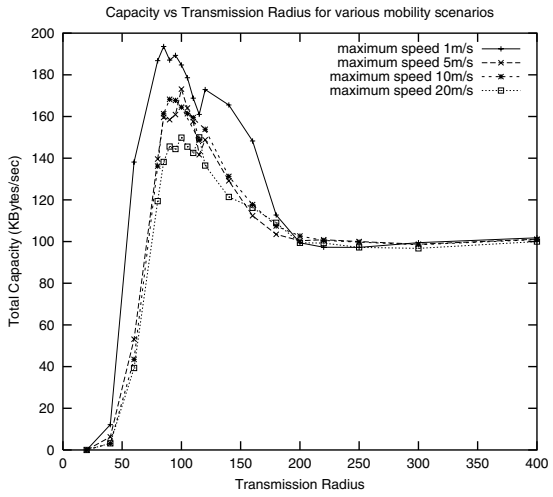

Figure 12. Network capacity vs. transmission radius with varying speed (FTP Traffic).

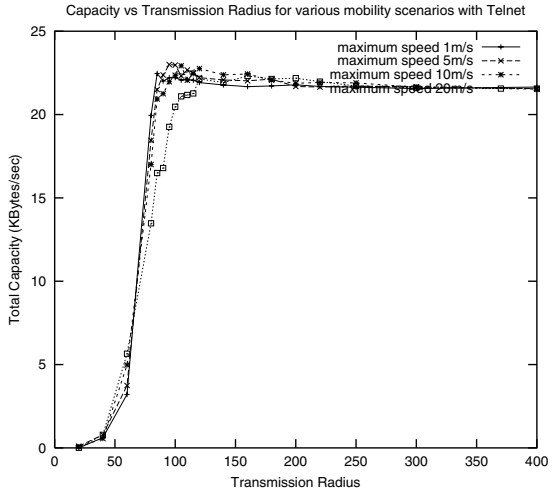

Figure 13. Network capacity vs. transmission radius with varying speed (Telnet traffic).

networks as well. Moreover, we observe that the increase in R^* with larger mobility rates is not very dramatic; accordingly, it appears that a single well-chosen value of R^* will ensure reasonably good (although not necessarily optimal) network performance, even if the node speeds cannot be predicted precisely.

For Telnet traffic (figure 13) and fixed node speeds, the capacity increases with increasing transmission radius till a value of R', after which the capacity does not change appreciably with R. Since telnet traffic is sporadic in nature, we do not observe the interference effect visible with FTP. In other words, due to the lower average packet arrival rate, the network is never MAC-layer constrained; even at large values of R, there are very few requests for concurrent access to the 802.11 channel. While the number of non-interfering concurrent transmissions possible in the network does dip with an increase in R, the telnet goodput remains unaltered. This is also the reason that the network capacity with Telnet application is significantly lower than that with the FTP (persistent) source. Increasing R beyond R' does not result in any further increase in the throughput; the number of packets transferred in a single burst is usually too small to allow TCP to take advantage of the smaller loss probability and round-trip delays. Hence for light non-persistent traffic, the TCP goodput depends solely on the connectivity of the network.

6.3. Varying number of TCP connections

We now study how changes to TA, the number of active TCP flows, affects the overall system throughput. In general, we can clearly expect the TCP goodput for an individual session to degrade with an increase in the offered load. In essence, an increase in TA leads to a potential increase in both p and RTT, since the larger load leads to more frequent buffer overflow and larger buffering delays. Accordingly, equation 8 implies a drop in the TCP throughput. However, the effect on the overall system capacity is unclear, since this reduction may or may not be offset by an increase in the number of distinct flows.

This phenomenon is studied in figure 14 where the number of FTP connections is varied from 5 to 2000 in an ad hoc network with 50 nodes. Node speeds are uniformly distributed between 0 m/s and 1 m/s. As the number of TCP connections increase, the network capacity increases initially. However, the capacity begins to degrade beyond 750 TCP connections in the ad hoc network. It is worth noting that the drop in throughput for values of R larger than R^* is more acute for larger values of TA. When the network becomes MAC-constrained and nodes must perform exponential backoff more frequently to access the channel, the individual nodes are unable to clear their packet buffers at a sufficiently high rate. Accordingly, the buffering losses and delays are higher for higher values of

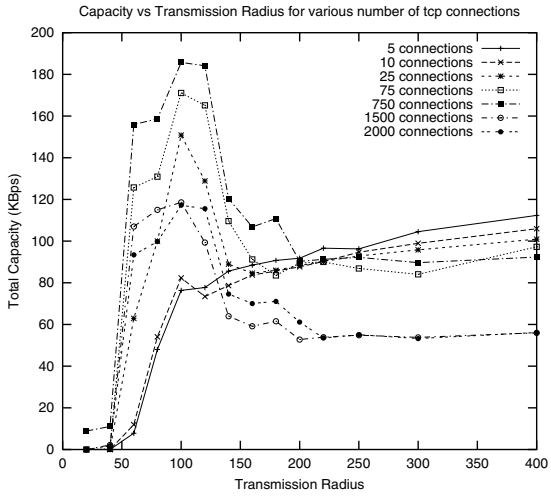

Figure 14. Capacity with varying FTP connections.

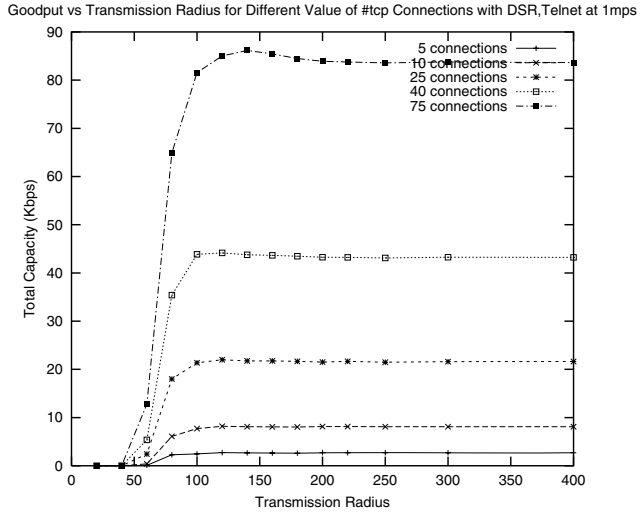

Figure 15. Capacity with varying Telnet connections.

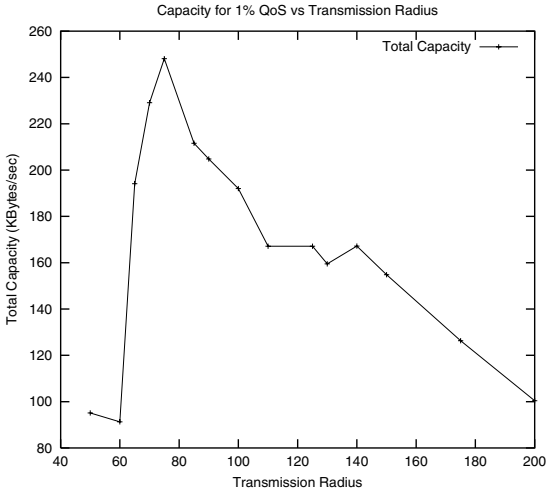

Figure 16. Total QoS-compliant network capacity vs. transmission radius.

TA, leading to a sharp drop in system throughput–in essence, the system is now nearer to *congestion collapse*. Figure 14 thus illustrates an important point: *if the number of persistent TCP flows cannot be accurately estimated in advance, it is better to adopt a conservative approach and set R to a smaller value. If the chosen value of R is larger than R^*, the network suffers a much stiffer penalty.*

Figure 15 plots the results obtained by varying the number of Telnet sessions. Due to the rather sporadic injection of packets, the overall traffic load is always rather low for Telnet sources. Accordingly, the network is always *source-constrained*, even at large values of R. Accordingly, the capacity of the network is seen to linearly increase with an increase in the number of TCP sessions. As seen earlier in figure 13, the system capacity saturates at a certain value R' of the transmission range.

6.4. QoS-compliant capacity

We have seen that an unbounded increase in the number of persistent (FTP) connections can ultimately lead to a drop in the system capacity. Figure 14, however, does not consider the associated issue of QoS; in particular, it does not incorporate the fact that an increase in the number of sessions, typically leads to a decrease in the performance metrics of an individual session. For a more accurate characteristics of the maximum *permissible* TCP throughput, we have to limit the maximum number of active sessions to ensure that the quality of an individual session does not degrade below an acceptable threshold.

We now attempt to answer the question—what is the maximum number of TCP connections (and the resulting QoS-compliant network throughput) that can be sustained in an ad hoc network with K nodes, without causing a violation of the QoS metrics of each individual connection? This leads to the notion of *QoS-compliant capacity*, or the maximum total throughput that may be achieved without causing the QoS bounds of any individual flow to be violated. As an illustration of possible QoS constraints, we consider the metric of *packet loss rate* and assume that the maximum acceptable loss rate is fixed at 1%. Metrics other than the loss rate can also be considered, but are often too application-specific. For example, different applications often have appreciably different bounds on the permissible end-to-end delay; moreover, most TCP applications are fairly insensitive to delay variations among individual packets.

In figure 16, we plot the maximal total capacity (subject to the constraint of an upper bound of 1% on the packet loss rate) versus the transmission radius R for our 50 node network. We see that capacity is maximum when R is approximately 75 meters. Figure 17 plots the maximum permissible number of TCP connections (subject to the 1% loss constraint) versus R.

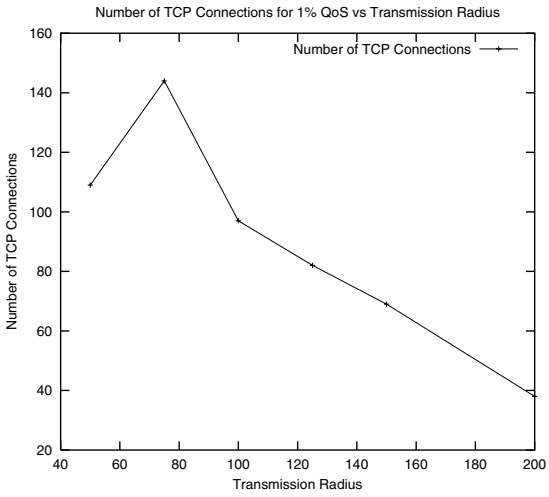

Figure 17. Maximum number of QoS-compliant TCP connections vs. transmission radius.

We see that the maximum acceptable number of TCP connections shows a peak as well, and begins to drop fairly sharply as R is increased beyond an optimal value. Further, we see that a transmission radius of ~ 75 meters corresponds to the maximum number of TCP connections, i.e., ~ 140. Note that figure 17 conveys more information than figure 16: it enables us to obtain an upper bound on the number of simultaneous TCP connections (for 1% loss bound) that can be permitted in a K node ad hoc network. These graphs also convey the appropriate value of the transmission radius $R^\star$ that yields maximum system capacity and maximum number of TCP connections. It is therefore clear that, even with QoS constraints imposed in the network, the total capacity versus R behavior exhibits the bell-shaped behavior seen earlier in figures 8 and 14.

6.5. Ad hoc routing protocols

In this section, we investigate whether the shape of the capacity versus transmission radius curve is affected by a change in the choice of the ad hoc routing protocol. We also examine whether the optimal transmission radius, R^*, is appreciably different for different ad hoc routing protocols.

Several earlier studies (e.g., [2,8]) have compared the performance of different ad hoc routing protocols. These studies have, however, primarily studied the variation in throughput and loss rates as a function of the node mobility rates and network density, but not as a function of the transmission range. Our primary aim in this section is not to perform a comparative study of the routing protocols with varying R. Instead, we investigate whether the bell-shaped curves obtained earlier using the DSR routing protocol hold true for other well known ad hoc routing protocols such as DSDV [15] and AODV [14]. We compare DSDV, DSR and AODV in a 50-node network with (i) low node mobility (figure 18) and (ii) high node mobility (figure 19). Both these graphs correspond to FTP traffic.

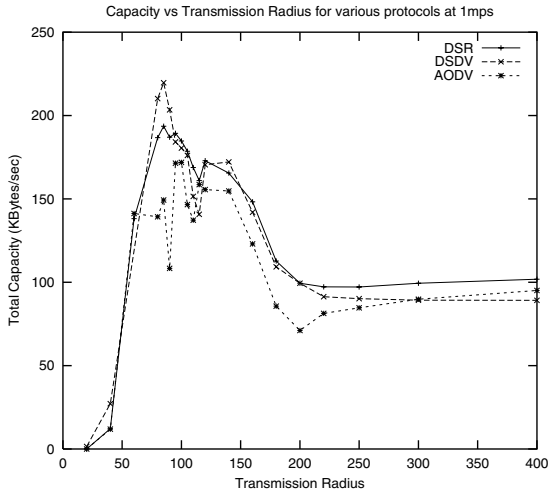

Figure 18. Capacity with different ad hoc routing protocols (speed: 1 m/s).

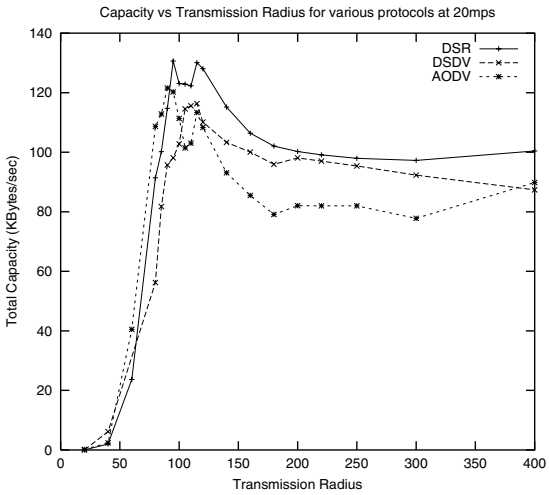

Figure 19. Capacity with different ad hoc routing protocols (speed: 20 m/s).

It can be seen that for almost all values of the transmission radii, AODV protocol yields the lowest capacity as compared to the DSDV and the DSR protocols. The difference in capacity of DSR, AODV and DSDV protocols is more pronounced at higher mobility and at higher transmission ranges. However, for our purposes, it is more important to note that *the shape of the capacity versus R curve and the value of the optimum transmission range R^*, is fairly similar for all three ad hoc routing protocols.* In other words, the results of our capacity analysis are fairly independent (at least qualitatively) of the precise choice of the ad hoc routing protocol. Since routing protocols will continue to evolve with time, verifying this protocol-independence is essential to making our results and observations meaningful for future ad hoc networks.

7. Conclusion

In this paper, we focus on the theoretical performance of TCP traffic over a multi-hop, wireless network where all links share the same physical channel. In contrast to earlier studies that largely focus on the throughput achievable at the MAC layer, we concentrate on the goodput achievable at the TCP layer. Our studies show that the TCP-layer throughput is a function of the transmission range, R, of an individual node, since this parameter directly affects the path length (in terms of hop count), and implicitly, the loss rate and round trip delay of a TCP session.

Our most important contribution lies in establishing how the TCP-centric capacity differs from the idealized MAC-layer capacity. The achievable TCP throughput depends on the tradeoff between two mutually antagonistic effects:

(a) the degree of spatial reuse, which determines the number of permitted concurrent transmissions, and

(b) TCP's flow control algorithm, which reduces the packet transmission rate in reaction to packet losses or increases in the round trip delay.

A smaller value of R results in a smaller interference area, and thus a larger value of the maximal MAC-layer throughput. However, a smaller transmission radius also increases the loss rate and RTT encountered by TCP packets, both of which lead to a reduction in TCP's transmission rate. In general, the goodput of an individual TCP flow decreases as either $O(\frac{1}{H^{\frac{3}{2}}})$, or as $O(\frac{1}{H})$ (where H is the number of hops), depending on the link error rate. The overall network throughput thus exhibits a bell-shaped curve with an optimal value R^* for R; for $R > R^*$,the network is MAC-constrained, while for $R < R^*$, the network is TCP-constrained. Due to these constraints, the maximal TCP-layer network throughput (when the number of sessions is variable) varies between between $O(N^{\frac{1}{4}})$ and $O(N^{\frac{1}{2}})$, in contrast to the MAC-layer bound of $O(N^{\frac{1}{2}})$ obtained in earlier studies. On the other hand, when the area of the network and the number of active sessions is fixed, the capacity is a concave function of the transmission range.

We have also showed how the communication energy-efficiency is also strongly dependent on the transmission radius R. When packet retransmissions, needed for reliable communication, and end-to-end latencies are taken into account, we see that it does not pay to reduce R beyond a certain value—any apparent reduction in the transmission energy of a single packet is swamped out by an increase in the number of retransmissions and the overall communication latency. Our simulation studies also show that the energy consumed per byte is minimized for a certain transmission range R_e, that can often be distinct from R^*. Accordingly, the choice of a transmission range R, at least in the range (R^*, R_e), essentially involves a *tradeoff between network throughput and energy efficiency*.

In the second part of the paper, we studied the sensitivity of TCP capacity to various network parameters (node mobility, number of TCP connections), different ad hoc routing protocols and different applications (Telnet, FTP). Our results show the existence of a sharply defined optimal transmission range R^* in the case of persistent (FTP) traffic; for Telnet traffic, the system capacity increases with increasing R and eventually saturate at a value R'. Moreover, we have observed that R^* is higher for higher mobility rates–clearly, a larger R helps to reduce the frequency of mobility-related link breakages and the consequent loss of data packets. By incorporating the notion of a minimal acceptable QoS metric (loss) for an individual session, we defined and studied the *QoS-compliant capacity* as a more accurate metric of network performance. Our simulations demonstrated that the QoS-compliant capacity is a bell-shaped function of the transmission range R and exhibits a *rapid decrease* if the transmission range exceeds an optimal value $R^\star$. Accordingly, if the network load cannot be estimated accurately in advance, it is *better to set the transmission range and power level of the ad hoc nodes to a smaller, rather than a larger, value.*

We expect that the work in this paper will yield useful insights into the performance of multi-hop wireless networks. In future work, we propose to study QoS-compliant capacity for additional applications such as HTTP, each of which has its own unique packet arrival pattern and distinct QoS constraints. Our studies also need to be extended to cover UDP traffic and UDP-based applications, which often have stringent constraints on additional QoS metrics such as delay, jitter and packet loss. The results in this paper assume the Random Waypoint model. In a recent paper [16], the authors explore the tradeoff between the number of hops in a traffic path and and the overall bandwidth available to individual nodes as the transmission power is varied. These results assumed UDP traffic and a *modified random direction* mobility model [16]. It will be interesting to study the sensitivity of our TCP-based studies to different mobility models.

Appendix

In this appendix, we derive the expression for the total number of packet transmissions necessary for reliable delivery of a packet over an H hop path. The packet error rate for each hop is p and the maximum number of retransmissions at the link layer is max.

Since reliable link forwarding fails only when all max transmissions fail, the unconditional probability of link packet transmission failure, which we call q, is given by $q = p^{\max}$; the corresponding probability of reliable link delivery (potentially using between $(1, \ldots, max)$ transmissions) is then $1 - q$. Since the total number of link transmissions, given that the link has reliably forwarded the packet, is a truncated geometric distribution with parameter p, the *conditional* expected number of transmissions, T_{good}, over a single link, is given by:

$$T_{\text{good}} = \sum_{i=1}^{\max} i * p^i * (1-p) = \frac{1}{1-p} - \frac{\max * p^{\max}}{1-p^{\max}}. \quad (19)$$

Since link packet delivery fails only after exactly max transmissions, the corresponding *conditional* number of transmissions, given forwarding failure is:

$$T_{\text{bad}} = max.$$

Now since each link fails to forward the packet independently with q, the unconditional probability of successful end-to-end delivery (without another source retransmission) is given by $P_{\text{succ}} = (1-q)^H$, and the unconditional probability of unsuccessful end-to-end delivery is given by

$$P_{\text{fail}} = 1 - (1-q)^H. \quad (20)$$

Next, we determine the expected number of total packet transmissions (over all the links that attempted to transmit a

packet), $T_{\text{bad}}^{\text{total}}$, *given that the end-to-end forwarding attempt was unsuccessful*. Since a downstream node forwards packets only when all the upstream nodes successfully transmitted the packet, it is easy to see that the conditional probability that failure occurs at the ith link is given by:

$$Prob_{fail}(i \mid \text{end-to-end failure}) = \frac{Prob_{\text{fail}}(i)}{P_{\text{fail}}} = \frac{(1-q)^{i-1} * q}{P_{\text{fail}}}.$$

If failure occurs at the ith link, the expected number of total link-layer transmissions (over all the upstream nodes) is $(i-1) * T_{\text{good}} + T_{\text{bad}}$. Accordingly, the conditional mean number of total link-layer transmissions during link failure is:

$$T_{\text{bad}}^{\text{total}} = T_{\text{bad}} + T_{\text{good}} * (1-q) * \left\{ \frac{1 - H * (1-q)^{H-1} + (H-1) * (1-q)^{H}}{q * \{1-(1-q)^{H}\}} \right\}. \quad (21)$$

On the other hand, if the packet has been successfully received at the end-destination, it is clear that the total expected transmission energy is

$$T_{\text{good}}^{\text{total}} = H * T_{\text{good}}. \quad (22)$$

Since each end-to-end transmission attempt (initiated at the transport layer by the source) is independent of prior end-to-end retransmissions, the total number of end-to-end transmissions for reliable delivery is geometrically distributed with a mean of $\frac{1}{1-P_{\text{fail}}}$; hence, on average, the successful transmission of a packet involves $\frac{1}{1-P_{\text{fail}}} - 1$ failed end-to-end transmissions, followed by the final successful one. Accordingly, the total effective number of distinct packet transmissions is

$$T = T_{\text{bad}}^{\text{total}} * \frac{P_{\text{fail}}}{1 - P_{fail}} + T_{\text{good}}^{\text{total}}, \quad (23)$$

where $T_{\text{bad}}^{\text{total}}$, $T_{\text{good}}^{\text{total}}$ and P_{fail} are given by equations (23), (22) and (20) respectively.

Acknowledgments

We would like to acknowledge the efforts of Imran Ali, Shobhit Chugh, Anurag Goel, Kapil Kumar (CSE department, Indian Institute of Technology, New Delhi), Ashu Razdan (UCLA, USA) and Rajeev Gupta (IBM India Research Laboratory, New Delhi) for helping us with many simulation results in the paper.

References

[1] S. Banerjee and A. Misra, Minimum energy paths for reliable communication in multi-hop wireless networks, to appear in *ACM Mobihoc* (June 2002).

[2] J. Broch, D. Maltz, D. Johnson, Y. Hu and J. Jetcheva, A performance comparison of multi-hop wireless ad hoc network routing protocols, Proceedings of ACM Mobicom (1998).

[3] S. Das, C. Perkins and E. Royer, Performance comparison of two on-demand routing protocols for ad-hoc networks, in: *Proceedings of IEEE INFOCOM 2000* (March 2000).

[4] S. Floyd, Connections with Multiple congested gateways in packet-switched networks Part 1: One-way traffic, in: *Computer Communication Review* (Oct. 1991).

[5] J. Gomaz, A. Campbell, M Naghdhineh and C. Bisdikian, Conserving transmission power in wireless ad hoc networks, in: *Proceeding of IEEE International Conference on Network Protocols (ICNP)* (Nov. 2001).

[6] P. Gupta and P.R. Kumar, The capacity of wireless networks, in: *IEEE Transactions on Information Theory* (March 2000).

[7] IEEE Computer Society LAN MAN Standards Committee, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, *IEEE Std. 802.11-1197*.

[8] P. Johansson, T. Larsson, N. Hedman and B. Mielczarek, routing protocols for mobile ad-hoc networks—a comparative performance analysis, in: *Proceedings of ACM MOBICOM'99* (Aug. 1999).

[9] D. Johnson and D. Maltz, Dynamic source routing in ad hoc wireless networks, in: *Mobile Computing*, chapter 5, Kluwer Academic Publishers (1996) pp. 153–181.

[10] L. Kleinrock and J. Silvester, Optimum transmission radii for packet radio networks or why six is a magic number, in: *Proceedings of the IEEE National Telecommunications Conference*, (Dec. 1978).

[11] J. Li, et al., Capacity of Ad Hoc Wireless Networks, in: *Proceedings of ACM MOBICOM '01* (July 2001).

[12] M. Matthis, J. Semke, J. Mahdavi and T. Ott, The macroscopic behavior of the TCP congestion avoidance algorithm, in: *Computer Communications Review* (July 1997).

[13] J. Padhye, V. Firiou, J. Kurose and D. Towsley, Modeling TCP throughput: A simple model and its empirical validation, in: *Proceedings of ACM SIGCOMM'98*, (Sept. 1998).

[14] C. Perkins, E. Belding-Royer and S. Das, Ad hoc on-deman distance vector (AODV) routing, draft-ietf-manet-aodv-09.txt, IETF, Work in Progress, (Nov. 2001).

[15] C. Perkins and P. Bhagwat, Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers, in: *Proceedings of ACM SIGCOMM '94* (Aug. 1994).

[16] E.M. Royer, P.M. Melliar-Smith and L.E. Moser, An analysis of the optimum node density for ad hoc mobile networks, in: *Proceedings of IEEE ICC'01*, Helsinki, Finland (June 2001).

[17] T. Shepard, A channel access scheme for large dense packet radio networks, in: *Proceedings of ACM SIGCOMM'96* (Aug. 1996).

[18] S. Singh and C.S. Raghavendra, PAMAS-poweraware multi-access protocol with signaling for ad hoc networks, *ACM Communication Review*, (July 1998).

[19] S. Singh, M. Woo and C.S. Raghavendra, Power-aware routing in mobile ad hoc networks, in: *Proceedings of ACM MOBICOM '98* (Oct. 1998).

[20] M. Stemm and R. Katx, Measuring and reducing energy consumption of network interfaces in hand-held devices, IEICE Transactions on Communication E80-B(8) (1997).

[21] The ns-2 network simulator, http://www.mash.cs.berkeley.edu/ns.

[22] C.K. Toh, H. Cobb and D. Scott, Performance evaluation of battery-life-aware routing schemes for wireless ad hoc networks, in: *Proceedings of IEEE ICC'2001* (June 2001).

Sorav Bansal is a graduate student in the Electrical Engineering Department at Stanford University, where he is currently working on Participation Incentives in Mobile Networks. His primary research interests lie in Mobile Computing, Embedded Devices and Sensor Networks. Sorav holds a B.Tech. in Computer Science from Indian Institute of Technology, Delhi, and spent a summer interning at IBM Almaden Research Center.
E-mail: soravban@in.ibm.com

Rajeev Shoreyis a research staff member at the IBM India Research Laboratory, New Delhi since March 1998. He received the Bachelor of Engineering (B.E) degree in Computer Science from the department of Computer Science and Automation, Indian Institute of Science, Bangalore, India in 1987. He received the M.S and Ph.D degrees in Electrical Communication Engineering from the Indian Institute of Science, Bangalore, India, in 1990 and 1996 respectively. Since March 1998, he is a Research Staff Member at the IBM India Research Laboratory, Indian Institute of Technology, New Delhi, India. His research interests include wireless LANs and wireless PANs, Internet protocols, performance modeling and analysis of wireline and wireless networks. Dr. Shorey has published numerous papers in international journals and conferences. He has to his credit one IBM US patent and around 8 US patents that are pending, all in the area of networking. He serves on the technical program committee of several international conferences in networking, namely, IEEE Infocom 2004 and IEEE ICC 2004. In the past, he was serving in the technical program committee for Infocom 2003 and Infocom 2002, Globecom 2002 and Globecom 2001 and ICC 2003. He is an adjunct faculty in the department of Computer Science and Engineering, Indian Institute of Technology, New Delhi where he actively teaches and guides undergraduate and graduate students. He is a senior member of IEEE.
E-mail: srajeev@in.ibm.com

E-mail: grajeev@in.ibm.com

Archan Misra is currently a Research Staff Member with the Pervasive Security and Networking Department at the IBM TJ Watson Research Center, Hawthorne, NY. He is presently working on services and mobility protocols for next-generation (4G) wireless networks, middleware for location and context-aware data composition, scalable infrastructure for on-demand distributed computing and MAC/routing protocols for energy-efficient, high-performance wireless networks. Before joining IBM in March 2001, Archan spent 3 1/2 years at Telcordia Technologies (formerly called Bellcore), where he was responsible for several initiatives in the areas of IP-based mobility management, congestion control, QoS architectures and autoconfiguration of heterogeneous networks. As part of his research efforts, Archan co-invented the IDMP mobility management and the DCDP autoconfiguration protocols. He has published over fifty papers in the areas of wireless networking, congestion control and mobility management and received the Best Paper awards in ACM WOWMOM 2002 and IEEE MILCOM 2001. Archan received his Ph.D. in Electrical and Computer Engineering from the University of Maryland at College Park in May, 2000, and his B.Tech in Electronics and Communication Engineering from IIT Kharagpur, India in July 1993.
E-mail: archan@in.ibm.com