A Distributed Three-hop Routing Protocol to Increase the Capacity of Hybrid Wireless Networks

Haiying Shen*, *Senior Member, IEEE*, Ze Li and Chenxi Qiu

Abstract—Hybrid wireless networks combining the advantages of both mobile ad-hoc networks and infrastructure wireless networks have been receiving increased attention due to their ultra-high performance. An efficient data routing protocol is important in such networks for high network capacity and scalability. However, most routing protocols for these networks simply combine the ad-hoc transmission mode with the cellular transmission mode, which inherits the drawbacks of ad-hoc transmission. This paper presents a Distributed Three-hop Routing protocol (DTR) for hybrid wireless networks. To take full advantage of the widespread base stations, DTR divides a message data stream into segments and transmits the segments in a distributed manner. It makes full spatial reuse of a system via its high speed ad-hoc interface and alleviates mobile gateway congestion via its cellular interface. Furthermore, sending segments to a number of base stations simultaneously increases throughput and makes full use of widespread base stations. In addition, DTR significantly reduces overhead due to short path lengths and the elimination of route discovery and maintenance. DTR also has a congestion control algorithm to avoid overloading base stations. Theoretical analysis and simulation results show the superiority of DTR in comparison with other routing protocols in terms of throughput capacity, scalability and mobility resilience. The results also show the effectiveness of the congestion control algorithm in balancing the load between base stations.

Index Terms—Hybrid wireless networks, Routing algorithm, Load balancing, Congestion control

1 Introduction

Over the past few years, wireless networks including infrastructure wireless networks and mobile ad-hoc networks (MANETs) have attracted significant research interest. The growing desire to increase wireless network capacity for high performance applications has stimulated the development of hybrid wireless networks [1–6]. A hybrid wireless network consists of both an infrastructure wireless network and a mobile ad-hoc network. Wireless devices such as smart-phones, tablets and laptops, have both an infrastructure interface and an ad-hoc interface. As the number of such devices has been increasing sharply in recent years, a hybrid transmission structure will be widely used in the near future. Such a structure synergistically combines the inherent advantages and overcome the disadvantages of the infrastructure wireless networks and mobile ad-hoc networks.

In a mobile ad-hoc network, with the absence of a central control infrastructure, data is routed to its destination through the intermediate nodes in a multi-hop manner. The multi-hop routing needs on-demand route discovery or route maintenance [7–10]. Since the messages are transmitted in wireless channels and through dynamic routing paths, mobile ad-hoc networks are not as reliable as infrastructure wireless networks. Furthermore, because of the multi-hop transmission feature, mobile ad-hoc networks are only suitable for local area data transmission.

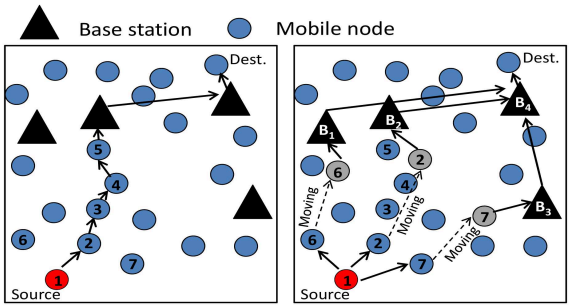
The infrastructure wireless network (e.g. cellular network) is the major means of wireless communication in our daily lives. It excels at inter-cell communication (i.e., communication between nodes in different cells) and Internet access. It makes possible the support of universal network connectivity and ubiquitous computing by integrating all kinds of wireless devices into the network. In an infrastructure network, nodes communicate with each other through base stations (BSes). Because of the long distance one-hop transmission between BSes and mobile nodes, the infrastructure wireless networks can provide higher message transmission reliability and channel access efficiency, but suffer from higher power consumption on mobile nodes and the single point of failure problem [11].

A hybrid wireless network synergistically combines an infrastructure wireless network and a mobile ad-hoc network to leverage their advantages and overcome their shortcomings, and finally increases the throughput capacity of a wide-area wireless network. A routing protocol is a critical component that affects the throughput capacity of a wireless network in data transmission. Most current routing protocols in hybrid wireless networks [1, 5, 6, 12–18] simply combine the cellular transmission mode (i.e. BS transmission mode) in infrastructure wireless networks and the ad-hoc transmission mode in mobile ad-hoc networks [8, 9, 7]. That is, as shown in Figure 1 (a), the protocols use the multi-hop routing to forward a message to the mobile gateway nodes that are closest to the BSes or have the highest bandwidth to the BSes. The bandwidth of a channel is the maximum throughput (i.e., transmission rate in bits/s) that can be achieved. The mobile gateway nodes then forward the messages to the BSes, functioning as bridges to connect the ad-hoc network and the infrastructure network.

However, direct combination of the two transmission modes inherits the following problems that are rooted in the ad-hoc transmission mode.

- *High overhead.* Route discovery and maintenance incur high overhead. The wireless random access medium access control (MAC) required in mobile ad-hoc networks, which utilizes control handshaking and a back-off mechanism, further increases overhead.
- *Hot spots.* The mobile gateway nodes can easily become hot spots. The RTS-CTS random access, in which most traffic goes through the same gateway, and the flooding employed in mobile ad-hoc routing to discover routes

- * *Corresponding Author. Email: shenh@clemson.edu; Phone: (864) 656 5931; Fax: (864) 656 5910.*
- *The authors are with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, 29634. E-mail: {shenh, zel, chenxiq}@clemson.edu*



(a) Traditional routing algorithm on the uplink direction.

(b) Our proposed routing algorithm on the uplink direction.

Fig. 1: Traditional and proposed routing algorithms on the uplink direction.

may exacerbate the hot spot problem. In addition, mobile nodes only use the channel resources in their route direction, which may generate hot spots while leave resources in other directions under-utilized. Hot spots lead to low transmission rates, severe network congestion, and high data dropping rates.

- *Low reliability.* Dynamic and long routing paths lead to unreliable routing. Noise interference and neighbor interference during the multi-hop transmission process cause a high data drop rate. Long routing paths increase the probability of the occurrence of path breakdown due to the highly dynamic nature of wireless ad-hoc networks.

These problems become an obstacle in achieving high throughput capacity and scalability in hybrid wireless networks. Considering the widespread BSes, the mobile nodes have a high probability of encountering a BS while moving. Taking advantage of this feature, we propose a Distributed Three-hop Data Routing protocol (DTR). In DTR, as shown in Figure 1 (b), a source node divides a message stream into a number of segments. Each segment is sent to a neighbor mobile node. Based on the QoS requirement, these mobile relay nodes choose between direct transmission or relay transmission to the BS. In relay transmission, a segment is forwarded to another mobile node with higher capacity to a BS than the current node. In direct transmission, a segment is directly forwarded to a BS. In the infrastructure, the segments are rearranged in their original order and sent to the destination. The number of routing hops in DTR is confined to three, including at most two hops in the ad-hoc transmission mode and one hop in the cellular transmission mode. To overcome the aforementioned shortcomings, DTR tries to limit the number of hops. The first hop forwarding distributes the segments of a message in different directions to fully utilize the resources, and the possible second hop forwarding ensures the high capacity of the forwarder. DTR also has a congestion control algorithm to balance the traffic load between the nearby BSes in order to avoid traffic congestion at BSes.

Using self-adaptive and distributed routing with high-speed and short-path ad-hoc transmission, DTR significantly increases the throughput capacity and scalability of hybrid wireless networks by overcoming the three shortcomings of the previous routing algorithms. It has the following features:

- *Low overhead.* It eliminates overhead caused by route discovery and maintenance in the ad-hoc transmission mode, especially in a dynamic environment.
- *Hot spot reduction.* It alleviates traffic congestion at mobile gateway nodes while makes full use of channel resources through a distributed multi-path relay.
- *High reliability.* Because of its small hop path length with a short physical distance in each step, it alleviates noise and neighbor interference and avoids the adverse effect of route breakdown during data transmission. Thus, it reduces the packet drop rate and makes full use of spacial reuse, in which several source and destination nodes can communicate simultaneously without interference.

The rest of this paper is organized as follows. Section 2 presents a review of representative hybrid wireless networks and multi-hop routing protocols. Section 3 details the DTR protocol, with an emphasis on its routing methods, segment structure, and BS congestion control. Section 4 theoretically analyzes the performance of the DTR protocol. Section 5 shows the performance of the DTR protocol in comparison to other routing protocols. Finally, Section 6 concludes the paper.

2 Related Work

In order to increase the capacity of hybrid wireless networks, various routing methods with different features have been proposed. One group of routing methods integrate the ad-hoc transmission mode and the cellular transmission mode [1, 5, 6, 14, 16–18]. Dousse *et al.* [6] built a Poisson Boolean model to study how a BS increases the capacity of a MANET. Lin *et al.* [5] proposed a Multihop Cellular Network and derived its throughput. Hsieh *et al.* [14] investigated a hybrid IEEE 802.11 network architecture with both a distributed coordination function and a point coordination function. Luo *et al.* [1] proposed a unified cellular and ad-hoc network architecture for wireless communication. Cho *et al.* [16] studied the impact of concurrent transmission in a downlink direction (i.e. from BSes to mobile nodes) on the system capacity of a hybrid wireless network. In [17, 18], a node initially communicates with other nodes using an ad-hoc transmission mode, and switches to a cellular transmission mode when its performance is better than the ad-hoc transmission.

The above methods are only used to assist intra-cell ad-hoc transmission rather than inter-cell transmission. In inter-cell transmission [1, 5, 6], a message is forwarded via the ad-hoc interface to the gateway mobile node that is closest to or has the highest uplink transmission bandwidth to a BS. The gateway mobile node then forwards the message to the BS using the cellular interface. However, most of these routing protocols simply combine routing schemes in ad-hoc networks and infrastructure networks, hence inherit the drawbacks of the ad-hoc transmission mode as explained previously.

DTR is similar to the Two-hop transmission protocol [19] in terms of the elimination of route maintenance and the limited number of hops in routing. In Two-hop, when a node's bandwidth to a BS is larger than that of each neighbor, it directly sends a message to the BS. Otherwise, it chooses a neighbor with a higher channel and sends a message to it, which further forwards the message to the BS. DTR is different from Two-hop in three aspects. First, Two-hop only considers the node transmission within a single cell, while DTR can also deal with inter-cell transmission, which is more challenging and more common than intra-cell communication in the real world. Second, DTR uses distributed transmission involving multiple cells, which makes full use of system resources and dynamically balances the traffic load between neighboring cells. In contrast, Two-hop employs single-path transmission.

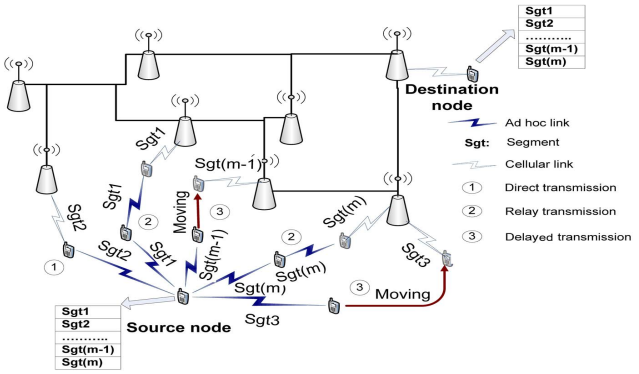


Fig. 2: Data transmission in the DTR protocol.

There are other methods proposed to improve routing performance in hybrid wireless networks. Wu *et al.* [3] proposed using ad-hoc relay stations to dynamically relay traffic from one cell to another in order to avoid traffic congestion in BSes. Li *et al.* [20] surveyed a number of multi-hop cellular network (MCN) architectures in literature, and compared and discussed methods to reduce the cost of deployment for MCNs. The work in [21] investigates how to allocate the bandwidth to users to improve the performance of hybrid wireless networks. Thulasiraman *et al.* [22] further considered the wireless interference in optimizing the resource allocation in hybrid wireless networks. The work in [23] proposes a coalitional game theory based cooperative packet delivery scheme in hybrid wireless networks. There are also some works [24–26] that study radio frequency allocation for direction transmission and relay transmission in hybrid wireless networks. These works are orthogonal to our study in this paper and can be incorporated into DTR to further enhance its performance.

The throughput capacity of the hybrid wireless network under different settings has also been an active research topic in the hybrid wireless network. The works in [17, 27] have studied the throughput of hybrid network with n nodes and m stations. Liu *et al.* [28] theoretically studied the capacity of hybrid wireless networks under an one-dimensional network topology and a two-dimensional strip topology. Wang *et al.* [29] studied the multicast throughput of hybrid wireless networks and designed an optimal multicast strategy based on deduced throughput.

3 Distributed Three-hop Routing Protocol

3.1 Assumption and Overview

Since BSes are connected with a wired backbone, we assume that there are no bandwidth and power constraints on transmissions between BSes. We use *intermediate nodes* to denote relay nodes that function as gateways connecting an infrastructure wireless network and a mobile ad-hoc network. We assume every mobile node is dual-mode; that is, it has ad-hoc network interface such as a WLAN radio interface and infrastructure network interface such as a 3G cellular interface.

DTR aims to shift the routing burden from the ad-hoc network to the infrastructure network by taking advantage of widespread base stations in a hybrid wireless network. Rather than using one multi-hop path to forward a message to one BS, DTR uses at most two hops to relay the segments of a message to different BSes in a distributed manner, and relies on BSes to combine the segments. Figure 2 demonstrates the process of DTR in a hybrid wireless network. We simplify the routings in the infrastructure network for clarity. As shown in the figure, when a source node wants to transmit a message stream to a destination node, it divides the message stream into a number of partial streams called segments and transmits each segment to a neighbor node. Upon receiving a segment from the source node, a neighbor node locally decides between direct transmission and relay transmission based on the QoS requirement of the application. The neighbor nodes forward these segments in a distributed manner to nearby BSes. Relying on the infrastructure network routing, the BSes further transmit the segments to the BS where the destination node resides. The final BS rearranges the segments into the original order and forwards the segments to the destination. It uses the cellular IP transmission method [30] to send segments to the destination if the destination moves to another BS during segment transmission.

Our DTR algorithm avoids the shortcomings of ad-hoc transmission in the previous routing algorithms that directly combine an ad-hoc transmission mode and a cellular transmission mode. Rather than using the multi-hop ad-hoc transmission, DTR uses two hop forwarding by relying on node movement and widespread base stations. All other aspects remain the same as those in the previous routing algorithms (including the interaction with the TCP layer). DTR works on the Internet layer. It receives packets from the TCP layer and routes it to the destination node, where DTR forwards the packet to the TCP layer.

The data routing process in DTR can be divided into two steps: uplink from a source node to the first BS and downlink from the final BS to the data's destination. Critical problems that need to be solved include how a source node or relay node chooses nodes for efficient segment forwarding, and how to ensure that the final BS sends segments in the right order so that a destination node receives the correct data. Also, since traffic is not evenly distributed in the network, how to avoid overloading BSes is another problem. Below, Section 3.2 will present the details for forwarding node selection in uplink transmission and Section 3.3 will present the segment structure that helps ensure the correct final order of segments in a message, and DTR's strategy for downlink transmission. Section 3.4 will present the congestion control algorithm for balancing a load between BSes.

3.2 Uplink Data Routing

A long routing path will lead to high overhead, hot spots and low reliability. Thus, DTR tries to limit the path length. It uses one hop to forward the segments of a message in a distributed manner and uses another hop to find high-capacity forwarder for high performance routing. As a result, DTR limits the path length of uplink routing to two hops in order to avoid the problems of long-path multi-hop routing in the ad-hoc networks. Specifically, in the uplink routing, a source node initially divides its message stream into a number of segments, then transmits the segments to its neighbor nodes. The neighbor nodes forward segments to BSes, which will forward the segments to the BS where the destination resides.Below, we first explain how to define capacity, then introduce the way for a node to collect the capacity information from its neighbors, and finally present the details of the DTR routing algorithm.

Different applications may have different QoS requirements, such as efficiency, throughput, and routing speed. For example, delay-tolerant applications (e.g. voice mail,

e-mail and text messaging) do not necessarily need fast real-time transmission and may make throughput the highest consideration to ensure successful data transmission. Some applications may take high mobility as their priority to avoid hot spots and blank spots. Hot spots are areas where BS channels are congested, while blank spots are areas without signals or with very weak signals. High-mobility nodes can quickly move out of a hot spot or blank spot and enter a cell with high bandwidth to a BS, thus providing efficient data transmission. Throughput can be measured by bandwidth, mobility can be measured by the speed of node movement, and routing speed can be measured by the speed of data forwarding. Bandwidth can be estimated using the non-intrusive technique proposed in [31]. In this work, we take throughput and routing speed as examples for the QoS requirement. We use a *bandwidth/queue* metric to reflect node capacity in throughput and fast data forwarding. The metric is the ratio of a node's channel bandwidth to its message queue size. A larger *bandwidth/queue* value means higher throughput and message forwarding speed, and vice versa.

When choosing neighbors for data forwarding, a node needs the capacity information (i.e., queue size and bandwidth) of its neighbors. Also, a selected neighbor should have enough storage space for a segment. To keep track of the capacity and storage space of its neighbors, each node periodically exchanges its current capacity and storage information with its neighbors. In the ad-hoc network component, every node needs to periodically send "hello" messages to identify its neighbors. Taking advantage of this policy, nodes piggyback the capacity and storage information onto the "hello" messages in order to reduce the overhead caused by the information exchanges. If a node's capacity and storage space are changed after its last "hello" message sending when it receives a segment, it sends its current capacity and storage information to the segment forwarder. Then, the segment forwarder will choose the highest capacity nodes in its neighbors based on the most updated information.

When a source node sends out message segments, it chooses the neighbors that have enough space for storing a segment, and then chooses neighbors that have the highest capacity. In order to find higher capacity forwarders in a larger neighborhood around the source, each segment receiver further forwards its received segment to its neighbor with the highest capacity. That is, after a neighbor node m_i receives a segment from the source, it uses either direct transmission or relay transmission. If the capacity of each of its neighbors is no greater than itself, relay node m_i uses direct transmission. Otherwise, it uses relay transmission. In direct transmission, the relay node sends the segment to a BS if it is in a BS's region. Otherwise, it stores the segment while moving until it enters a BS's region. In relay transmission, relay node m_i chooses its highest-capacity neighbor as the second relay node based on the QoS requirement. The second relay node will use direct transmission to forward the segment directly to a BS. As a result, the number of transmission hops in the ad-hoc network component is confined to no more than two. The small number of hops help to increase the capacity of the network and reduce channel contention in ad-hoc transmission. Algorithm 1 shows the pseudo-code for neighbor node selection and message forwarding in DTR.

The purpose of the second hop selection is to find a higher capacity node as the message forwarder in order to improve the performance of the QoS requirement. As the neighborhood scope of a node for high-capacity node searching grows, the probability of finding higher capacity nodes increases. Thus, a source node's neighbors are more likely to find neighbors with higher capacities than the source node. Therefore, transmitting data segments to neighbors and enabling them to choose the second relays help to find higher capacity nodes to forward data. If a source node has the highest capacity in its region, the segments will be forwarded back to the source node according to the DTR protocol. The source node then forwards the segments to the BSes directly due to the three-hop limit. Though sending data back and forth leads to latency and bandwidth wastage, this case occurs only when the source nodes is the highest capacity node within its two-hop neighborhood. Also, this step is necessary for finding the highest capacity nodes within the source's two-hop neighborhood, and ensures that the highest capacity nodes are always selected as the message forwarders. If the source node does not distribute segments to its neighbors, the higher-capacity node searching cannot be conducted. Note that the data transmission rate of the ad-hoc interface (e.g. IEEE 802.11) is more than 10 times faster than the cellular interface (e.g. GSM, 3G). Thus, the transmission delay for sending the data back and forth in the ad-hoc transmission is negligible in the total routing latency.

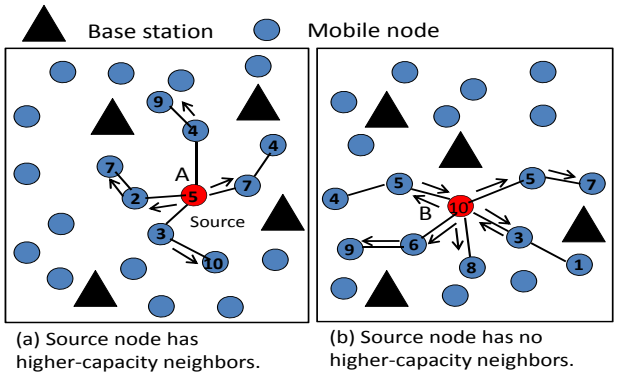


Fig. 3: Neighbor selection in DTR.

By distributing a message's segments to different nodes to be forwarded in different directions, our algorithm reduces the congestion in the previous routing algorithms in the hybrid wireless networks. When a node selects a relay to forward a segment, it checks the capacity of the node. Only when a node, say node m_i, has enough capacity, the node will forward a segment to node m_i. Therefore, even though the paths are not node-disjoint, there will be no congestion in the common sub-paths.

Figure 3 shows examples of neighbor selection in DTR, in which the source node is in the transmission range of a BS. In the figures, the value in the node represents its capacity. In scenario (a), there exist nodes that have higher capacity than the source node within the source's two-hop neighborhood. If a routing algorithm directly let a source node transmit a message to its BS, the high routing performance cannot be guaranteed since the source node may have very low capacity. In DTR, the source node sends segments to its neighbors, which further forward the segments to nodes with higher capacities. In scenario (b), the source node has the highest capacity among the nodes in its two-hop neighborhood. After receiving segments from the source node, some neighbors forward the segments back to the source node, which sends the message to its BS. Thus, DTR always arranges data to be forwarded by nodes with high

capacity to their BSes. DTR achieves higher throughput and faster data forwarding speed by taking into account node capacity in data forwarding.

Algorithm 1 Pseudo-code for neighbor node selection and message forwarding.

```
ChooseRelay( ) {
//choose neighbors with sufficient caches and bandwidth/queue (b/q) rates
Query storage size and QoS requirement info. from neighbors
for each neighbor n do
  if n.cache.size>segment.length && n.b/q>this.b/q then
    Add n to R = {r_1, ... r_m, ...} in a descending order of b/q
  end if
end for
Return R
}
Transmission( ) {
if it is a source node then
  //routing conducted by a source node
  //choose relay nodes based on QoS requirement
  R=ChooseRelay( );
  Send segments to {r_1, ... r_m} in R
else
  //routing conducted by a neighbor node
  if this.b/q ≤ b/q of all neighbors then
    //direct transmission
    if within the range of a BS then
      Transmit the segment directly to the BS
    end if
  else
    //relay transmission
    node_i=getHighestCapability(ChooseRelay( ))
    Send a segment to node_i
  end if
end if
}
```

Algorithm 2 Pseudo-code for a BS to reorder and forward segments to destination nodes.

```
//a cache pool is built for each data stream
//there are n cache pools currently
if receives a segment (S,D,m,s,q) then
  if there is no cache pool with msg sequence num equals m then
    Create a cache pool n + 1 for the stream m
  else
    //the last delivered segment of stream m has sequence num i − 1
    if s == i then
      Send out segment (S,D,m,s,q) to D
      i++;
    else
      Add segment (S,D,m,s) into cache pool m
    end if
  end if
end if
```

3.3 Downlink Data Routing and Data Reconstruction

As mentioned above, the message stream of a source node is divided into several segments. After a BS receives a segment, it needs to forward the segment to the BS, where the destination node resides (i.e., the destination BS). We use the mobile IP protocol [32] to enable BSes to know the destination BS. In this protocol, each mobile node is associated with a home BS, which is the BS in the node's home network, regardless of its current location in the network. The home network of a node contains its registration information identified by its home address, which is a static IP address assigned by an ISP. In a hybrid wireless network, each BS periodically emits beacon signals to locate the mobile nodes in its range. When a mobile node m_i moves away from its home BS, the BS where m_i currently resides detects m_i and sends its IP address to the home BS of m_i. When a BS wants to contact m_i, it contacts the home BS of m_i to find the destination BS where m_i currently resides at.

However, the destination BS recorded in the home BS may not be the most up-to-date destination BS since destination mobile nodes switch between the coverage regions of different BSes during data transmission to them. For instance, data is transmitted to BS B_i that has the data's destination, but the destination has moved to the range of BS B_j before the data arrives at BS B_i. To deal with this problem, we adopt the Cellular IP protocol [30] for tracking node locations. With this protocol, a BS has a home agent and a foreign agent. The foreign agent keeps track of mobile nodes moving into the ranges of other BSes. The home agent intercepts in-coming segments, reconstructs the original data, and re-routes it to the foreign agent, which then forwards the data to the destination mobile node.

After the destination BS receives the segments of a message, it rearranges the segments into the original message and then sends it to the destination mobile node. A vital issue is guaranteeing that the segments are combined in the correct order. For this purpose, DTR specifies the segment structure format. Each segment contains eight fields, including: (1) source node IP address (denoted by S); (2) destination node IP address (denoted by D); (3) message sequence number (denoted by m); (4) segment sequence number (denoted by s); (5) QoS indication number (denoted by q); (6) data; (7) length of the data; and (8) checksum. Fields (1)-(5) are in the segment head.

The role of the *source IP address* field is to inform the destination node where the message comes from. The *destination IP address* field indicates the destination node, and is used to locate the final BS. After sending out a message stream to a destination, a source node may send out another message stream to the same destination node. The *message sequence number* differentiates the different message streams initiated by the same source node. The *segment sequence number* is used to find the correct transmission sequence of the segments for transmission to a destination node. The *data* is the actual information that a source node wants to transmit to a destination node. The *length* field specifies the length of the DTR segment including the header in bytes. The *checksum* is used by the receiver node to check whether the received data has errors. The *QoS indication number* is used to indicate the QoS requirement of the application.

Thus, each segment's head includes the information represented by $(S, D, m, s, q)(m, s = 1, 2, 3, ...)$. When a segment with head (S, D, m, s, q) arrives at a BS, the BS contacts D's home BS to find the destination BS where D stays via the mobile IP protocol. It then transmits the segment to the destination BS through the infrastructure network component. After arriving at the BS, the segment waits in the cache for its turn to be transmitted to its destination node based on its message and segment sequence numbers. At this time, if another segment comes with a head labelled $(S, D, (m + 1), s, q)$, which means that it is from the same source node but belongs to another data stream, the BS will put it to another stream. If the segment is labeled as $(S, D, m, (s+1), q)$, it means that this segment belongs to the same data stream of the same source node as segment (S, D, m, s, q). The combination of the source node's sequence number and segment sequence number helps to locate the stream and the position of a segment in the steam. In order to integrate the segments into their correct order to retrieve the original data, the segments in the BS are transmitted

to the destination node in the order of the segments' sequence in the original message. If a segment has not arrived at the final BS, its subsequent segments will wait in the final BS until its arrival. Algorithm 2 shows the pseudo-code for a BS to reorder and forward segments to their destinations. Note that in the cache, we can set the timer based on the packet rate and storage limit. In other words, the timer should be set as large as possible to fully utilize the storage on BSes to ensure that a message has a high probability to be recovered.

3.4 Congestion Control in Base Stations

Compared to the previous routing algorithms in hybrid wireless networks, DTR can distribute traffic load among mobile nodes more evenly. Though the distributed routing in DTR can distribute traffic load among nearby BSes, if the traffic load is not distributed evenly in the network, some BSes may become overloaded while other BSes remain lightly loaded. We propose a congestion control algorithm to avoid overloading BSes in uplink transmission (e.g., B_1, B_2 and B_3 in Figure 1 (b)) and downlink transmission (e.g., B_4 in Figure 1 (b)), respectively.

In the hybrid wireless network, BSes send beacon messages to identify nearby mobile nodes. Taking advantage of this beacon strategy, once the workload of a BS, say B_i, exceeds a pre-defined threshold, B_i adds an extra bit in its beacon message to broadcast to all the nodes in its transmission range. Then, nodes near B_i know that B_i is overloaded and will not forward segments to B_i. When a node near B_i, say m_i, needs to forward a segment to a BS, it will send the segment to B_i based on the DTR algorithm. In our congestion control algorithm, because B_i is overloaded, rather than targeting B_i, m_i will forward the segment to a lightly loaded neighboring BS of B_i. To this end, node m_i first queries a multi-hop path to a lightly loaded neighboring BS of B_i. Node m_i broadcasts a query message into the system. We set the TTL for the path query forwarding step to a constant (e.g., 3). The query message is forwarded along other nodes until a node (say m_j) near a lightly loaded BS (say B_j) is reached. Due to broadcasting, a node may receive multiple copies of the same queries. Each node only remembers m_i and the node that forwards the first query (i.e., its preceding node), and ignores all other the same queries. In this way, a multi-hop path between the source node and the lightly loaded base station can be formed. Node m_j responds to the path query by adding a reply bit and the address of m_i into its beacon message to its preceding node in the path. This beacon receiver also adds a reply bit and the address of m_i into its beacon message to its preceding node in the path. This process repeats until m_i receives the beacon. Thus, each node knows its preceding node and succeeding node in the path from m_i and m_j based on the address of m_i. Then, m_i's message can be forwarded along the observed path along the nodes. The observed path can always be used by m_i for any subsequent messages to B_j as long as it is not broken. The neighboring BSes of an overloaded BS may also be overloaded. As the mobile nodes near an overloaded BS know that the BS is overloaded, when they receive a query message to find a path to an underloaded BS, they do not forward the message towards their overloaded BSes.

Node m_i may receive responses from a few nodes near BSes. It can choose b ($b \geq 1$) neighboring BSes of the destination to forward the segment. The redundant transmissions enhance the data transmission reliability while also increase the routing overhead. Thus, the value of b should be carefully determined based on the available resources for routing and the reliability demand. If b is set to a large value, the routing reliability is high at the cost of high overhead. If b is set to a small value, the routing reliability is low while the overhead is reduced. After the neighboring BSes receive the segments, they further forward the segments to the destination BS, which forwards the segments to the destination node. In this way, the heavy traffic from mobile nodes to a BS can be distributed among neighboring BSes quickly.

Next, we discuss how to handle the case when the destination BS is congested. If a BS has not received confirmation from the destination BS during a certain time period after it sends out a segment, it assumes that the destination BS is overloaded. Then, it sends the segment to b ($b \geq 1$) lightly loaded neighboring BSes of the destination BS from its routing table. If an attempted neighboring BS does not respond during a certain time period, it is also considered as overloaded. Then, the BS keeps trying other neighboring BSes until finding lightly loaded BSes. Redundant neighboring BSes are selected in order to increase routing reliability. The constant b should be set to an appropriate value considering factors such as the network size and the amount of traffic in order to achieve an optimal trade-off between overhead and reliability.

After receiving the message, each lightly loaded neighboring BS of the destination BS finds a multi-hop path to the destination mobile node. It broadcasts a path query message, which includes the IDs of the destination BS and the destination node, to the mobile nodes in its region. The path querying process is similar to the previous path querying for a lightly loaded BS. The nodes further forward the path query to their neighbors until the query reaches the destination node. Here, we do not piggyback the query to beacon messages because this querying is for a specific mobile node rather than any mobile node near a lightly loaded BS. Including the mobile node's ID into beacon messages generates very high overhead.

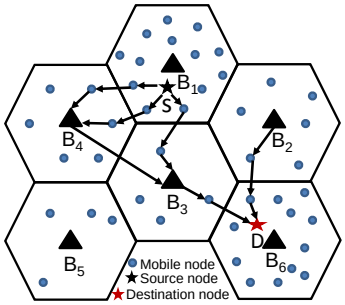


Fig. 4: Congestion control on BSes.

In order to reduce the broadcasting overhead, a mobile node residing in the region of a BS not close to the destination BS drops the query. The nodes can determine their approximate relative positions to BSes by sensing the signal strengths from different BSes. Each node adds the strength of its received signal into its beacon message that is periodically exchanged between neighbor nodes so that the nodes can identify their relative positions to each other. Only those mobile nodes that stay farther than the query forwarder from the forwarder's BS forward the queries in the direction of the destination BS. In this way, the query can be forwarded to the destination BS faster. After the multi-hop path is discovered, the neighboring BS sends the segment to the destination node along the path. Since the destination node is in the neighboring BS's region, the overhead to identify a path to the destination node is small. Note that our methods for congestion control in base stations involve query broadcasting. However, it is used only when some base stations are overloaded rather than in the normal DTR routing

algorithm in order to avoid load congestion in BSes.

Figure 4 shows an example of the congestion control on BSes when $b = 2$. As shown in figure, BS B_1 is congested. Then, the relay nodes of the source node's message broadcast locally by beacon piggybacking to find multi-hop paths which lead to B_3 and B_4. The relay nodes then send segments along the paths. In this way, the traffic originally targeting overloaded B_1 can be spread out to the neighboring BSes B_3 and B_4. B_3 and B_4 further forward the segments to the destination BS B_6 if B_6 is not congested. If B_6 is also congested, B_3 and B_4 send the segments to the neighboring BSes of B_6. Specifically, B_4 sends the segment to B_3. B_3 does not forward the segment to another BS since it already is close to B_6. B_3 then finds a multi-hop path to the destination node and uses ad-hoc transmission to forward the segments to the destination node. Similarly, when B_2 wants to send a segment to the destination node, it also uses a multi-hop path for the segment transmission.

4 Performance Analysis of the DTR Protocol

σ	Node density	M	Number of BSes
l	Segment's length	s_h	Area size of a cell
$n(S)$	Number of nodes in area S	R	Transmission range
W_i	Bandwidth of a node v_i	m_i	Mobile node i
$P(\sigma, M)$	Throughput	$n(\sigma, M)$	Number of nodes

TABLE 1: Parameter table.

In this section, we analyze the effectiveness of the DTR protocol at enhancing the capacity and scalability of hybrid wireless networks. In our analysis, we use the same scenario in [17] for hybrid wireless networks, and use the same scenario in [33] for the ad-hoc network component. We present the scenarios and some concepts below. We consider a large number of mobile nodes uniformly and randomly deployed over a 2-D field. The moving directions of the nodes are independent and identically distributed (i.i.d.). The distribution of mobile nodes can be modeled as a homogeneous Poisson process with node density σ [34]. That is, given an area of size S in the field, the number of nodes in the area, denoted by $n(S)$, follows the Poisson distribution with the parameter σS,

$$\Pr\left(n(S) = k\right) = \frac{(\sigma S)^k e^{-\sigma S}}{k!},\ k = 0, 1, 2, ... \qquad (1)$$

Besides mobile nodes, there are M BSes regularly deployed in the field. The BSes divide the area into a hexagon tessellation, in which each hexagon has side length h. The BSes are assumed to be connected together by a wired network. We assume that the link bandwidths in the wired network are large enough so that there are no bandwidth constraints between BSes. In *single-path transmission*, a message is sequentially transmitted in one routing path. In *multi-path transmission*, a message is divided into a number of segments that are forwarded along multiple paths in a distributed manner. We assume each segment has the same length l. Table 1 lists the notations used in our analysis.

We assume that the transmission range of all mobile nodes and all BSes is R ($R > h$). In this paper, we use *protocol model* [17, 33] to describe the interference among nodes; that is, a transmission from a node (here "node" can be either mobile node or BS) v_i to another node v_j is successful if the following two conditions are satisfied:
1) v_j is within the transmission range of v_i, i.e.,

$$|v_i - v_j| \le R \qquad (2)$$

where $|v_i - v_j|$ represents the Euclidean distance between v_i and v_j in the plane.
2) For any other node v_k that is simultaneously transmitting over the same channel,

$$|v_k - v_j| \ge (1 + \Delta)|v_i - v_j|. \qquad (3)$$

Formula (3) guarantees a guard zone around the receiving node to prevent a neighboring node from transmitting on the same channel at the same time. The radius of the guard zone is $(1 + \Delta)$ times the distance between the sender and the receiver. The parameter Δ defines the size of the guard zone and we require that $\Delta > 0$.

We first adopt a concept called *aggregate throughput capacity* introduced in [17, 33] to measure the throughput of the network.

Definition (Aggregate Throughput Capacity of Hybrid Networks) The aggregate throughput capacity of a hybrid wireless network is of order $\Theta(f(\sigma, M))$ if there are deterministic constants $\alpha > 0$, and $\alpha' < +\infty$ such that

$$\lim_{M\to\infty} \Pr\left(P(\sigma, M) = \alpha f(\sigma, M) \text{ is feasible}\right) = 1 \qquad (4)$$

$$\liminf_{M\to\infty} \Pr\left(P(\sigma, M) = \alpha' f(\sigma, M) \text{ is feasible}\right) < 1. \qquad (5)$$

Since the working frequency of infrastructure networks is around 700MHz while that of ad-hoc networks is 2.4 GHz, the communications in infrastructure mode (between mobile nodes and BSes through cellular interface) would not generate interference to ad-hoc mode. We divide the channel for infrastructure mode transmissions into uplink and downlink parts, according to the transmission direction relative to the BSes. Accordingly, in the DTR protocol, the traffic of each S-D pair is composed of at most two intra-cell traffics, one uplink traffic and one download traffic. Since uplink traffic and downlink traffic use different sub-channels, there is also no interference between these two types of traffics. For each node v_i, we denote the bandwidth assigned to intra-cell, uplink, and downlink sub-channels by W_i^{int}, W_i^{up} and W_i^{down}, respectively. We let $W_i^{\text{up}} = W_i^{\text{down}}$ because there are the same amount of uplink traffic and downlink traffic. The transmission rates should sum to W_i, i.e., $W_i^{\text{int}} + W_i^{\text{up}} + W_i^{\text{down}} = W_i$. Though no interference exists between intra-cell, uplink, and downlink traffics, interference exists between the same type of traffic in a cell and between different cells. Fortunately, there is an efficient spatial transmission schedule that can prevent such interferences [17]. First, to avoid the interference in a cell, any two nodes within the cell are not allowed to transmit with the same traffic mode at the same time. Second, to avoid the interference between different cells, the cells are spatially divided into a number of groups and transmissions in the cells of the same group do not interfere with each other. If the groups are scheduled to transmit in a round robin fashion, each cell will be able to transmit once every fixed amount of time without interfering with each other.

Below, we show how many groups we need to divide the cells to prevent interference. We adopt the notion of *interfering neighbors* introduced in [17], and give the number of cells that can be affected by a transmission in one cell. Two cells are defined to be interfering neighbors if there is a point in one cell which is within a distance $(2+\Delta)R$ of a point in the other cell. Accordingly, if two cells are not interfering neighbors, transmissions in one cell do not interfere with transmissions in the other cell. [17] has proved that (1) each cell has no more than c_1 interfering neighbors (Lemma 1 in [17]), where c_1 is a constant

$$c_1 = \frac{4}{3}\left(\frac{3l + 2R + \Delta R}{3l}\right)^2, \qquad (6)$$

and (2) all cells should be divided into c_1+1 groups and the whole channel should be divided into c_1+1 subchannels, where each subchannel is allocated to the cells in one group. Thus, the number of group we need to divide the cells to prevent interference is c_1+1.

Before calculating the aggregate throughput capacity of DTR, we first introduce Lemma 4.1.

Lemma 4.1: The number of cells that have mobile nodes is $\Theta(M)$.

Proof: Denote the number of cells having mobile nodes by M_1. To prove $M_1=\Theta(M)$, we need to prove that there exists deterministic constants $\alpha>0$ and $\alpha'<+\infty$ such that

$$\lim_{M\to\infty}\Pr\left(M_1=\alpha M\right)=1, \tag{7}$$

$$\liminf_{M\to\infty}\Pr\left(M_1=\alpha' M\right)<1. \tag{8}$$

For Formula (8), let $\alpha'=2$. Because the number of cells having mobile nodes is upper bounded by M, then

$$\liminf_{M\to\infty}\Pr\left(M_1=2M \text{ is feasible}\right)=0. \tag{9}$$

Now, we prove that Formula (7) can also be satisfied for some constant α. Because the number of nodes in a cell follows a Poisson distribution and the size of each cell (hexagon) is $s_h=3\sqrt{3}h^2$, then we can derive the probability that no mobile node is in a cell equals

$$\Pr\left(n(S_h)=0\right)=\frac{\sigma^0 e^{-s_h}}{0!}=e^{-s_h}. \tag{10}$$

Consider an arbitrary cell k, let X_1, X_2, ..., X_k,..., X_M be i.i.d. random variables, where X_k represents whether cell k has mobile nodes. Then, X_k is defined as follows:

$$X_k=\begin{cases}1 & \text{cell } k \text{ has mobile nodes}\\ 0 & \text{cell } k \text{ does not have mobile nodes}\end{cases} \tag{11}$$

and $E(X_k)=e^{-s_h}$. For simplicity, let $c_2=1-e^{-s_h}$. Then, $M_1=\sum_{k=1}^{M}X_k$. By the *Strong Law of Large Number (SLLN)* [34],

$$\Pr\left(\lim_{M\to\infty}\frac{\sum_{k=1}^{M}X_k}{M}=c_2\right)=1, \tag{12}$$

which implies that $\lim_{M\to\infty}\Pr\left(M_1=c_2M\right)=1$, which indicates that when $\alpha=c_2$, Formula (7) can also be satisfied. □

Lemma 4.2: Let $n(\sigma,M)$ denote the number of mobile nodes in the whole network. Then,

$$\lim_{M\to\infty}\Pr\left(n(\sigma,M)=s_hM\right)=1. \tag{13}$$

Proof: Let Z_1, Z_2, ..., Z_M be i.i.d. random variables representing the number of nodes in cell 1, 2,..., M, respectively. Then, $n(\sigma,M)=\sum_{k=1}^{M}Z_k$. Because each Z_k follows a Poisson distribution with parameter s_h, $E(Z_k)=s_h$, $\forall 1\le k\le M$. According to SLLN,

$$\Pr\left(\lim_{M\to\infty}\frac{\sum_{k=1}^{M}Z_k}{M}=s_h\right)=1, \tag{14}$$

which implies that $\lim_{M\to\infty}\Pr\left(\sum_{k=1}^{M}Z_k=s_hM\right)=1$, and hence $\lim_{M\to\infty}\Pr\left(n(\sigma,M)=s_hM\right)=1$. □

Theorem 4.1: For a hybrid network of M BSes and σ mobile node density, where each node has the intra-cell, uplink and downlink sub-channel bandwidth satisfying

$$W_i^{\text{down}}=W_i^{\text{up}}=W^{\text{up}}=W/4,\ W_i^{\text{int}}=W^{\text{int}}=W/2 \tag{15}$$

the aggregate throughput capacity of DTR is

$$P(\sigma,M)=\Theta(MW). \tag{16}$$

Proof: To prove $P(\sigma,M)=\Theta(MW)$, we need to prove that there exists deterministic constants $\alpha>0$ and $\alpha'<\infty$ such that

$$\lim_{M\to\infty}\Pr\{P(\sigma,M)=\alpha MW \text{ is feasible}\}=1 \tag{17}$$

$$\liminf_{M\to\infty}\Pr\{P(\sigma,M)=\alpha' MW \text{ is feasible}\}<1. \tag{18}$$

Recall that any two nodes within a cell cannot transmit simultaneously in the same traffic mode, the throughput P is upper bounded by $MW/4$, which can be achieved only if each cell has one node to send the message. Hence, Formula (18) can be satisfied by setting α' to $1/2$.

Then, we will show how Formula (17) can be satisfied. Since the same message has to go through an uplink and a downlink and it is counted only once in the throughput capacity, calculating the throughput of the whole network is equivalent to calculating the throughput of uplink traffic P_{up} or the throughput of downlink traffic P_{down}. Notice calculating intra-cell traffic throughput is not accurate because a message may transmit twice with intra-cell mode. In this proof, we calculate P_{up}.

First, we consider the throughput of the uplink traffic of an arbitrary cell k, denoted by P_{up}^k. Since the schedule allocates $1/(c_1+1)$ time slots to this cell, then

$$P_{\text{up}}^k=\frac{W^{\text{up}}}{c_1+1}. \tag{19}$$

Then, we consider the throughput of the whole network. Let $P_{\text{up}}=\sum_{i=1}^{M}P_{\text{up}}^iX_i$ represent the throughput of uplink traffic, then we have

$$\begin{aligned}&\lim_{M\to\infty}\Pr\left(P_{\text{up}}=\frac{c_2MW}{3(c_1+1)}\right)\\ =&\lim_{M\to\infty}\Pr\left(\sum_{i=1}^{M}P_{\text{up}}^iX_i=\frac{c_2MW^{\text{up}}}{c_1+1}\right)\\ =&\lim_{M\to\infty}\Pr\left(\sum_{i=1}^{M}X_i=c_2M\right)=1 \quad \text{(By Lemma 4.1)}\end{aligned}$$

Accordingly, Formula (17) can be satisfied when α is set to $\frac{c_2}{3(c_1+1)}$. □

Corollary 4.1: With the restriction in Theorem 4.1, DTR can achieve $\Theta(W)$ throughput per S-D pair.

Proof: Denote the throughput of per S-D pair by $\overline{P}$, which equals

$$\overline{P}=\frac{P(\sigma,M)}{n}. \tag{20}$$

Obviously, $\overline{P}$ is upper bounded by $\frac{W}{4}$ because each node has at most $\frac{W}{4}$ for uplink traffic (or downlink traffic), which equals its S-D pair throughput. By Lemma 4.2 and Theorem 4.1, we can derive that

$$\begin{aligned}&\lim_{M\to\infty}\Pr\left(\overline{P}=\frac{c_2W}{3(c_1+1)s_h}\right)\\ =&\lim_{M\to\infty}\Pr\left(\frac{P(\sigma,M)}{n(\sigma,M)}=\frac{c_2W}{3(c_1+1)s_h}\right)\\ \ge&\lim_{M\to\infty}\Pr\left(P(\sigma,M)=\frac{c_2WM}{3(c_1+1)}\right)\Pr\left(n(\sigma,M)=s_hM\right)\\ =&\lim_{M\to\infty}\Pr\left(P(\sigma,M)=\frac{c_2WM}{3(c_1+1)}\right)=1\end{aligned}$$

which implies that $\lim_{M\to\infty}\Pr\left(\overline{P}=\frac{c_2W}{3(c_1+1)s_h}\right)=1$. □

Corollary 4.1 shows that DTR produces a constant throughput for each pair of nodes regardless of the number of nodes in each cell due to its spacial reuse of the system. Theorem 4.1 and Corollary 4.1 show that the aggregate throughput capacity and the throughput per S-D pair of DTR are $\Theta(MW)$ and $\Theta(W)$, respectively. The work in [17] proves that DHybrid achieves $\Theta(MW)$ infrastructure aggregate throughput, and the work in [33] proves that the pure ad-hoc transmission achieves $\Theta(\frac{W}{\sqrt{n\cdot logn}})$ throughput per S-D pair. The results demonstrate that the throughput rates of DTR and DHybrid are higher than that of the pure ad-hoc transmission. This is because the pure ad-hoc transmission is not efficient in a large scale network [35]. A large network size reduces the path utilization efficiency and increases node interference. Facilitated by the infrastructure network,

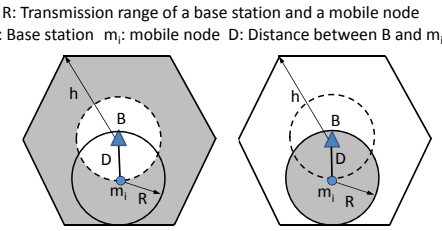


Fig. 5: The traffic load in DHybrid and DTR.

DTR and DHybrid avoid long distance transmissions, leading to a higher transmission throughput.

Proposition 4.1: Suppose a mobile node needs to allocate totally U segments with the same length to L neighboring mobile nodes m_1, ..., m_L, which has uplink bandwidth W_1^{up}, ..., W_L^{up}, respectively. Let U_i denote the number of segments to be allocate to m_i $(i = 1, 2, ..., L)$. To minimize the average latency of these segments, the optimal allocation should satisfy $\frac{U_1}{W_1^{\text{up}}} = ... = \frac{U_L}{W_L^{\text{up}}}$. The minimized average latency equals $\frac{Ul}{2\sum_{i=1}^{L} W_i^{\text{up}}}$.

Proof: Recall that each segment has length l. Then, for each mobile node m_i it requires $\frac{l}{W_i^{\text{up}}}$ time to transmit a segment. Therefore, the j^{th} segment that m_i needs to transmit has to wait $\frac{(j-1)l}{W_1^{\text{up}}}$ slots. Hence, the total latency of the segments that m_i needs to transmit to its BS equals

$$\sum_{j=1}^{U_i} \frac{(j-1)l}{W_i^{\text{up}}} = \frac{(0+1+...+(U_i-1))l}{W_i^{\text{up}}} \approx \frac{U_i^2 l}{2W_i^{\text{up}}}. \quad (21)$$

Hence, the average latency of transmitting all the messages should be $\sum_{i=1}^{K} \frac{U_i^2 l}{2W_i^{\text{up}}}/U$. According to Cauchy-Schwarz inequality [34], the average latency is lower bounded

$$\begin{aligned} \frac{1}{U}\sum_{i=1}^{L} \frac{U_i^2 l}{2W_i^{\text{up}}} &= \frac{l}{2U\sum_{i=1}^{L} W_i^{\text{up}}} \sum_{i=1}^{L} \frac{U_i^2}{W_i^{\text{up}}} \sum_{i=1}^{L} W_i^{\text{up}} \\ &\geq \frac{l}{2U\sum_{i=1}^{L} W_i^{\text{up}}} \left(\sum_{i=1}^{L} \sqrt{\frac{U_i^2}{W_i^{\text{up}}}}\sqrt{W_i^{\text{up}}}\right)^2 \\ &= \frac{Ul}{2\sum_{i=1}^{L} W_i^{\text{up}}}. \end{aligned} \quad (22)$$

When $\frac{\sqrt{\frac{U_1^2}{W_1^{\text{up}}}}}{\sqrt{W_1^{\text{up}}}} = ... = \frac{\sqrt{\frac{U_L^2}{W_L^{\text{up}}}}}{\sqrt{W_L^{\text{up}}}}$, or equivalently, $\frac{U_1}{W_1^{\text{up}}} = ... = \frac{U_L}{W_L^{\text{up}}}$, the average segment latency $\sum_{i=1}^{L} \frac{U_i^2 l}{2W_i^{\text{up}}}/U$ can achieve the minimum value $\frac{Ul}{2\sum_{i=1}^{L} W_i^{\text{up}}}$. □

Proposition 4.1 indicates that forwarding segments to the nearby nodes with the highest capacity can minimize the average latency of messages in the cell. It also balances the transmission load of the mobile nodes within a cell.

Proposition 4.2: A source node in DTR can find relay nodes for message forwarding with probability $\sum_{k=1}^{\infty} \frac{k-1}{k} \frac{c_r^k e^{-c_r}}{k!}$, where $c_r = \pi R^2$.

Proof: Let m denote the number of nodes within m_i's transmission area and define the indicator variable Q_i by

$$Q_i = \begin{cases} 1 & m_i \text{ is the highest capacity node} \\ 0 & m_i \text{ is not the highest capacity node} \end{cases} \quad (23)$$

then,

$$\begin{aligned} &\Pr\{m_i \text{ can find relays for message forwarding}\} \\ &= \sum_{k=0}^{\infty} \Pr(Q_i = 0 | m = k) \Pr(m = k) = \sum_{k=1}^{\infty} \frac{k-1}{k} \frac{c_r^k e^{-c_r}}{k!} \end{aligned}$$ □

Proposition 4.2 indicates that in a high-density network, a source node in DTR can find relay nodes for message forwarding with a high probability. For example, assume the average number of neighbor nodes of a source node is 10. With the daily increasing number of mobile devices, such an assumption is realistic. Then, the probability of not being able to find any node in the range of a node is $1 - \sum_{k=1}^{\infty} \frac{k-1}{k} \frac{10^k e^{-10}}{k!} \approx 0.12$, which is very small. Therefore, in a high-density network, a source node can find neighbors for message forwarding with a high probability.

We use DHybrid to denote the group of routing protocols in hybrid wireless networks that directly combine the ad-hoc transmission mode and the infrastructure transmission mode [1, 5, 6, 12–18].

Proposition 4.3: In a hybrid wireless network, the D-Hybrid routing protocol leads to load imbalance among the mobile nodes in a cell.

Proof: Figure 5 (a) shows a cell with a BS and a randomly picked mobile node m_i in the range of the BS. The shaded region represents all possible positions of the source nodes that choose m_i as the relay node in DHybrid. The total traffic passing through node m_i is the sum of the traffic generated by the nodes in the shaded region. The area of shaded region is

$$S = s_h - \pi D^2 \quad (0 < D < h), \quad (24)$$

where D is the distance between the BS and relay node m_i and s_h is the area size of a cell. Therefore, the expected value of traffic passing through node m_i is

$$W \cdot \sigma \cdot (s_h - \pi D^2) \quad (0 < D < h), \quad (25)$$

where W is the data transmission rate of a source node, and σ is the density of the nodes in a region. Equation (25) shows that the traffic passing through node m_i decreases as D increases. That is, the nodes closer to the BS have a higher load than the nodes staying at the brim of the cell. □

Proposition 4.4: In a hybrid wireless network, DTR achieves more balanced load distribution among the mobile nodes in each cell.

Proof: The shaded region in Figure 5 (b) represents all possible positions of the source and relay nodes that choose node m_i as relay node. Suppose m neighbor nodes are chosen as relay nodes, then the expected traffic passing through node m_i is $\frac{W}{m} \cdot \sigma \cdot \pi R^2$ which shows that the traffic going through node m_i is independent of its location relative to its BS. Since every node in the cell has an equal probability of generating traffic, the traffic load is balanced among the nodes in the cell. □

5 Performance Evaluation

This section demonstrates the properties of DTR through simulations on NS-2 [36] in comparison to DHybrid [17], Two-hop [19] and AODV [8]. In DHybrid, a node first uses broadcasting to observe a multi-hop path to its own BS and then forwards a message in the ad-hoc transmission mode along the path. During the routing process, if the transmission rate (i.e., bandwidth) of the next hop to the BS is lower than a threshold, rather than forwarding the message to the neighbor, the node forwards the message directly to its BS. The source node will be notified if an established path is broken during data transmission. If a source sends a message to the same destination next time, it uses the previously established path if it is not broken. In the Two-hop protocol, a source node selects the better transmission mode between direct transmission and relay transmission. If the source node can find a neighbor that has higher bandwidth to the

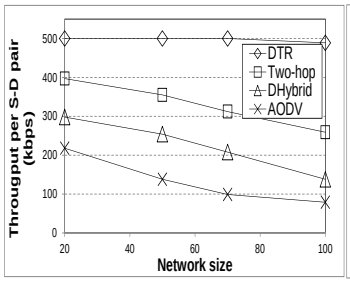


Fig. 6: Throughput vs. network size (simulation).

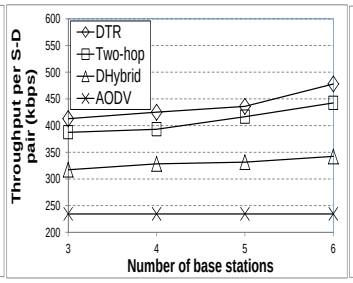


Fig. 7: Throughput vs. number of BSes.

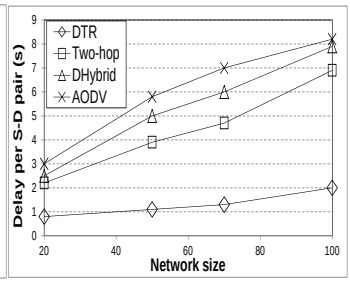


Fig. 8: Delay vs. network size.

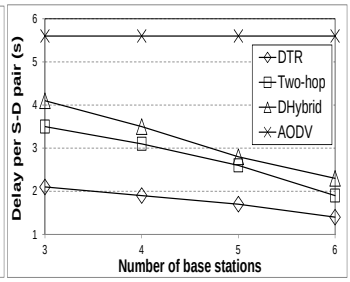


Fig. 9: Delay vs. number of BSes.

BS than itself, it transmits the message to the neighbor. Otherwise, it directly transmits the message to the BS.

Unless otherwise specified, the simulated network consists of 50 mobile nodes and 4 BSes. In the ad-hoc component of the hybrid wireless network, mobile nodes are randomly deployed around the BSes in a field of 1000×1000 square meters. We used the Distributed Coordination Function (DCF) of the IEEE 802.11 as the MAC layer protocol. The transmission range of the cellular interface was set to 250 meters, and the raw physical link bandwidth was set to 2Mbits/s. The transmission power of the ad-hoc interface was set to the minimum value required to keep the network connected for most times, even when nodes are in motion in the network. Then, the influence of the transmission range on different methods' performance is controlled. Specifically, we set the transmission range through the ad-hoc interface to 1.5 times of the average distance between neighboring nodes, which can be obtained by measuring the simulated network. We used the two-ray propagation model for the physical layer model. Constant bit rate (CBR) was selected as the traffic mode in the experiment with a rate of 640kbps. In the experiment, we randomly chose 4 source nodes to continuously send messages to randomly chosen destination nodes. The number of channels for each BS was set to 10. We set the number of redundant routing paths b in Section 3.4 to 1. We assumed that there was no capacity degradation during transmission between BSes. This assumption is realistic considering the advanced technologies and hardware presently used in wired infrastructure networks. There was no message retransmission for failed transmissions in the experiments.

We employed the random way-point mobility model [37] to generate the moving direction, speed, and pause duration of each node. In this model, each node moves to a random position with a speed randomly chosen from $(1-20)$m/s. The pause time of each node was set to 0. We set the number of segments of a message to the connection degree of the source node. The simulation warmup time was set to 100s and the simulation time was set to 1000s. We conducted the experiments 5 times and used the average value as the final experimental result. To make the methods comparable, we did not use the congestion control algorithm in DTR unless otherwise indicated.

5.1 Scalability

Figure 6 shows the average throughput measured in kbps per S-D pair of different routing protocols versus the number of mobile nodes in the system. The figure shows the throughput of DTR remains almost the same with different network sizes. This result conforms to Corollary 4.1. DTR uses distributed multi-path routing to fully take advantage of the spatial reuse and avoid transmission congestion in a single path. Unlike the multi-hop routing in mobile ad-hoc networks, DTR does not need path query and maintenance. Also, it limits the path length to three to avoid problems in long-path transmission. The throughput of DHybrid and AODV decreases as the number of nodes in the network increases. This is mainly because when the network size increases, more beacon messages are generated in the network. Also, the long transmission path also leads to high transmission interference. Then, nodes in these methods suffer from intense interference, leading to more transmission failure and degraded overall throughput. Also, the mobile node increase in the system leads to high network dynamism, resulting in frequent route re-establishments.

The short routing paths in Two-hop reduce congestion and signal interference, thus enabling better spatial reuse as in DTR. Meanwhile, Two-hop enables nodes to adaptively switch between direct transmission and relay transmission. Hence, part of the transmission load is transferred to relay nodes, which carry the messages until meeting the BSes. As a result, gateway nodes connecting mobile nodes and BSes are not easily overloaded. Therefore, the throughput of Two-hop is higher than DHybrid. However, since the number of message routing hops is confined to one, Two-hop may not find the node with the best transmission rate to the BSes because of the short transmission range of the ad-hoc interface. Therefore, the throughput of Two-hop is lower than DTR, especially in a network with high node density. The reason that AODV has the lowest throughput per S-D pair is its long transmission paths.

Figure 7 shows the throughput per S-D pair versus the number of BSes in different routing protocols. The number of BSes was varied from 3 to 6. The BSes are uniformly distributed in the network. We can see from the figure that as the number of BSes increases, the throughputs of DTR, Two-hop, and DHybrid increase while the throughput of AODV stays nearly constant. In DTR, Two-hop, and DHybrid, as the number of BSes increases, the total number of nodes close to the BSes increases. Then, more nodes have high transmission rates to the BSes, leading to a throughput increase. In AODV, since the traffic between S-D pairs does not travel though BSes, the throughput between an S-D pair is not affected by the increased number of BSes in the network. The figure also shows that the throughput of DTR is constantly larger than Two-hop and the throughput of Two-hop is constantly larger than DHybrid. AODV constantly has the lowest transmission delay due to the same reasons as in Figure 6.

5.2 Transmission Delay

Figure 8 shows the average transmission delay of S-D pairs for successfully delivered messages in different routing protocols versus network size. The network size was varied from 20 to 100 with 20 increase in each step.

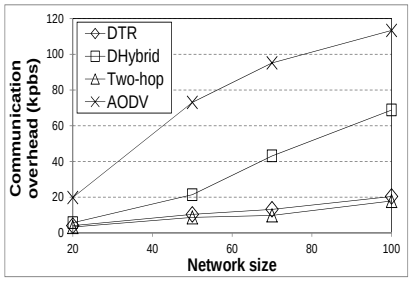


Fig. 10: Overhead vs. network size

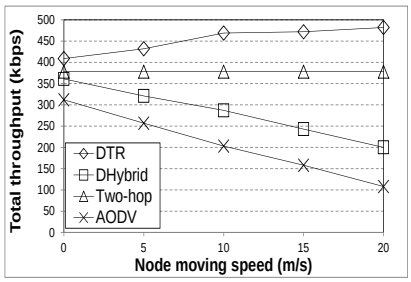


Fig. 11: Throughput vs. mobility.

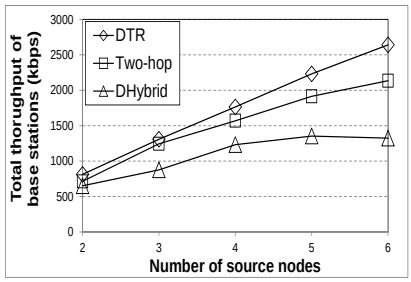


Fig. 12: Throughput of BSes vs. number of source nodes.

Transmission delay is the amount of time it takes for a message to be transmitted from its source node to its destination node. From the figure, we see that DTR generates the smallest delay. In DTR, each source node first divides its messages into smaller segments and then forwards them to the nearby nodes with the highest capacity, which leads to more balanced transmission load distribution among nodes than the previous methods. According to Proposition 4.1, average latency can be minimized when the transmission loads of all the nodes are balanced. Hence, DTR has smaller latency than the previous methods. The delay of DHybrid is 5-6 times larger than DTR. DHybrid uses a single transmission path, while DTR uses multiple paths. Recall that we set the number of segments of a message to the connection degree of the source node in DTR. Thus, the ratio of delay time of DHybrid to that of DTR equals the average connection degree. As the number of nodes in the system increases, the connection degree of each node increases, and the increase rate of the ratio grows. This is caused by two reasons. First, a higher node density leads to longer path lengths in DHybrid, resulting in a longer delay because of a higher likelihood of link breaks. Second, a higher node density enables a node to quickly find relay nodes to forward messages in DTR, as indicated in Proposition 4.2.

DTR also produces a shorter transmission delay than Two-hop for two reasons. First, the multi-path parallel routing of DTR saves much transmission time as shown in Proposition 4.1. Second, the distributed routing of DTR enables some messages to be forwarded to the destination BS's neighboring cells with high transmission rates rather than waiting in the current hot cell for a transmission channel. We can also observe that Two-hop produces lower delay than DHybrid. This is because the delay of DHybird includes the time for establishing a path and for data transmission. Also, the multi-hop transmission component of DHybrid results in a higher delay due to the queuing delay in each hop. Because of the long distance transmissions without support from an infrastructure network, AODV generates the longest delay.

Figure 9 plots the average communication delay per S-D pair for successfully delivered messages versus the number of BSes in different routing protocols. The figure shows that the increasing number of BSes in the system leads to a communication delay decrease between nodes in DTR, Two-hop, and DHybrid, but does not affect the communication delay in AODV. In DTR, Two-hop, and DHybrid, as the number of BSes increases, more nodes can stay close to the BSes, leading to fewer communication hops and better transmission links between nodes and BSes. Thus, the transmission delay between the nodes is reduced. Since the communication between S-D pairs in AODV does not rely on BSes, AODV maintains constant communication delay. The figure also shows that the communication delay between S-D pairs follows DTR<Two-hop<DHybrid<AODV for the same reason as in Figure 8.

5.3 Communication Overhead

We use the generation rate of control messages in the network and MAC layers in kbps to represent the communication overhead of the routing protocols. Figure 10 illustrates the communication overhead of DTR, Two-hop, DHybrid, and AODV versus network size. We can see that the communication overheads of DTR and Two-hop are very close. This is because both DTR and Two-hop are transmission protocols of short distance and small hops. DTR has slightly higher communication overhead than Two-hop because DTR utilizes three hop transmission, which has one more hop than two hop transmission. However, the marginal overhead increase leads to a much higher transmission throughput as shown in Figure 6. DHybrid generates much higher overhead than DTR and Two-hop because of the high overhead of routing path querying. The pure AODV routing protocol results in much more overhead than the others. This is because without an infrastructure network, the messages in AODV travel a long way from the source node to the destination node through much longer paths.

5.4 Effect of Mobility

In order to see how the node mobility influences the performance of the routing protocols, we evaluated the throughput of these four transmission protocols with different node mobilities. Figure 11 plots the throughput of DTR, DHybrid, Two-hop, and AODV versus node moving speed. From the figure, we can see that the increasing mobility of the nodes does not adversely affect the performance of DTR and Two-hop. It is intriguing to find that high mobility can even help DTR to increase its throughput and that Two-hop generates constant throughput regardless of the mobility. This is because the DTR and Two-hop transmission modes do not need to query and rely on multi-hop paths; thus, they are not affected by the network partition and topology changes. Moreover, since DTR transmits segments of a message in a distributed manner, as the mobility increases, a mobile node can meet more nodes in a shorter time period. Therefore, DTR enables the segments to be quickly sent to high-capacity nodes. As node mobility increases, the throughput of DHybrid decreases. In DHybrid, the messages are routed in a multi-hop fashion. When the links between nodes are broken because of node mobility, the messages are dropped. Therefore, when nodes have smaller mobility, the links between the mobile nodes last longer and more messages can be transmitted. Hence, the throughput of DHybrid is adversely affected by node mobility. However, since DHybrid can adaptively adjust

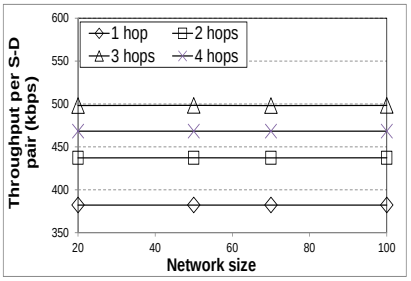


Fig. 13: Throughput vs. number of hops.

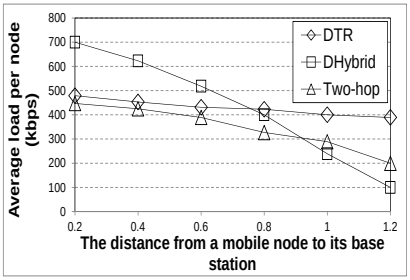


Fig. 14: Load distribution in a cell.

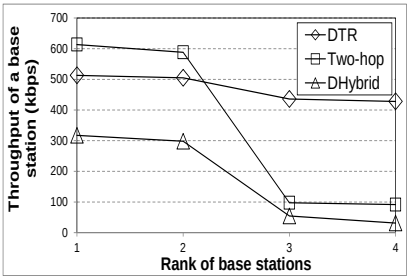


Fig. 15: Load distribution among BSes.

the routing between the ad-hoc transmission and cellular transmission, the throughput of DHybrid is much higher than AODV's. With no infrastructure network, AODV produces much lower throughput than the others. Its throughput also drops as node mobility increases for the same reasons as DHybrid.

5.5 Effect of Workload

We measured the total throughput of BSes on the messages received by BSes. Figure 12 shows the total throughput of the BSes versus the number of source nodes. We can see that DTR and Two-hop have much higher throughput increase rates than DHybrid. This is because in DTR and Two-hop, the number of transmission hops from a source node to a BS is small. Meanwhile, each node can adaptively switch between relay transmission and direct transmission based on the transmission rate of its neighbors. Hence, part of a source node's transmission load is transferred to a few relay nodes, which carry the messages until meeting the BSes. Therefore, the gateway mobile nodes are less likely to be congested. However, nodes in DHybrid cannot adaptively adjust the next forwarding hop because it is predetermined in the routing path. Messages are always forwarded to the mobile gateway nodes that are closer to the BSes or that have higher transmission rates. Therefore, these mobile gateway nodes can easily become congested as the workload of the system increases, leading to many message drops. Therefore, when the number of the source nodes is larger than 4, the throughput of DHybrid remains nearly constant. This is also the reason that the throughput of DHybrid is constantly lower than those of DTR and Two-hop. Additionally, the figure shows that the overall throughput of Two-hop is lower than that of DTR. This is because most of the traffic in Two-hop is confined to a single cell. When a BS in a cell is congested, the traffic cannot be transferred to other cells. In contrast, DTR's three-hop distributed forwarding mechanism enables it to distribute the traffic among the BSes in a balance. Therefore, the BSes in DTR will not become congested easily. In addition, as the forwarding mechanism gives nodes more flexibility in choosing relay nodes with higher transmission rates for message forwarding to the BSes, the overall BS throughput in DTR is larger than in Two-hop.

5.6 Effect of the Number of Routing Hops

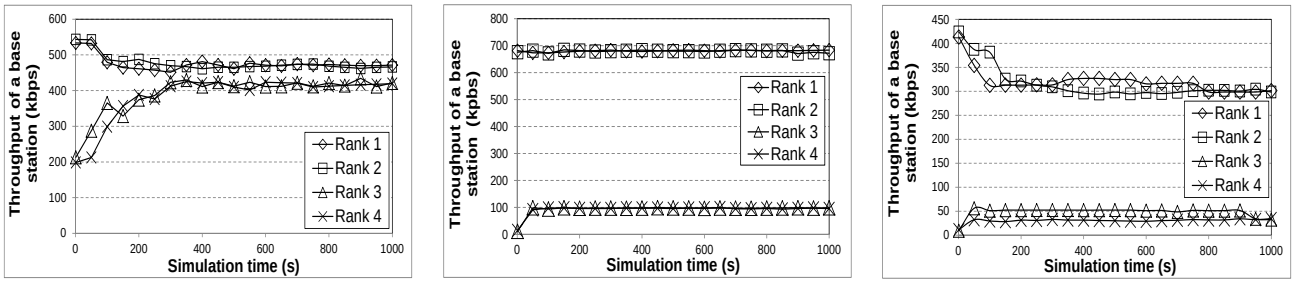
We conducted experiments to show the optimal number of routing hops for the routing in hybrid wireless networks. We tested the throughput per S-D pair for x-hop DTR, where x was varied from 1 to 4. In the 1-hop routing, a node directly transmits a message to the BS without message division. In the other routing protocols, the $(x-1)^{th}$ hop chooses the best transmission mode between direct transmission and relay transmission. Also, in the 4-hop routing, the second relay node randomly chooses the third relay node.

Figure 13 shows the average throughput per S-D pair versus network size in DTR. As the figure shows, as the network size increases, the node throughput keeps constant regardless of the number of forwarding hops in a routing. The reason is the same as in Figure 6. We can also see from the figure that the throughput of the four protocols follows 3-hop>4-hop>2-hop>1-hop. In the 1-hop routing, each node only transmits segments directly to a BS regardless of its current transmission rate. In the 2-hop routing, if the transmission rate of a node's neighbor is higher than that of the node, it asks its neighbor node to forward the segment to a BS. Therefore, the 2-hop routing has higher throughput than the 1-hop routing. The 3-hop routing can greatly increase the number of node options for segment routing since the number of nodes that the source node can encounter increases from d to d^2, where d is the average node degree. Thus, a node with a greater transmission rate can be chosen as the forwarding node. Meanwhile, the 3-hop routing can greatly facilitate inter-cell communication because a node has a higher probability of reaching a neighboring BS within a 3-hop path length than within a 2-hop path length. Therefore, the throughput of the 3-hop routing is much higher than that of the 2-hop routing. The figure also shows that the 4-hop routing produces lower throughput than the 3-hop routing. The reason is that 3 hops are enough to find a hop with high transmission rate and achieve inter-cell communication because of widespread BSes. The 4-hop routing increases the forwarding delay due to the greater number of nodes in a route; thus, it cannot increase the uploading transmission rate of messages.

5.7 Load Distribution Within a Cell

In this experiment, we tested the load distribution of mobile nodes in a randomly chosen cell in the hybrid wireless network that employs each of the DTR, DHybrid, and Two-hop protocols. We normalized the distance from a mobile node to its base station according to the function $\frac{D}{R_b}$, where D is the actual distance and R_b is the radius of its cell. We divided the space of the cell into several concentric circles and measured the loads of the nodes on each circle to show the load distribution.

Figure 14 shows the average load of a node corresponding to the normalized distance from itself to the BS in the chosen cell. The figure shows that most of the traffic load of DHybrid is located at nodes near the BS. The nodes far from the BS have very low load. The results conform to Proposition 4.3. In DHybrid, if a source node wants to access the Internet backbone or engage in inter-cell communication, it transmits the messages to the BSes in a multi-hop fashion. Therefore, the nodes near the BSes will have the highest load. On the other hand, since there is little traffic going through the nodes at the brim of a cell, the load of these nodes is small. As a result, some nodes can easily become hot spots while the resources of other nodes are not



(a) DTR (b) Two-hop (c) DHybrid

Fig. 16: Base station load vs. simulation time.

fully utilized. This load imbalance prevents DHybrid from fully utilizing system resources. The traffic load of DTR is almost evenly distributed in the system, which is in line with Proposition 4.4. In DTR, the traffic from a source node is distributed among a number of relay neighbors for further data forwarding. The nodes at the brim of the cell also take responsibility for message forwarding, since the neighbor nodes of the brim nodes could be located in other cells with good transmission channels. In Two-hop, the source node considers direct transmission or one-hop relay transmission based on the channel condition. Since the node is chosen within one hop, the messages will not gather close to the BS due to the limited transmission range. However, because of its sequential transmission, Two-hop cannot achieve load balance among nodes in a cell as well as DTR.

5.8 Load Balance Between Cells

In this experiment, we tested the effectiveness of the congestion control algorithm in DTR. We also added a congestion control algorithm to DHybrid. In the algorithm, when a node receives beacon messages from its BS indicating that it is overloaded, the node broadcasts a query message to find a path to a nearby uncongested BS. We selected two BSes out of the total four BSes. In the range of each of the two selected BSes, we randomly selected one mobile node as the source node to send messages to a randomly selected destination node in the network. Once the source node moves out of the range of the selected BS, another mobile node in the range was selected as the source node. In order to show the load distribution of the BSes in different protocols, we ranked the BSes based on BS throughput. The BS with the highest throughput has a rank of 1.

Figure 15 shows the throughput of each BS versus the BS rank. We can see from the figure that in Two-hop, the throughput of the first two BSes is extremely high while the throughput of the last two BSes is extremely small. This is because the two hop routing path length in Two-hop is not long enough to forward messages from a congested BS to a lightly loaded BS. Therefore, the traffic cannot be shifted to the neighboring lightly loaded BSes, leading to an unbalanced load distribution. We can also see from the figure that in DTR, the variance of the throughputs in different BSes is small. The reason is that three forwarding hops are enough for a mobile node to reach a neighboring BS and hence to balance the load between the BSes. Meanwhile, the congestion control algorithm in DTR can effectively switch the traffic from a highly loaded cell to a lightly loaded cell. Because the BSes of ranks 1 and 2 in DTR are not congested, their throughput is less than the corresponding BSes in Two-hop; also, the throughput of the BSes of ranks 3 and 4 in DTR is much higher than that of the corresponding BSes in Two-hop. DHybrid achieves more balanced load distribution between BSes than Two-hop since it employs a congestion control algorithm. In DHybrid, if a previously established path to a destination is not broken, a node still uses this path to transmit messages to the same destination. Thus, the nodes cannot dynamically balance load between BSes. Also, when a node finds that its current BS is congested, it takes a long time for it to find a path to a non-congested BS by re-issuing a query message to the neighboring non-congested BS, which greatly reduces the throughput of the system.

Figure 16 further shows the throughput of the BSes versus simulation time in the three routing protocols. At the beginning, the BSes with ranks 1 and 2 are congested and those with ranks 3 and 4 do not have much traffic. Thus, the three figures show that the BSes with ranks 1 and 2 have high throughput but those with ranks 3 and 4 have extremely low throughput at the beginning in all three protocols. Figure 16 (a) shows the throughput of the BSes in DTR. As shown in the figure, since DTR can adaptively adjust the traffic among the BSes using its congestion control algorithm, the throughput of the two highly congested BSes is distributed to the neighboring BSes. As the traffic is forwarded from the BSes of ranks 1 and 2 to the BSes of ranks 3 and 4, the throughputs of these BSes are very similar later in the simulation. This result indicates the effectiveness of the congestion control algorithm in DTR for load balance between cells.

Figure 16 (b) shows the throughput of the BSes in Two-hop. In Two-hop, since the source nodes cannot effectively move the traffic between BSes, the BSes with rank 1 and rank 2 constantly have the highest throughput, while the BSes with rank 3 and rank 4 constantly have low throughput. The low throughput is produced when the immediate neighbors of the source node are in the range of the neighboring BSes of the source node's BS. However, the probability of such cases is very small. Figure 16 (c) shows the throughput of the BSes in DHybrid. As the nodes in DHybrid cannot effectively balance the load between the BSes, the throughput of the BSes of rank 1 and rank 2 is much larger than that of the BSes of rank 3 and rank 4. Comparing Figure 16 (b) and Figure 16 (c), we can find that the throughput in DHybrid is lower than that in Two-hop. This is because the multi-hop transmission in the ad-hoc network in DHybrid greatly reduces the throughput. Meanwhile, the mobile gateway nodes in DHybrid easily become congested, leading to more message drops.

6 Conclusions

Hybrid wireless networks have been receiving increasing attention in recent years. A hybrid wireless network combining an infrastructure wireless network and a mobile ad-hoc network leverages their advantages to increase the throughput capacity of the system. However, current hybrid wireless networks simply combine the routing protocols in the two types of networks for

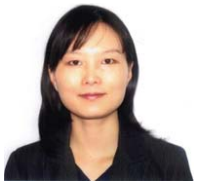
data transmission, which prevents them from achieving higher system capacity. In this paper, we propose a Distributed Three-hop Routing (DTR) data routing protocol that integrates the dual features of hybrid wireless networks in the data transmission process. In DTR, a source node divides a message stream into segments and transmits them to its mobile neighbors, which further forward the segments to their destination through an infrastructure network. DTR limits the routing path length to three, and always arranges for high-capacity nodes to forward data. Unlike most existing routing protocols, DTR produces significantly lower overhead by eliminating route discovery and maintenance. In addition, its distinguishing characteristics of short path length, short-distance transmission, and balanced load distribution provide high routing reliability and efficiency. DTR also has a congestion control algorithm to avoid load congestion in BSes in the case of unbalanced traffic distributions in networks. Theoretical analysis and simulation results show that DTR can dramatically improve the throughput capacity and scalability of hybrid wireless networks due to its high scalability, efficiency, and reliability and low overhead.

ACKNOWLEDGEMENTS

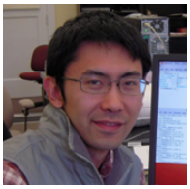
This research was supported in part by U.S. NSF grants IIS-1354123, CNS-1254006, CNS-1249603, CNS-1049947, CNS-0917056 and CNS-1025652, Microsoft Research Faculty Fellowship 8300751, and the United States Department of Defense 238866. We would like to thank Mr. Kang Chen for his help in addressing review comments.

REFERENCES

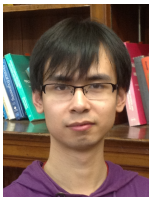
[1] H Luo, R. Ramjee, P. Sinha, L. Li, and S. Lu. Ucan: A unified cell and ad-hoc network architecture. In *Proc. of MOBICOM*, 2003.
[2] P. K. McKinley, H. Xu, A. H. Esfahanian, and L. M. Ni. Unicast-based multicast communication in wormhole-routed direct networks. *TPDS*, 1992.
[3] H. Wu, C. Qiao, S. De, and O. Tonguz. Integrated cell and ad hoc relaying systems: iCAR. *J-SAC*, 2001.
[4] Y. H. Tam, H. S. Hassanein, S. G. Akl, and R. Benkoczi. Optimal multi-hop cellular architecture for wireless communications. In *Proc. of LCN*, 2006.
[5] Y. D. Lin and Y. C. Hsu. Multi-hop cellular: A new architecture for wireless communications. In *Proc. of INFOCOM*, 2000.
[6] P. T. Oliver, Dousse, and M. Hasler. Connectivity in ad hoc and hybrid networks. In *Proc. of INFOCOM*, 2002.
[7] E. P. Charles and P. Bhagwat. Highly dynamic destination sequenced distance vector routing (DSDV) for mobile computers. In *Proc. of SIGCOMM*, 1994.
[8] C. Perkins, E. Belding-Royer, and S. Das. RFC 3561: Ad hoc on demand distance vector (AODV) routing. Technical report, Internet Engineering Task Force, 2003.
[9] D. B. Johnson and D. A. Maltz. Dynamic source routing in ad hoc wireless networks. *IEEE Mobile Computing*, 1996.
[10] V. D. Park and M. Scott Corson. A highly adaptive distributed routing algorithm for mobile wireless networks. In *Proc. of INFOCOM*, 1997.
[11] R. S. Chang, W. Y. Chen, and Y. F. Wen. Hybrid wireless network protocols. *IEEE Transaction on Vehicular Technology*, 2003.
[12] G. N. Aggelou and R. Tafazolli. On the relaying capacity of next-generation gsm cellular networks. *IEEE Personal Communications Magazine*, 2001.
[13] T. Rouse, I. Band, and S. McLaughlin. Capacity and power investigation of opportunity driven multiple access (ODMA) networks in TDD-CDMA based systems. In *Proc. of ICC*, 2002.
[14] H. Y. Hsieh and R. Sivakumar. On Using the Ad-hoc Network Model in Wireless Packet Data Networks. In *Proc. of MOBIHOC*, 2002.
[15] L. M. Feeney, B. Cetin, D. Hollos, M. Kubisch, S. Mengesha, and H. Karl. Multi-rate relaying for performance improvement in ieee 802.11 wlans. In *Proc. of WWIC*, 2007.
[16] J. Cho and Z. J. Haas. On the throughput enhancement of the downstream channel in cellular radio networks through multihop relaying. *IEEE JSAC*, 2004.
[17] B. Liu, Z. Liu, and D. Towsley. On the capacity of hybrid wireless networks. In *Proc. of INFOCOM*, 2003.
[18] H. Y. Hsieh and R. Sivakumar. A hybrid network model for wireless packet data networks. In *Proc. of GLOBECOM*, 2002.
[19] Y. Wei and D. Gitlin. Two-hop-relay architecture for next-generation WWAN/WLAN integration. *IEEE Wireless Communication*, 2004.
[20] X. J. Li, B. C. Seet, and P. H. J. Chong. Multihop cellular networks: Technology and economics. *Computer Networks*, 2008.
[21] B. Bengfort, W. Zhang, and X. Du. Efficient resource allocation in hybrid wireless networks. In *Proc. of WCNC*, 2011.
[22] P. Thulasiraman and X. Shen. Interference aware resource allocation for hybrid hierarchical wireless networks. *Computer Networks*, 54(13):2271–2280, 2010.
[23] K. Akkarajitsakul, E. Hossain, and D. Niyato. Cooperative packet delivery in hybrid wireless mobile networks: A coalitional game approach. *IEEE Trans. Mob. Comput.*, 12(5):840–854, 2013.
[24] T. Liu, M. Rong, P. Li, D. Yu, Y. Xue, and E. Schulz. Radio resource allocation in two-hop cellular relaying network. In *Proc. of VTC*, 2006.
[25] T. Liu, M. Rong, H. Shi, D. Yu, Y. Xue, and E. Schulz. Reuse partitioning in fixed two-hop cellular relaying network. In *Proc. of WCNC*, 2006.
[26] L. Guan, J. Zhang, J. Li, G. Liu, and P. Zhang. Spectral efficient frequency allocation scheme in multihop cellular network. In *Proc. of VTC*, 2007.
[27] D. M. Shila, Y. Cheng, and T. Anjali. Throughput and delay analysis of hybrid wireless networks with multi-hop uplinks. In *Proc. of INFOCOM*, 2011.
[28] B. Liu, P. Thiran, and D. Towsley. Capacity of a wireless ad hoc network with infrastructure. In *Proc. of Mobihoc*, 2007.
[29] C. Wang, X. Li, C. Jiang, S. Tang, and Y. Liu. Multicast throughput for hybrid wireless networks under gaussian channel model. *TMC*, 10(6):839–852, 2011.
[30] A. G. Valko. Cellular ip: A new approach to internet host mobility. *ACM Computer Comminication*, 1999.
[31] C. Sarr, C. Chaudet, G. Chelius, and I. G. Lassous. A Node-Based Available Bandwidth Evaluation In IEEE 802.11 Ad Hoc Networks. *IJPEDS*, 00(00):1–21, 2005.
[32] X. P. Costa, M. T. Moreno, and H. Hartenstein. A simulation study on the performance of hierarchical mobile ipv6. In *Proc. of ITC*, 2003.
[33] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE TIT*, 2000.
[34] L. B. Koralov and Y. G. Sinai. Theory of probability and random processes. *Berlin New York Springer*, 2007.
[35] H. Y. Hsieh and R. Sivakumar. Performance comparison of cellular and multi-hop wireless networks: A quantitative study. In *Proc. of SIGMETRIC*, 2001.
[36] The network simulator - ns-2. http://www.isi.edu/nsnam/ns/.
[37] M. Grossglauser and D. Tse. Mobility increases the capacity of ad hoc wireless networks. In *Proc. of TON*, 2002.



Haiying Shen received the BS degree in Computer Science and Engineering from Tongji University, China in 2000, and the MS and Ph.D. degrees in Computer Engineering from Wayne State University in 2004 and 2006, respectively. She is currently an Associate Professor in the Department of Electrical and Computer Engineering at Clemson University. Her research interests include distributed computer systems and computer networks, with an emphasis on content delivery networks, mobile computing, wireless sensor networks, and cloud. She is a Microsoft Faculty Fellow of 2010, a senior member of the IEEE and a member of the ACM.



Ze Li received the BS degree in Electronics and Information Engineering from Huazhong University of Science and Technology, China, in 2007 and the Ph.D. degree in the Department of Electrical and Computer Engineering of Clemson University, in 2012. His research interests include distributed networks, with an emphasis on content delivery networks, wireless multi-hop cellular networks, game theory and data mining.



Chenxi Qiu received the BS degree in Telecommunication Engineering from Xidian University, China, in 2009. He currently is a Ph.D. student in the Department of Electrical and Computer Engineering at Clemson University, SC, United States. His research interests include sensor networks and wireless networks.