Dept - Reseaux et Services des Telecommunications

Institut National des Telecommunications

9 rue Charles Fourier, 91011 Evry, France

Universitatea POLITEHNICA din Bucureşti

Facultatea de ELECTRONICĂ, TELECOMUNICAŢII

şi TEHNOLOGIA INFORMAŢIEI

Bd. Iuliu Maniu nr. 1-3, Sector 6, Bucureşti,

Challenge in Mars Planetary MANETs

In partial fulfillment of the requirements for the

Diploma Degree

CONSTANTINESCU MIHAI MARIUS

Under the Direction of

Joséphine Kohlenberg

Serban Obreja

[2009]

ABSTRACT: Mobile Ad hoc NETworks (MANETs) are very useful in such situations and are currently being researched for their autonomous characteristics and well-suited characteristics for deployment in adverse conditions such as the earth's upper atmosphere or outer space. However, traditional MANET protocols are not directly applicable to this situation as Inter-Planetary Area Networks (IPAN) tends to be extremely sparse and the node velocities are very high. We propose to make a comparison study on Earth MANET with Mars Surface Exploration Ad-Doc network and mapping the QoS requirement between two with possible solutions.

Keywords: MANETS; IPAN; QoS;

Table of Contents

CHAPTER 1

1. Introduction

1.1 Research motivation

Wireless mobile ad hoc networks (MANET) attracted recently more and more interest as an infrastructureless approach for military use, emergency use, wireless sensor networks or other situations where it is required that a wireless network can be set up on-demand, automatically, and instantly.

Mobile Ad hoc NETworks (MANETs) are very useful in such situations and are currently being researched for their autonomous characteristics and well-suited characteristics for deployment in adverse conditions such as the earth's upper atmosphere or outer space. But we propose to use MANET protocols in Inter-Planetary Area Networks (IPAN) where it cannot directly be applicable to this situation as tends to be extremely sparse and the node velocities are very high. We propose a comparison study on Earth MANET with Mars Surface Exploration Ad-Doc network and mapping the QoS requirement between two with possible solutions.

Furthermore, the developments in space technologies enable the realization of deep-space missions such as Mars exploration. The Interplanetary Internet is envisioned to provide communication services for scientific data delivery and navigation services for the explorer spacecrafts and orbiters of future deep-space missions. The unique characteristics posed by deep-space communications call for different research approaches from those in terrestrial networks.

In the future, it is expected to send more probes for outer space exploration, and also to make earth-science mission more sophisticated by placing huge groups of satellites, called constellations, orbiting earth and other planets to facilitate scientific research. Like the mars lander, these satellites gather data from outer space about distant planets and distant galaxies and beam the data down to earth by suitable means. The challenge is to manage such a grand scheme considering the distances between the planets and the diverse conditions of operations in outer space.

1.2 Thesis outline

This paper is organized as follows – In section II we discuss about the most important Concepts and Characteristics for Mobile Ad Hoc Networks, MANET topologies, Applications, Routing protocols for MANETs and Metrics. We also refer to the Challenges in MANETs routing and routing problems.

In section III, we talk about the Mars Surface Exploration Ad hoc network: History of Mars Exploration, Architecture of a Mars Surface Exploration Ad hoc network and Challenges imposed by this type of environment.

We then propose in IV some solutions to the problems and refer to DTN and QoS protocols for MANETs.

We conclude with DTN simulations made in the ONE environment and an overview.

CHAPTER 2

2. Overview of MANETs

2.1 Concepts and Characteristics for Mobile Ad Hoc Networks

A mobile ad-hoc network (MANET) is a collection of nodes, which have the possibility to connect on a wireless medium and form an arbitrary and dynamic network with wireless links. This collection of wireless mobile nodes forming a temporary network without the support of any established infrastructure or centralized administration it is composed from links that can change during time, new nodes that can join the network, and other nodes that can leave it. A MANET is expected to be of larger size than the radio range of the wireless antennas, because of this fact it could be necessary to route the traffic through a multi-hop path to give two nodes the ability to communicate. There are neither fixed routers nor fixed locations for the routers as in cellular networks (infrastructure networks). Cellular networks consist of a wired backbone which connects the base-stations. The mobile nodes can only communicate over a one-hop wireless link to the base station; multi-hop wireless links are not possible. By contrast, a MANET having no permanent infrastructure at all permits all mobile nodes to act as mobile routers.

2.2 History

The term „MANET" [TOPOLOGY CONTROL, ROUTING PROTOCOLS AND PERFORMANCE EVALUATION FOR MOBILE WIRELESS AD HOC NETWORKS] is relatively new, but the concept of mobile packet radio networks, where every node in the network is mobile and where multi hop (store-and-forward) routing wireless is utilized, dates back to the seventies. In 1972 DARPA initiated a research effort to develop and demonstrate a packet radio network (PRNet), which was to provide an efficient means of sharing a broadcast radio channel as well as coping with changing and incomplete connectivity. The initial PRNet protocols adopted a centralized control station, but the core PRNet concept quickly evolved into a distributed architecture consisting of a network of broadcast radios with minimal central control, using multi hop store-and-forward routing techniques to create complete end-user connectivity from incomplete radio connectivity. The original ideas about MANET came out during the research of PRNet. In order to enhance the robustness and scalability, DARPA initiated the Survivable Radio Networks (SURAN) program in 1983. With the widespread implementation of Internet and Web technology, both commercial and defense sectors are motivated to extend the global information infrastructure into the mobile wireless environment, DARPA funded the Global Mobile (GloMo) Information Systems program in 1994 that has just recently concluded. The goal of the GloMo program was to make the mobile, wireless environment a first-class citizen in the defense information infrastructure by providing user friendly connectivity and access to services for mobile users. Interest in the improvements based packet radio networks was not limited to the United States. Many countries have proposed their own solutions.

MANET is a new network architecture that absorbed the advantages of the previous research results.

So far, many international conferences and workshops on MANET have been held by e.g. IEEE and ACM. For instance, MobiHoc (The ACM Symposium on Mobile Ad Hoc Networking & Computing) has been one of the most important conferences of ACM SIGMOBILE (Special Interest Group on Mobility of Systems, Users, Data and Computing). Among the many research issues, it has been widely recognized that routing strategy is the most important.

Therefore, the Internet Engineering Task Force MANET Working Group (IETF MANET WG) has been formed since 1996, aiming to investigate and develop candidate standard Internet routing support for mobile, wireless IP autonomous segments and develop a framework for running IP based protocols in ad hoc networks. There are nearly a dozen candidate routing protocols currently being discussed within the MANET WG for achieving this goal.

2.3 Topology

Figure 1 [Ad-hoc and Hybrid Networks Performance Comparison of MANET Routing Protocols in Ad-hoc and Hybrid Networks] describes an example of MANET including four mobile nodes. In this example, nodes S, B, and C are within the transmission ranges of each other. Thus, nodes S, B, and C are named neighbors, and they can communicate directly with each other. However, node D does not reside in the transmission range of node S. If node S wants to send data to node D, the data must be routed through the intermediate node, such as node C, which acts as the router between node S and node D. Routes between a source and destination may potentially contain multiple hops.

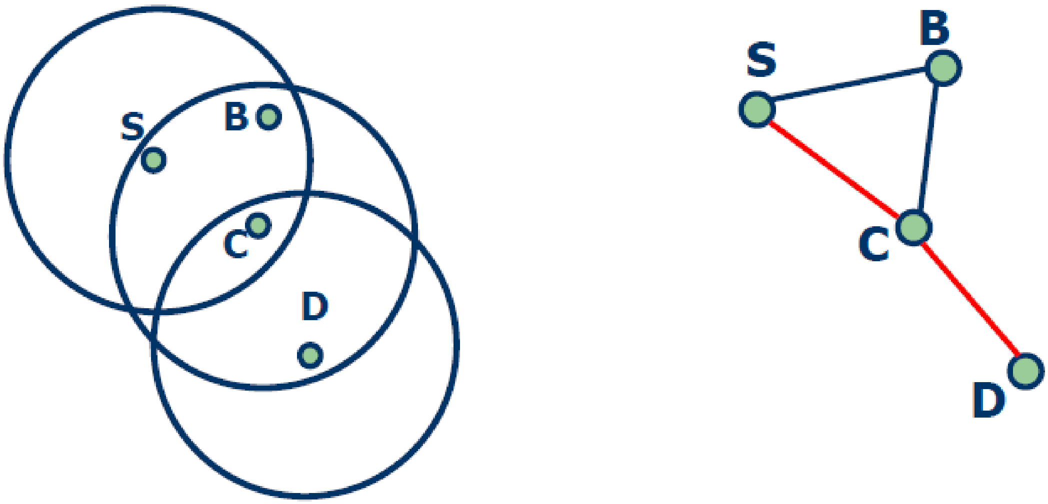
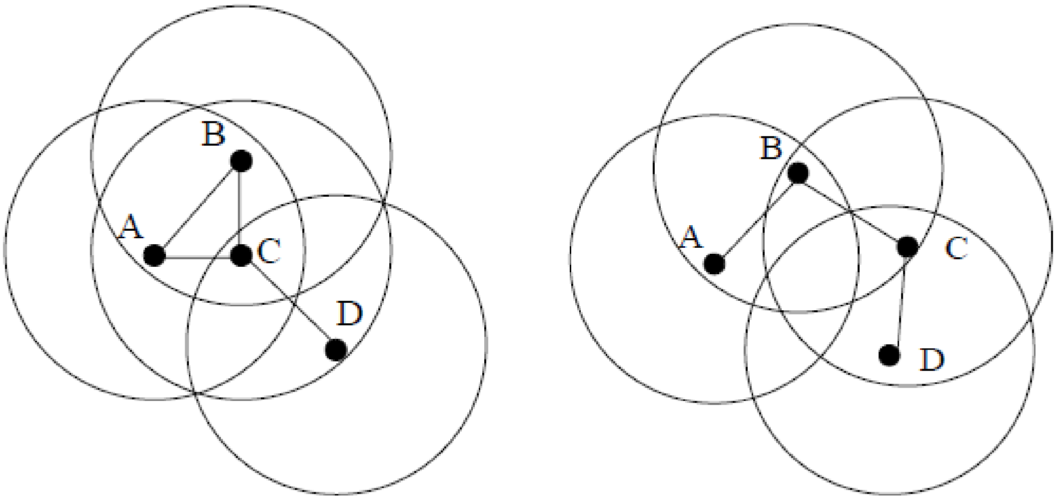


Figure 1: Example for mobile ad hoc networks



(a) A Mobile Ad-Hoc Network Before Movement

(b) A Mobile Ad-Hoc Network After Movement

Figure 2 shows a simple ad-hoc network of three nodes. Nodes A-B and B-C are in communication range of each other so they can send data directly to each other, if the data has to be send from A-C, then node B has to forward the data to C on behalf of node A.

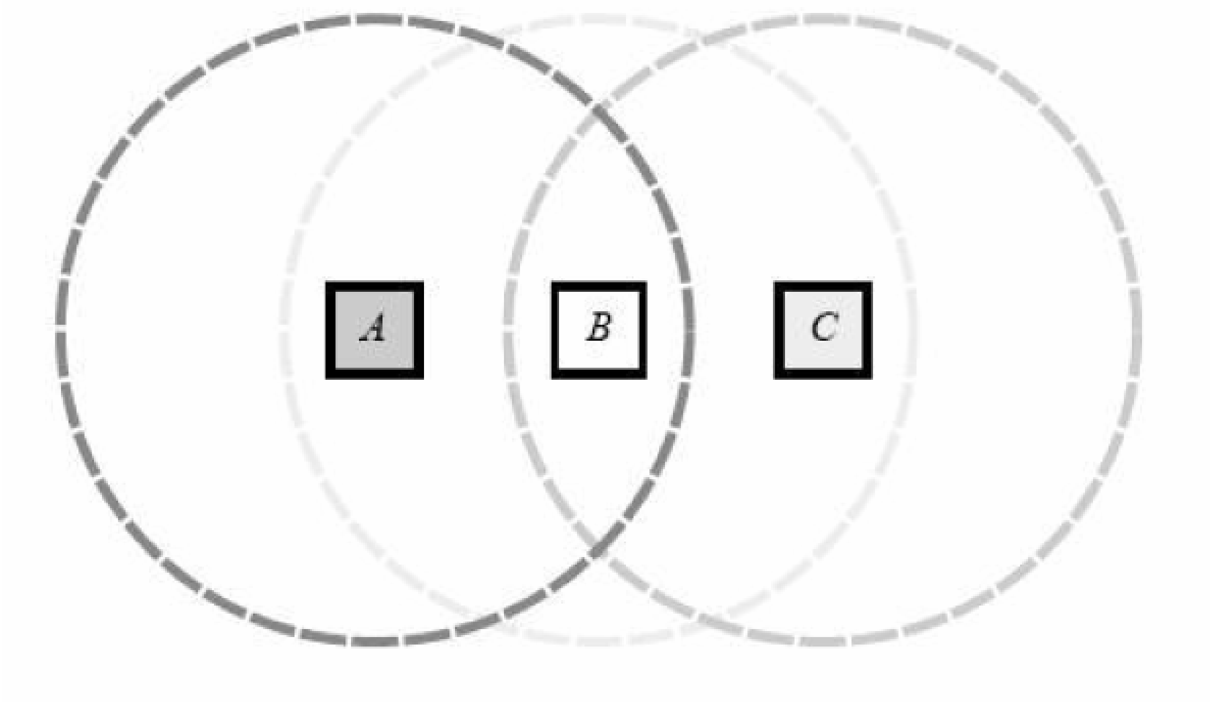


Figure 2: A simple MANET consisting of 3 nodes

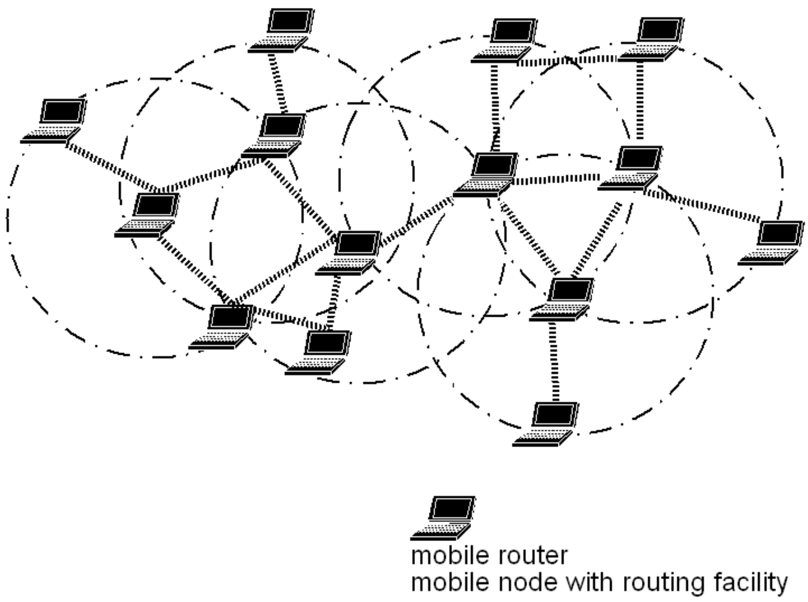


Figure 3: Mobile Ad hoc Network

Hybrid Networks

In hybrid networks the concepts of cellular networks and mobile ad-hoc networks are mixed. On one side we have a cellular network, on the other side there are mobile nodes with routing facilities.

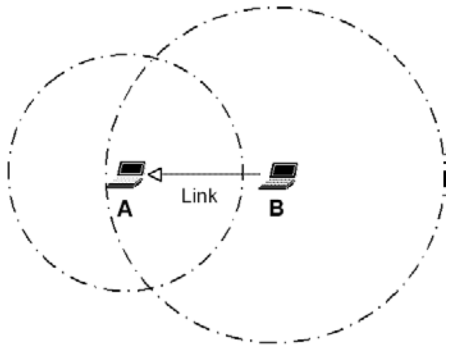


Figure 4: Asymmetric link

With this approach it is possible to have multi-hop routes between mobile nodes and the base-station. The covered area of a base-station becomes larger. The idea is to gain more efficiency out of the existing infrastructure, to cover wider areas with less fixed antennas and base-stations and to reduce power consumption. There are many benefits of enhancing cellular network with ad-hoc technologies.

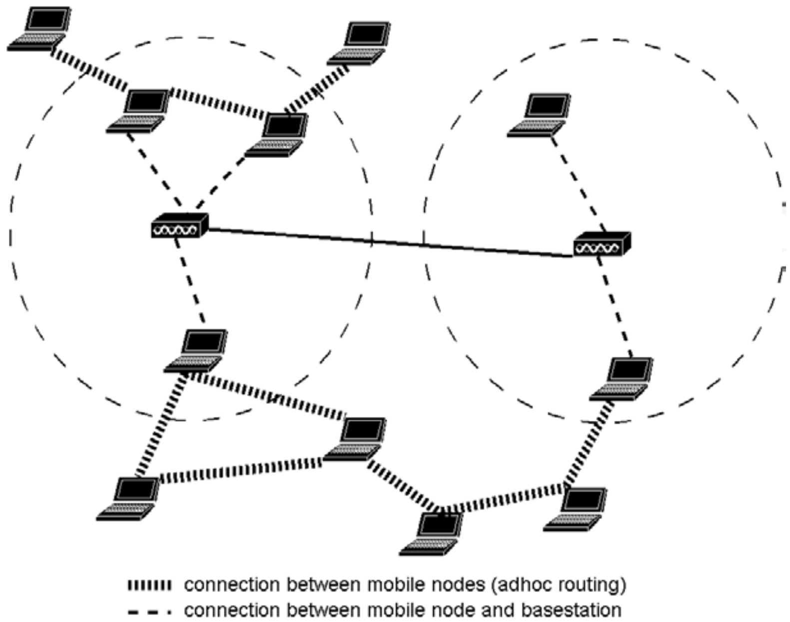


Figure 5: Hybrid network

2.4 Application

There are many potential applications based the techniques of MANETs, such as disaster rescue, personal area networking, wireless conference, military applications and many others.

But in this paper, we present besides traditional MANET applications some other research for adverse conditions such as the earth's upper atmosphere or outer space.

2.4.1 Military application

The research of MANETs originated from military application. In the future providing services for military field is still an important topic because MANETs perfectly satisfy military needs like battlefield survivability. Wireless electronic devices carried in soldiers, tanks, airplane and other military equipment can form MANETs to support communication among them in order to collaboratively achieve military goals since there is not any pre-placed infrastructure and connectivity in battlefield environments.

2.4.2 Emergency services

There are situations in which loss of local power causes loss of electricity, and each year natural disasters destroy people lives around the world. As the Internet grows in importance, the loss of network connectivity during such natural disasters will become an ever more noticeable consequence of the calamity. Furthermore, network applications will become increasingly important for emergency services, and thus it will be important to find ways to enable the operations of networks even when infrastructure elements have been disabled as part of the effects of a disaster.

An example [Quality of Service Aware Routing Protocols for Mobile Ad Hoc Networks] is the design for future public safety communications. A European project called Wireless Deployable Network System (WIDENS) concentrated their work on this field. WIDENS have an idea that using ad hoc network to interoperate with existing TETRA network which is used for public safety

2.4.3 Mobile conferencing

When mobile computer users gather outside their normal office environment, the business network infrastructure is often missing. But the need for collaborative computing might be even more important here than in the everyday office environment. As it turns out, the establishment of a MANET for collaborative mobile computer users is needed even when there may be Internet infrastructure support available.

2.4.4 Personal area networks

The idea of a personal area network (PAN) is to create a much localized network populated by some network nodes that are closely associated with a single person. Some visions of the future include virtual reality devices attached and other devices more oriented toward the sense of touch. These devices may or may not need to have an attachment to the Internet, but they will almost certainly need to communicate with each other while they are associated with their user's activities. In this scenario, mobility is not the overriding consideration.

Actually Bluetooth is a commercial application of MANETs that is designed to eliminate wires between personal devices. Currently Bluetooth only provides short distance communication. However, when interactions between several PANs are needed, mobility becomes suddenly much more important. In other words, when people meet in real life, their PANs are likely to become aware of each other also. Methods for establishing communications between nodes on separate PANs could benefit from the technologies of MANETs.

2.4.5 Embedded computing applications

Some researchers predict a world in which computers will be around us, constantly performing tasks to make our lives a little easier. These computers will often react to the changing environment in which they are situated and will themselves cause changes to the environment in ways that are predictable and planned. Intelligent internetworking devices that detect their environment interact with each other, and respond to changing environmental conditions will create a future that is as challenging to imagine as a science fiction scenario. These capabilities can be provided with the use of MANETs.

2.4.6 Wireless mesh networks

Wireless mesh networks are ad hoc wireless networks which are formed to provide communication infrastructure using mobile or fixed nodes/users. The mesh topology provides alternative path for data transmission from the source to the destination. It gives quick re-configuration when the firstly chosen path fails. Wireless mesh network should be capable of self-organization and self-maintenance. The main advantages of wireless mesh networks are high speed, low cost, quick deployment, high scalability, and high availability. For example, if IEEE 802.11a is used, the speed can be up to 54 Mbps. An application example of wireless mesh network could be a wireless mesh networks in a residential zone, which the radio relay devices are built on top of the rooftops. In this situation, once one of the nodes in this residential area is equipped with the wired link to the internet, this node could be the gateway node. Others could connect to the internet from this node. Other possible deployments are highways, business zones, and university campus.

2.4.7 Wireless sensor networks

Wireless sensor networks use sensors to provide a wireless communication infrastructure. Sensor nodes are tiny devices used for sensing physical parameters, processing data, and communicating over the networks to the monitoring station. The application areas are military, health care, home security and environmental monitoring. There are some special characteristics which make sensor network different from other ad hoc networks.

In the sensor network, nodes could be assumed to be static, that is, sensor networks need not to be in all cases designed to support the mobility. In addition, power constraint is one of the most important factors that have to be considered carefully. The limitation of power is mainly caused by the working environment of sensor network which is often harsh. As a result, it is impossible to recharge a sensor node battery, so effective protocols are required. For example, in the network layer, people need to design a low power consumption routing protocol, and power consumption will give the first priority to be considered during the route selection phase.

The deployment and operation of self-organizing sensor networks [Load Balanced, Energy-Aware Communications for Mars Sensor Networks] is envisioned to play a key role in space exploration, such as for future in situ exploration of Mars. Sensors are equipped with several measurement instruments and are able to cooperate autonomously and to collect scientific measurements (seismic, chemicals, temperature, etc.). One or more landers or rovers functioning as base stations periodically (or on demand) collect measurements and relay the aggregated sensor field results to an orbiter and from the orbiter back to Earth.

2.4.8 Mobile Satellite Earth-stations

[A Mobile Ad Hoc Network with Mobile Satellite Earth-stations] Ad hoc networks face the problem of improving networks' capacity and scalability. The scalability problem can be properly solved through physical hierarchy networking. A new type of hierarchal mobile ad hoc network is called the Mobile Ad hoc Network, with Mobile Satellite Earth stations (MANMSE), which can effectively improve the network's scalability. MANMSE has some advantages over other physically hierarchical ad hoc networks supported by Aerial Platforms and other mobile backbones in characteristics such as greater flexibility, large coverage and higher scalability.

2.4.9 Mars Proximity Networks

The challenges [Proximity Networks Technology Assessment] of designing network architectures and protocols for mobile ad hoc networks are similar to those presented by proximity networks. The participants in an ad hoc network are not known in advance. Rather, the network is composed of the nodes in the same general area (or "proximity") that wish to communicate. The nodes must identify their neighbors and determine routes within the network. In some cases, some of the nodes are able to communicate with, and act as gateways for, external networks, such as the Internet. In these configurations, information about external connectivity must be propagated throughout the ad hoc network.

2.4.10 Inter-Planetary Area Networks (IPANs)

The need for networks that span inter-planetary distances [Challenges in Interplanetary MANETs], or Inter-Planetary Area Networks (IPANs), may be reality. These networks pose unique challenges on end-to-end communications with QoS, reliable information exchange under frequent topology changes, and rapid service provisioning. Mobile Ad hoc NETworks (MANETs) are very useful in such situations.

Inter-planetary spacecraft are key players in the deployment of IPAN MANETs because they bridge the communication between smaller, less powerful spacecraft in orbit around neighboring planets. Planetary exploration has opened huge possibilities in the area of IPAN networking. For example, scientists are exploring Mars as a potential habitat for humans in the future, and in the future, a constellation of spacecraft (which could be any combination of LEO/MEO/GEO satellites, ULDB and MEMS devices) which gather real-time data at Mars will need to relay the information to earth-based stations.

2.5 Challenges in MANETs

MANETs are distinguished from other communication networks by the following features:

2.5.1 Limited Resources

Battery is the only source of power for nodes in many ad hoc network environments, and the need to keep these nodes compact, light and even wearable, imposes limitation on their storage and processing capabilities. This is again very different from conventional wired networks, wherein the network nodes seldom depend on batteries as their sole source of energy, and typically have significant storage and processing capacity.

2.5.2 Multiple Roles

In most wired networks, network nodes play distinct roles, such as sources, destinations or routers. Also, nodes are typically dedicated to specific network operations and their characteristics well suited to the role they play. For example, machines are specifically designed to operate as servers, and dedicated high-end routers and switches are used to handle network traffic. On the other hand, in ad hoc networks, most nodes are expected to route packets for other nodes in the network, while they themselves may also be a source or destination for one or more application flows. A management framework must account for the multiple roles played by network nodes.

2.5.3 Dynamic Topology

The topology of a MANET can change very dynamically for various reasons. In MANETs, the topology changes as nodes move out of range of one or more nodes with which they were connected, and move closer and connect to other nodes. In addition, even in fixed wireless ad hoc networks (wireless sensor fields), due to limited survivability of wireless links (subject to fading or jamming) or of the nodes themselves (damaged due to hostile conditions, or discharged battery), the logical topology of the network may change.

Keeping current knowledge of the network topology is an important requirement in any network management system. In fixed wired networks, this is a relatively simple task since the changes in topology (mainly due to node or link failure, or addition/removal of a node) are infrequent. In a wireless ad hoc environment, it is crucial that the management system keep up with the frequent topology changes. However, the frequent exchange of topology information may lead to considerable signaling overhead, congesting low bandwidth wireless links, and possibly depleting the limited battery life of the nodes involved. Hence, the choice of mechanism used to collect or manage topology information is critical.

2.5.4 Low Bandwidth, Variable Capacity Links

Wireless links are typically more bandwidth-constrained than their wired counterparts. Fading, interference or jamming may cause intermittent link failures or considerable variation in the channel error rate. In addition, the diverse nature of nodes (with different transmission power levels), and communication technologies (IEEE 802.11, direct line of sight UHF/VHF links, satellite links) being used may lead to links of varying capacity in multi-hop wireless networks.

2.5.5 Heterogeneity

Heterogeneity is inherent to most ad hoc networks due to the diverse nature of communication technologies (IEEE 802.11, line of sight UHF/VHF, etc.) that may be used and the different types of nodes' ranging from sensors, palmtops, laptops to mobile networks hosted on a ship, a tank or an airplane's that may form the network. The heterogeneity of nodes can be a criterion to assign roles (management server versus client) to the various nodes. An ad hoc network may also be a result of a multi organization consortium and present additional interoperability challenges for a management system.

2.5.6 Limited Survivability

One of the major challenges in using ad hoc networks is their limited survivability and vulnerability to security attacks. The use of a wireless medium for communication opens another issue of initiating link level attacks ranging from passive eavesdropping to message replay and message distortion. In addition, deployment of wireless ad hoc networks in diverse and often hostile environments (rapidly deployed military battle-site network, sensor fields used to collect sensitive data in remote, unmanned locations) makes these networks even more prone to network security attacks leading to failure of network elements.

2.6 Routing Problem of MANETs

Routing protocols are used to determine paths through the network so a data packet can get from its source, hop by hop, to its destinations. In general, one goal of routing is to choose a suitably efficient path, where efficiency can be measured in terms of end-to-end delay, packet delivery ratio, power expended, and amount of self interference.

Routes between a source and a destination may potentially contain an ordered series of intermediate nodes that act as the routers. The multiple hops communication has some performance advantages compared with single hop communication solution:

1) Adaptability: by deploying a multiple hop data forwarding network, packets can be routed around obstructions or areas that are not anymore available (captured by enemy units for a battlefield scenario).2) Spatial reuse: packet forwarding over multiple hops via small radii transmissions will exploit spatial reuse, by allowing multiple concurrent packet transmissions in different regions of the network, and maximize throughput.

3) Energy consumption efficiency: packet forwarding via multiple small radii transmissions as opposed to a single large radius transmission will improve the throughput per unit energy.

In traditional hop-by-hop solutions to the routing problem, each node in the network maintains a routing table containing each destination with a corresponding next hop node and link cost. Packets are forwarded by consulting the routing table for the next-hop node leading to the shortest path to the destination. In MANETs, the routing table at each node can be thought of as a view into part of a distributed data structure that, when taken together, describes the topology of the network. The goal of the routing protocol is to ensure that the overall data structure contains a consistent and correct view of the actual network topology.

One challenge in creating a routing protocol for ad hoc networks is to design a single protocol that can adapt to the wide variety of conditions that can be present in any ad hoc network over time. For example, the bandwidth available between two nodes in the network may vary from more than 10 Mbps to 10 Kbps or less. The highest speeds are achieved when using high-speed network interfaces with little interference, and the extremely low speeds may arise when using low-speed network interfaces or when there is significant interference from outside sources or other nodes transmitters. Similar to the potential variability in bandwidth, nodes in an ad hoc network may alternate between periods during which they are stationary with respect to each other and periods during which they change topology rapidly. Conditions across a single network may also vary, so while some nodes are slowly moving, others change location rapidly. Another challenge for routing is that mobility causes the next-hop node to be disconnected as nodes move in and out of transmission range. The result is that routes are frequently broken causing extra network traffic to reconstruct the routing table. If there is a high frequency of broken links, the overhead cost of routing can dominate the traffic load causing congestion and consuming precious energy in an attempt to discover unstable pathways. Thus routing protocols in MANET should have the capability to handle dynamic connectivity.

The mobility of a given host causes some problems such as co-channel interference, dynamic topology and the multi-hop route. MANET faces the challenge of improving the network’s capacity and enlarging its scalability. The scalability problem can be properly solved through physical hierarchy networking.

The dynamic nature of MANETs leads to the difficulty in addressing, routing and data delivering. A solution [A Novel Solution for Global Connectivity in MANET] for Internet access is by using multiple access routers (ARs) simultaneously. This brings a lot of benefits such as effective soft-handoff, throughput improvement, and load balancing between ARs. In addition, hierarchical mobile IP (HMIP) is incorporated into the solution to speed up handoff, and to maintain the Internet connection using the regional care of address (RCoA). The simulation results for both proactive and reactive ad hoc routing protocols with two ARs show an improvement of network throughput, especially in the small size network.

Internet services for MANET nodes are provided by attachment points which are called Internet gateways or access routers (AR). Nodes in MANET requiring connection to the Internet have to register with ARs and obtain a global routable address. This is done through address auto configuration, gateway discovery and registration processes. Depending on what type of ad hoc routing protocol used in MANETs, these processes could be manual or proactive or integrated with the ad hoc routing protocol.

However, it's not that simple in reality because of the dynamic nature of MANET where nodes could join and leave the network at any time. It becomes even worse when more than one ARs and/or large and dynamic nodes movement are involved. This will lead to problems such as which ARs to choose, how to maintain the connection when nodes are moving or how to make a handoff decision.

The vast majority of mobile ad hoc networking research makes a very large assumption: that communication can only take place between nodes that are simultaneously accessible within in the same connected cloud (that communication is synchronous). In [Adaptive Routing for Intermittently Connected Mobile Ad Hoc Networks], this assumption is dismantled, particularly for sparsely or irregularly populated environments.

In [Adaptive Routing for Intermittently Connected Mobile Ad Hoc Networks], it is presented the Context-Aware Routing (CAR) algorithm. CAR is a novel approach to the provision of asynchronous communication in partially-connected mobile ad hoc networks, based on the intelligent placement of messages. The assumption that all communication is synchronous in nature is overly constrained if one considers that there is a strong requirement for communication that is asynchronous in nature, as argued. In such a case, the delay tolerant character of the traffic allows useful communication to still occur by using nodes moving between disconnected groups of nodes (clouds) to transport messages from one cloud to another.

Thus, it is perfectly possible that two nodes may never be part of the same connected cloud and yet may still be able to exchange delay tolerant information by making use of predicted mobility patterns as an indicator of which other nodes might make good carriers for this information.

For our discussion about Mars surface another critical issue in routing strategy design that sets the Mars sensor network apart from conventional ad hoc networks is energy conservation and prolonging network lifetime while maintaining connectivity and satisfying latency constraints.

2.7 Routing protocols for MANETs

Many routing protocols have been proposed for MANETs. Those routing protocols are classified into two main categories: topology-based and position-based.

Topology-based routing protocols are based on the information concerning links. In position-based routing protocols, mobile nodes know physical position information by geo location techniques such as GPS.

Topology based routing protocols can be generally grouped by the routing strategy into one of three categories. The first type of topology based routing is a proactive routing protocol, which maintains network topology through the periodic exchange of control information. A proactive protocol might be appropriate in a network in which non-local communications are normal and route maintenance must be rapid. Some of proactive routing protocols are distance vector typed. Pure distance vector algorithms do not perform very well in MANETs due to the count-to-infinity problem, where two nodes both believe their route back to the reference node is through each other upon loss of a link. Thus, newly proposed protocols modify and enhance the distance vector algorithms (Distributed Bellman Ford, Routing Internet Protocol (RIP)). Proactive routing protocols of this type include the Destination-Sequenced Distance Vector (DSDV) Routing, Wireless Routing Protocol (WRP), and Least Resistance Routing (LRR). Other proactive routing protocols are based on link state algorithms.

Link state based algorithms consist of Global State Routing (GSR), Fisheye State Routing (FSR), Adaptive Link-State Protocol (ALP), Source Tree Adaptive Routing (STAR), Optimized Link State Routing (OLSR) protocol, and Landmark Ad Hoc Routing (LANMAR).In general, proactive protocols are not efficient enough due to much communication overhead when nodes are mobile.

The second type of topology based routing is a reactive routing protocol, which does not require maintaining a route to each destination of the network on a continual basis. Instead, routes are established on demand by the source. When a route is needed by the source, it floods a route request packet to construct a route. Upon receiving route requests, the destination selects the best route based on route selection metrics. In reactive routing protocols, control communication overhead is greatly reduced compared with proactive routing protocols since no effect is made to maintain the total network topology. Numerous protocols of this type have been proposed, such as Lightweight Mobile Routing (LMR), Dynamic Source Routing (DSR), Ad-Hoc On Demand Distance Vector (AODV) routing, Associative -Based Routing (ABR), Signal Stability-Based Adaptive (SSA) routing, Routing On-demand Acyclic Multipath (ROAM) algorithm, Multipath Dynamic Source Routing (MDSR), Relative Distance Micro-discovery Ad Hoc Routing (RD-MAR) protocol, Efficient Secure Dynamic Source Routing (ESDSR).

The third type of topology based routing is hybrid in nature, which combines proactive and reactive techniques. As an example of a hybrid MANET routing protocol, ZPR defines a zone around each node where the local topology is proactively maintained via the Intrazone Routing Protocol (IARP). When routes are required outside the local zone, a reactive route discovery mechanism is used via the Interzone Routing Protocol (IERP). This type routing protocols include the Zone Routing Protocol (ZPR), the Bordercast Resolution Protocol, the Temporally Ordered Routing Algorithm (TORA), and the Landmark Routing Protocol (LAN-MAR). The results show that the use of location information in MANETs significantly improves routing performance. With the knowledge of a nodes geographical location, position based routing can be more effective at the cost of overhead required for exchanging location information. In position based routing, nodes maintain a location table that records the location of some other nodes and the time at which that location information was received. A sender node then uses this information to improve the efficiency in the transmission of data packets. Examples of position based routing protocols include the Location-Aided Routing algorithm (LAR), the Distance Routing Effect Algorithm for Mobility (DREAM), the Greedy Perimeter Stateless Routing algorithm (GPSR), the Geographical Routing Algorithm (GRA), the Geographic Distance routing protocol (Gedir), and the GRID protocol.

2.7.1 Proactive (Table driven) Routing

Proactive routing tries to keep up to date information about the entire network, therefore when there is a routing request, the request is fulfilled without any delay (Destination Sequenced Distance Vector (DSDV), link state routing protocol). In simple link state routing protocol each node will maintain a table containing routes to all the nodes in the network. This is done by transmitting the node’s neighbor list to all the nodes in the network. Figure 6 shows the snapshot of a network using link state routing protocol. When a route is required, the routing table is looked up and since the routing table has routes to all the nodes in the network the route is satisfied without any delay. For example, if node 0 has to find the route to destination 3, then node 0 does a lookup in its routing table for node 3. According to the routing table the data packet is forwarded to node 2. The same procedure is repeated at node 2 and finally the packet reached the destination node 3.

Such a protocol will work very effectively and efficiently in a small network scenario but when the network is scaled it's difficult for each node to keep information about each and every other node. If the topology of the network changes very frequently then the number of control messages which are exchanged will increase. Thus, most of the bandwidth will be used up by these control messages rather than the actual messages which is not acceptable in a bandwidth limited scenario. To reduce the control message traffic and for achieving better convergence rate new algorithms combining the features of distance vector and link state protocols, like the wireless routing protocol (WRP) are being used. In WRP, routing nodes communicate the distance and second-to-last hop information for each destination in the wireless networks. It avoids the "count-to-infinity" problem by forcing each node to perform consistency checks of predecessor information reported by all its neighbors. This ultimately eliminates looping situations and provides faster route convergence when a link failure event occurs.

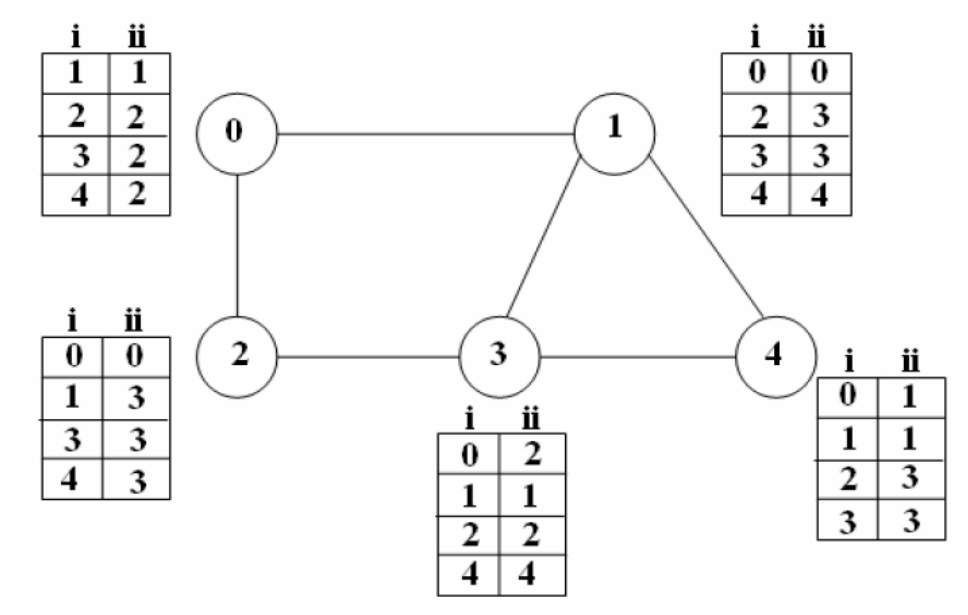


Figure 6: Example of a simple Link State Routing protocol. Col i – represents the destination id, Col. ii represents the next-hop id. Each node has route information for all other nodes in the network

2.7.2 Reactive (On-demand) Routing

Reactive routing does not gather and keep information regarding the network topology. When a message has to be routed, a route discovery procedure is invoked in which a route discovery message is flooded throughout the network. The original message is delayed till the source gets the reply for the route discovery message. The reply message for the route discovery message contains the path from the source to the destination.

Examples of reactive routing protocols are Dynamic Source Routing (DSR) and Ad Hoc On-demand Distance Vector (AODV) protocol. Once a path is discovered the path is saved in the source nodes cache with a time stamp. If any of the nodes in the path disappear, if a node is switched off or moved away) then the path until that node is retained and a new route discovery procedure is invoked to find the rest of the path to the destination.

The most popularly used ad hoc routing protocol is the DSR protocol. The source node starts the route discovery process by sending a route request to find the destination node. Figure 7 shows and example of this process to find a route form source node 1 to destination node 8. Whenever a route request packet visits a node, the packet adds that node to its visited list, as indicated by the link tags of figure 7.

The first route request packet which reaches the destination replies to the route request. Figure 8 shows how the route reply is sent back through the accumulated path (visited nodes) from node 8 back to the source.

In case of Ad Hoc On-demand Distance Vector (AODV) a similar procedure is followed, but instead of storing the routing path in the packet, each node on the path saves the next hop address for the destination.

The main advantage of reactive routing is that the control traffic in the network is minimized, but this is the cost of long setup delay therefore this scheme is not suitable for routing real-time traffic. Another drawback of this scheme is that the message size increases because the entire path information is in the message.

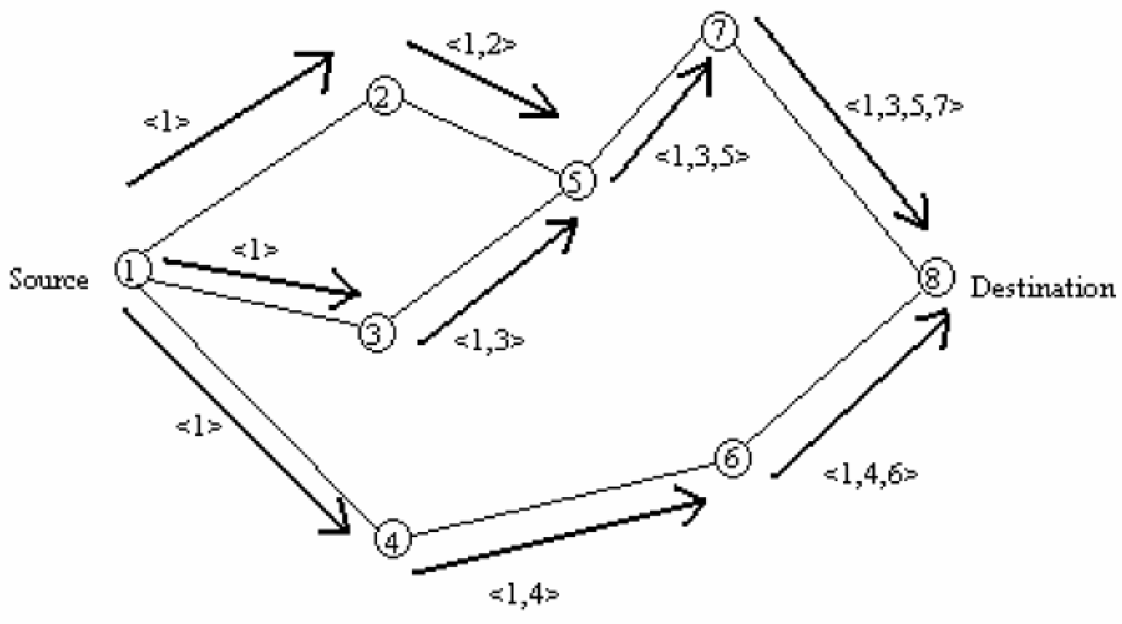
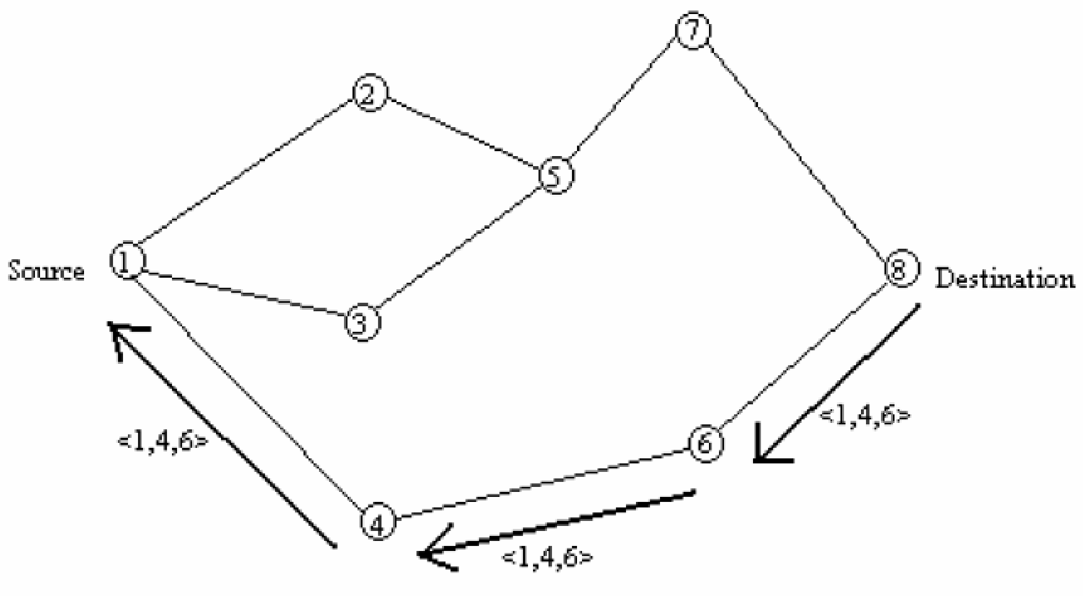


Figure 7: Building Record Route during Route Discovery

Figure 8: Propagation of Route Reply with the Route Record



2.7.3 Hybrid Routing

Ad hoc networks are used in diverse applications and it's not possible to create a single protocol which will work efficiently for all the applications. Proactive and reactive routing schemes operate efficiently in small regions of the ad hoc network design space.

Proactive routing is suited for areas where the ratio of node mobility to call rate is low, where the network topology does not change frequently, whereas reactive routing is suited for areas where the ratio of node mobility to call rate is high. The performance of both the protocols suffer when call rate and node mobility increase. Therefore new protocols have to be created which combine the strengths of existing protocols and a solution is hybridization. A lot of hybrid routing protocols [DYNAMIC VIRTUAL BACKBONE ROUTING PROTOCOL: A HYBRID ROUTING PROTOCOL FOR ADHOC NETWORKS] have been suggested for ad hoc networks: Zone Routing Protocol (ZRP), Independent Zone Routing Protocol, and Routing via dynamic group construction, AntHocNet and many more.

Most of the hybrid routing protocols have a common structure, they contain two sub protocols. A sub protocol covers some local area and the other covers some global area. The basic idea of most of the protocols is to organize a group of nodes or a single node to maintain some minimal information about the topology of its neighbors; a sub protocol governs the messages passing in this small group and to pass messages between such groups the other sub protocol (global) is used. In most of the hybrid protocols, the first sub protocol will be some kind of proactive (table based) protocol and the other will be some kind of reactive (on-demand) protocol. The main factor which affects the performance of the hybrid protocols is the decision of when to switch between proactive and reactive protocols.

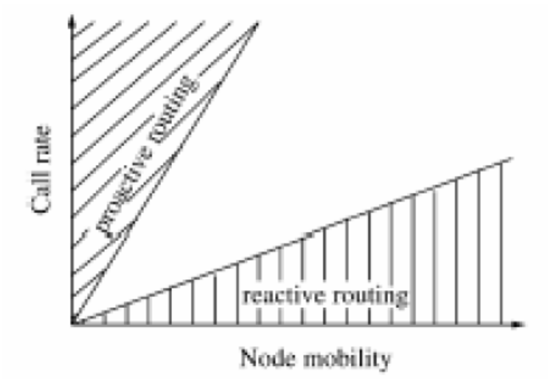


Figure 9: Ad hoc network design space

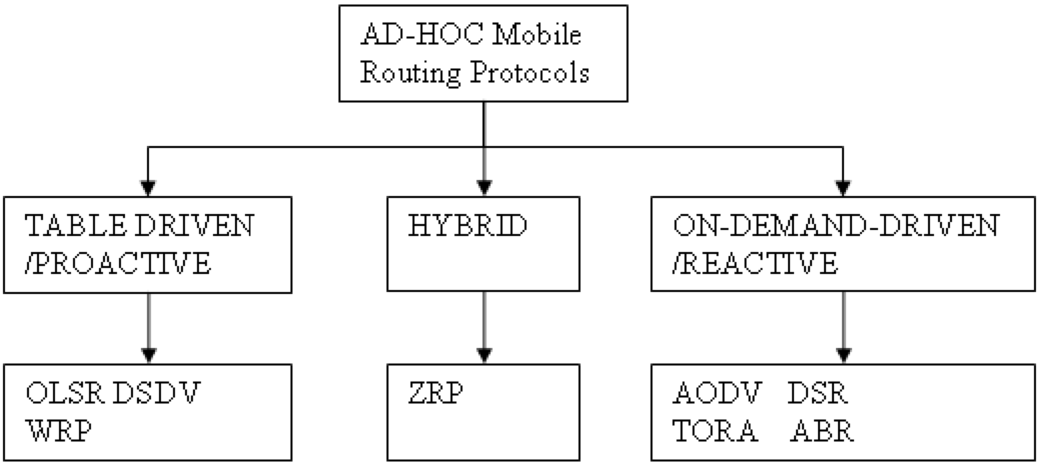


Figure 10: Classification of non-location-based ad-hoc routing protocols

Routing protocols in ad hoc networks vary depending on the type of the network and the decisions we make are depending on our needs and the comparison factors.

2.7.4 Comparison (Metrics)

The comparisons among the protocols are based on some factors and the following metrics are often chosen to compare the different routing protocols:

2.7.4.1 Packet delivery ratio

The packet delivery ratio is defined as the ratio between the number of packets sent by constant bit rate sources (CBR"application layer") and the number of received packets by the CBR sink at destination.

It describes percentage of the packets which reach the destination.

2.7.4.2 Node Density

It is defined as the number of nodes per square unit. In an ideal network the nodes should be evenly distributed throughout the network. But there can be situation when nodes are concentrated in some regions of the network. This can lead to partition of the network and uneven usage of the network bandwidth therefore it is assumed in all the four protocols that the node density is evenly distributed. Partition of the network can lead to situation in which nodes cannot communicate with each other because there are no intermediate nodes. Node density also affects bandwidth because the uneven usage of the bandwidth can affect the throughput of the entire network. Higher node density can lead to more overhead for the proactive routing protocols.

2.7.4.3 Scalability (Number of nodes)

As the number of nodes increases (scalability) the routing protocol should be able to handle the traffic efficiently. Protocols have some limitations affecting their scalability. The problem with scalability is that as the number of nodes increases the routing protocol has to search more nodes to reach the destination, thus increasing the convergence time. As the network size increases the convergence time also increases.

2.7.4.4 Node Mobility

It defines the rate of movement of the nodes. In a network where the nodes move around frequently the node mobility is high. Some protocols are increasing their size of reactive zone as the mobility increases. In case of other protocols as the mobility increases there is more link formation and breakage and this make the stored routing information invalid. Therefore more time and bandwidth is spent in control traffic than in data traffic. In dynamic group routing protocols high node mobility is a problem because link breakages lead to a lot of computation and restructuring of the network. As there is a lot of routing information which is stored in the routing group, frequent mobility makes this information invalid. In VB protocol mobility can be a problem as the local routing information which is stored will become invalid, but the good aspect of VB protocol is that to reconstruct the VB local information is used which is obtained quickly.

2.7.4.5 Call rate frequency

It defines the number of routing queries generated per unit time. A high call rate can be easily handled if the nodes keep some local routing information. If no routing information is saved then the time required to answer a route request query will increase thus degrading the performance of the network. Some protocols save local routing information which helps in answering the route request query, but in the case of others a reactive protocol is used therefore every time a new route request will result, in degraded performance of the network.

2.7.4.6 Routing overhead

The sum of all transmissions overhead of routing packets sent during a session. For packets transmitted over multiple hops, each transmission over one hop, counts as one transmission. Routing overhead is important to compare the scalability of the routing protocols, the adaption to low-bandwidth environments and its efficiency in relation to node battery power (in that sending more routing packets consumes more power). Sending more routing packets also increases the probability of packet collision and can delay data packets in the queues.

In an ideal routing protocol the routing overhead should be to the minimum. The proactive protocols will always produce more overhead than the reactive ones. In the case of reactive protocols we have less overhead because the route request queries are directed to specific parts of the network through other nodes. Using flooding technique to spread the route request query to all parts of the network could affect also the overhead in a network.

2.7.4.7 Traffic load

In the communication between source and destination multiple links are used, these links are also used for communication between different sets of sources and destinations. So there is a possibility that as the usage of the link increases, this leads to a decrease in the link bandwidth. Hence in spite of having an optimal path between source and destination, the data traffic might take more time to reach the destination. Multiple paths can be provided between the source and the destination; the most optimal path is selected, but as the traffic load increases the other routes can also be used thus balancing the traffic load over multiple links.

2.7.4.8 Path Optimality

Optimum path is the best path between the source and destination. In all the routing protocols the main aim is to provide this shortest path, but some of them just provide the first available path rather than the optimal path between the source and destinations. If a proactive protocol is used it can be guaranteed that the path which is obtained is the best path, as the routing information stored by the proactive protocols will be used to determine the best path. In case of a reactive protocol there is no routing information which is stored therefore there is no guarantee that the path obtained is the best path.

2.7.5 Representative Routing Protocols

2.7.5.1 OLSR (Optimized Link State Routing Protocol)

One of the most representative ad hoc proactive routing protocols OLSR is described. OLSR protocol is a proactive routing protocol. Due to its proactive nature, it has a low setup time when a route is asked. In addition, it employs an efficient link state packet forwarding mechanism called multipoint relaying, so this protocol is an optimization of the pure link state protocol. The optimization is achieved by reducing the size of the control packets and by reducing the number of links that are used to forward the link state packets. The reduction of number of links is achieved by announcing only a subset of neighbors of a node that will be used for forwarding the packets for this node. This subset of neighbors which has the responsibility to forward packets is called Multi Point Relays (MPRs). MPR is a key point in OLSR protocol. Each node selects its own MPR set. With this MPR set, the node should cover all the nodes which are two hops away. Link state packets generated by a node will be forward by its MPRs to all its two hop neighbors.

The neighbors that are not the member of MPRs of the considered node can only process the link state packet but cannot forward it (Figure 11). The considered node will calculate the route to all destinations through the members of MPR set. The smaller the MPR set, the more efficient the protocol is, compared with the pure link state protocol.

Similarly, each node has a set of MPR selectors. The MPR selector set includes those who have selected this node as MPR. When the node receives link state packets from its neighbors, it will check whether the originator of this link state packet is in the set of MPR selectors. If yes, link state packets will be forwarded. Otherwise link state packets will only be processed. The member of MPR set and MPR selectors keep changing (updating) over time.

The selection of MPR is one of the most important issues in OLSR, since only the selected node can be the relay point. The requirement for this MPR set is that node, through the neighbors in the MPR set, can read all symmetric strict 2-hop neighbors.

When there is any change detected in the symmetric neighborhood or in a symmetric strict 2-hop neighborhood, MPR set of the node should be recalculated.

Every node periodically originates Topology Control (TC) packets that contain the topology information. This TC information contains the list of neighbors which have selected the sender node as a multipoint relay and will be flooded throughout the network by using the MPR mechanism. The MPRs have the responsibility to broadcast the TC messaged to all the nodes in the network. Nodes will use TCs received from others to calculate routing table. An entry in a routing table includes the MPR selector (the destination) and a last hop node to that destination (the node who originates the TC packet for this destination).

What is more, each node periodically originates Hello message to its neighbors to declare the neighbor nodes that it hears. The MPR set selected by this node will also be included in this Hello message. To use the network more efficiently, data rate should be considered during the selection of MPRs. If nodes with low data rate are selected, there will be higher possibility of overloading at this node. The link with larger data rate should have more probability to be involved in the MPR set.

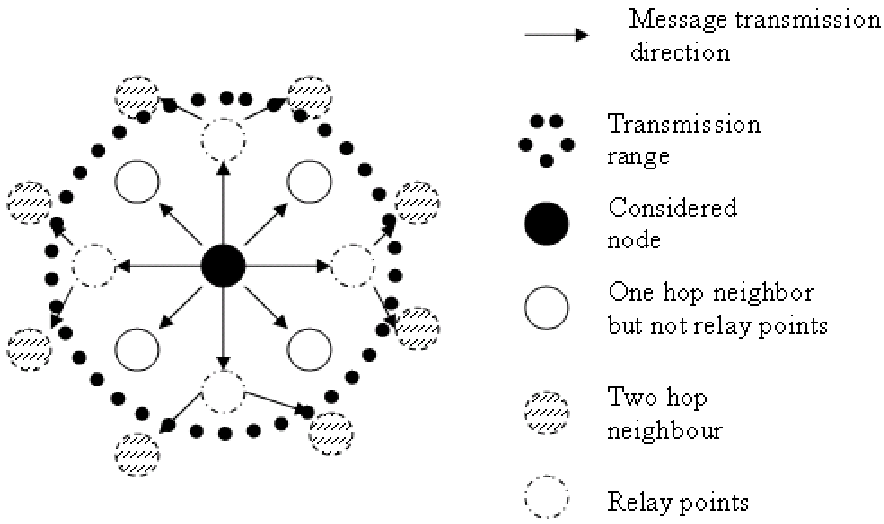


Figure 11: OLSR MPR set

2.7.5.2 Destination Sequenced Distance Vector Protocol

The Destination Sequenced Distance Vector Protocol (DSDV) [Ad-hoc and Hybrid Networks Performance Comparison of MANET Routing Protocols in Ad-hoc and Hybrid Networks] is a proactive, distance vector protocol which uses the Bellman-Ford algorithm. Compared to RIP one more attribute is added to the routing table. The sequence number as new attribute guarantees loop-freedom. It makes it possible for the mobile to distinguish stale routes from new ones and that is how it prevents loops. DSDV can only handle bidirectional links.

The routing table in each node consists of a list of all available nodes, their metric, the next hop to destination and a sequence number generated by the destination node. The routing table is used to transmit packets through the ad hoc network. In order to keep the routing table consistent with the dynamically changing topology of an ad hoc network the nodes have to update the routing table periodically or when there is a significant change in the network. Therefore mobile nodes advertise their routing information by broadcasting a routing table update packet. The metric of an update packet starts with metric one for one-hop neighbors and is incremented by each forwarding node and additionally the original node tags the update packet with a sequence number. The receiving nodes update their routing tables if the sequence number of the update is greater than the current one or it is equal and the metric is smaller than the current metric. Delaying the advertisement of routes until best routes have been found may minimize fluctuations of the routing table. On the other hand the spreading of the routing information has to be frequent and quick enough to guarantee the consistency of the routing tables in a dynamic network. There exist two types of update packets. One is the full dump which contains the entire routing table and must be periodically exchanged. The other is an incremental update which only consists of the information changed since the last full dump.

DSDV responds to broken links by invalidating all routes that contain this link. The routes are immediately assigned an infinite metric and an incremented sequence number. Broken links can be detected by link and physical layer components or if a node receives no broadcast packets from its next neighbors for a while. Then the detecting node broadcasts immediately an update packet and informs the other nodes with it. If the link to a node is up again, the routes will be re-established when the node broadcasts its routing table.

2.7.5.3 Dynamic Source Routing Protocol

As this on demand protocol is called, dynamic source routing protocol uses source routing. It means that the source station knows the whole route to the destination. A complete list of intermediate stations to the destination kept in the header of each data packet.

In route discovery process, when a source node attempt to send packets to a destination, it first checks whether it has a route to this destination in the route cache. If not, the source node will initiate a route request (RREQ) packet for broadcast. Nodes receiving this RREQ will first see whether the destination is itself. If yes, it will reply with a route reply (RREP) packet unicast to the source node with the reverse path the RREQ traversed.

Otherwise this node is an intermediate node. Intermediate nodes who receive this packet should first check the freshness of this RREQ. If the intermediate nodes have received this RREQ recently, it will ignore the packet; or else the intermediate node will rebroadcast this RREQ, another alternative is that the intermediate node replies the RREQ when it knows a route to the destination.

Figure 12 shows the formation of route record as the RREQ broadcasted through the network. Figure 13 shows the unicast of RREP with route record from destination to the source. Source node should be notified using route error (RERR) packets when there is a break on any link on the route which is in use. Source node receiving the RERR will delete all the routes which contain the reported link. A new RREQ will be generated when the route is still needed.

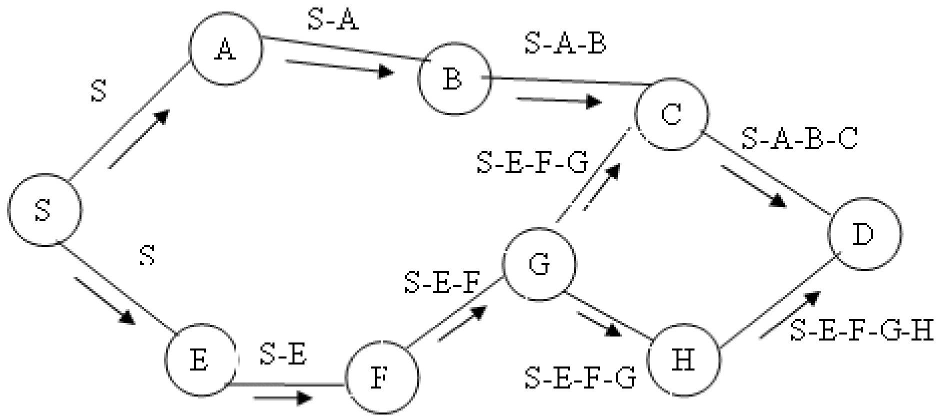


Figure 12: Building of route record during route discovery

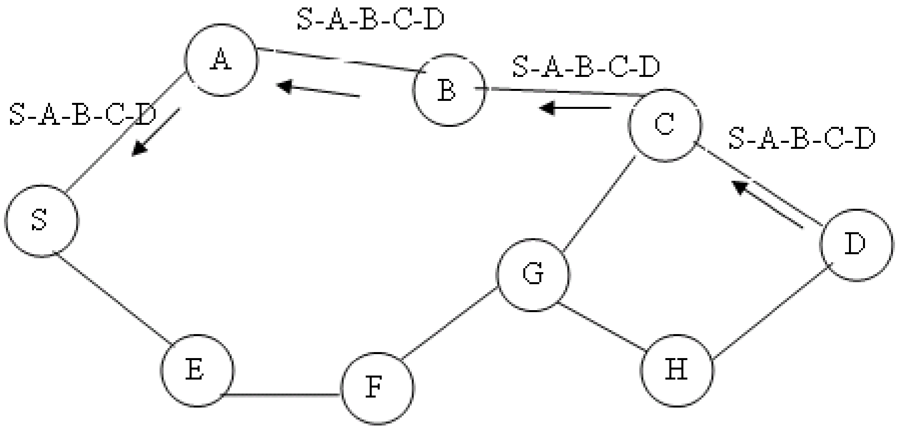


Figure 13: Propagation of RREP with route record

2.7.5.4 Ad Hoc On-Demand Distance Vector Routing Protocol

AODV [Quality of Service Aware Routing Protocols for Mobile Ad Hoc Networks] routing protocol is another on demand routing protocol. Routes are established when they are required. Not like in DSR (Dynamic Source Routing) that stations have all the route information in the routing table, in the routing table of AODV, the station only has the information of the next hop and destination pair.

Route Discovery

1: Generate RREQs: When a source node intends to send packets, it checks its routing table to see whether it has a valid route to that destination. If so, it could begin to send packet to the next hop towards the destination. Or else, it does not have the information about a route to the destination, a RREQ packet is sent as a broadcast message. The route request message includes source identifier (SrcID), the destination identifier (DestID), the source sequence number (SrcSeqNum), the destination sequence number (DestSeqNum), the broadcast identifier (BcastID) and the value of Time to Live (TTL). RREQ is broadcasted to all neighbors of the node. Neighbors are those who have sent Hello message during the last Hello interval.

SrcID and BcastID pair is unique for each route request generated by the source node. The pair should be buffered at the source node for PATH_DISCOVERY_TIME. When the source receives this RREQ packet again, it will not forward it. The TTL in the RREQ is used to control the number of hops this RREQ message could be broadcasted. The mechanism used in AODV for control RREQ broadcast range is called expanding ring search technique.

For the first time of RREQ broadcast, the value of TTL at the source node is set to TTL_START. When no reply is received within the discovery period, another RREQ is send with TTL equal to the TTL_START +TTL_INCREMENT. If there is still no reply got, this process will be repeated again and again until TTL reaches the value of TTL_THRESHOLD. When the incremented value TTL is equal or larger than TTL_THRESHOLD, the value of TTL will be set to the value of NETWORK_DIAMETER. RREQ will be broadcasted at a maximum of RREQ_RETRIES times with the value of NETWORK_DIAMETER. If there is still no reply got, all the packets that destine for this destination will be dropped. The advantage of this technique is to control the flooding of RREQ in the whole network. This will help to save more data rate especially when the network is large.

2: Nodes receive RREQs

Nodes receiving RREQ will first check whether it is the source that sends this RREQ or whether the node has recently heard (during the last PATH_DISCOVERY_TIME) this RREQ. If one of these holds true, the newer RREQ will be dropped silently to prevent the flood of RREQ in the network. RREQ is identified by the unique SrcID and BcastID pair which will store in the nodes for PATH_DISCOVERY_TIME after nodes first received this RREQ. If the node receiving the RREQ has not heard this previously and it is not originated by itself, it will be preceded by this node further on which will be described in the following paragraphs. The node will set up a reverse route entry for the source node in its routing table. By storing this reverse route in the routing table, the node knows how to forward the RREP to the source if it received a RREP for this source node at a later time. If the entry for this source node already exists, this entry will be updated by refreshing the sequence number of the record. The sequence number is decided by choosing a larger one between the sequence number in RREQ and the one in the already existing route record. Lifetime of this route entry should be updated. When this route entry is not used until the end of lifetime, it will be deleted from the routing table.

After processing with the reverse path, the node checks whether it is the destination node. If the node is an intermediate node and it may rebroadcast RREQ when it does not have a fresh enough route. In this case (step 4): Intermediate node forwards RREQ. Otherwise, if the intermediate node has a fresh enough route and it could send RREP (D in RREQ is set to 0), (step 3): Intermediate node sends RREP. If the node is the destination node (step 4): Destination node sends RREP.

3: Intermediate node sends RREP

Intermediate node could send RREP to the source node when two of the following conditions are satisfied. The first one is that the node has a fresh enough route to the destination. The second condition is the bit D in RREQ is not set to 0.

For the first condition, the DestSeqNum of the intermediate node will be compared with the corresponding one in RREQ message. The route is valid only when DestSeqNum of the intermediate node is bigger than the corresponding one in RREQ message. The DestSeqNum is used to help find the latest route.

The bit D in RREQ is a destination only flag. It tells whether the intermediate node could reply the RREP. When this is set to 0, intermediate node could send RREP. Otherwise, intermediate node must not send RREQ and should propagate the route request further by adding its own address in the RREQ. The reason why D is needed is explained as follows. There is one possibility that all the RREQs are replied by intermediate nodes and no RREQ can reach the destination node. When the user is expecting a bi-directional link, the destination node has to initiate another route discovery and it costs more network resource. If D is set to 1, it can be ensured that the RREQ could reach destination when there is a route from source to the destination. In addition, the RREP gotten from the destination is always as fresh as or fresher than the RREP gotten from the intermediate node.

Another alternative to ensure destination know how to get to source is to generate gratuitous RREQ.A bit called G which is a gratuitous RREP flag can be set in the RREQ. When this bit is 1, intermediate node which generates RREP to the source node will at the same time unicast a RREP to the destination node. In this way, destination node knows the route to source.

4: Intermediate node forwards RREQ

Intermediate node forwards the RREQ packet when it does not have fresh enough routes to the destination or D is set to 1. Before forwarding, the value of TTL is examined to see whether the RREQ could be rebroadcasted further. If the value of TTL stored in the receiving RREQ is 1, the RREQ is not going to be forwarded any further and will be discarded. Otherwise, the RREQ is broadcasted to all its neighbors who send Hello messages during the last Hello interval.

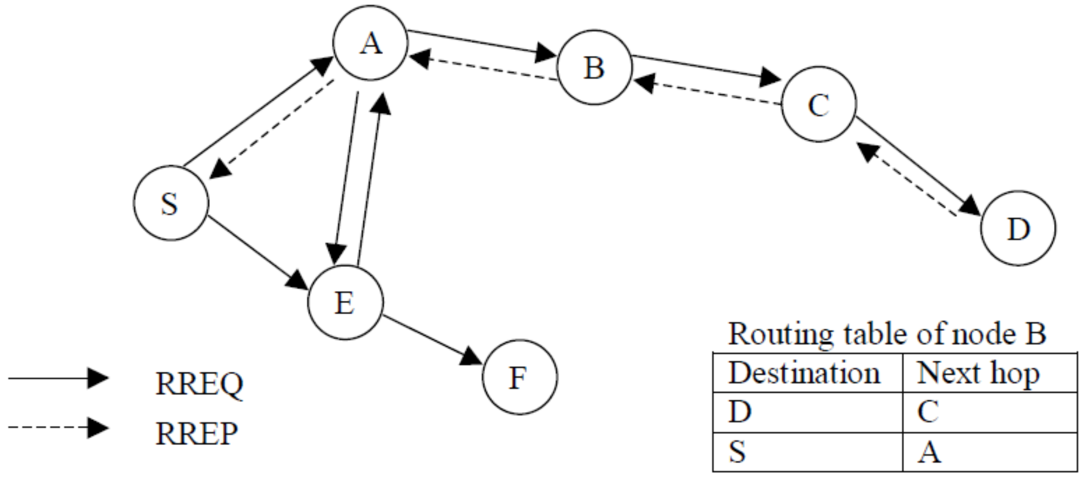
5: Destination node sends RREP

The destination node sends a RREP when it receives a RREQ. The highest Sequence No. of the destination node is increased by 1 firstly. In the RREP message, the Destination Sequence Number field is fill by own highest Sequence No. of the destination node. Lifetime of this route is set to MY_ROUTE_TIMEOUT. The RREP is unicasted to the next hop towards the source of the RREQ. The hop count is incremented by one at each hop. In this way, when the RREP arrive the source, this hop count field is the distance (in hops) from the source to the destination.

6: Receive RREP

Nodes receive RREP have the responsibility to forward RREP along the path to the source node.

Besides route discovery process, the already set up routes also need to be repaired when the promised provided route cannot be guaranteed. When there is disconnection of a link on the route (some transient node may move out of range) the neighbor nodes will notice the absence of this connection. If so, the neighbor nodes will check if there is any route in its routing table which the next hop is this disconnected neighbor. If there is any, all the sources that send traffic going through this disconnected node will be noticed by sending RERRs. A new route request will be generated by the source node if there is still need of transmission from sources. For example, node S intends to find a route to node D, the process is shown in Figure 14.



Destination	Next hop
D	C
S	A

Figure 14: AODV route discovery

2.7.5.5 Comparison between DSR and AODV routing protocols

There are a few differences between DSR and AODV routing protocols. Firstly, compared with DSR, the source node of AODV only knows the route to the destination, but in DSR, source node knows the route to intermediate node also. On the other hand, because in DSR, each data packet has to take the whole route information in the header, it costs large overhead which will waste data rate. Another difference between DSR and AODV is the usage of timer. In the DSR protocol, there is no timer used for the validation of routes. Stale routes could be used for routing. In AODV, timer is used for the freshness of a route. In DSR, with stale route, it is possible that the route is not validated. It will cause the loss the packets before source node is notified that the route is invalid. On the other hand, if the route is still valid, route overhead is saved for route discovery process. In AODV, route which has not been used for a period of expire time will be deleted. The set of expire time is important since short expire time may lead to the deletion of still valid route and long expire time will give more non-fresh routes. Thirdly, it is about RREP. The destination of DSR could send RREP more than once for the same RREQ without considering whether the same RREQ had been replied just a second ago for other routes. In addition, multiple route entry could be store in the route table for the same destination in the same source node. That is, a source node has multiple route entries for one destination. While, in AODV, only one route reply can be sent by the destination and only one route entry per destination is stored in the route table. The advantage of having more than one route entry per destination is that it can provide backup routes when there is a break on the current route. It means that there is no need to send a RREQ again to

search for the route. On the other side, if the network has high mobility, the freshness of the backup route could be a problem and reinitiate a new RREQ should be a better choice.
In the [Quality of Service Aware Routing Protocols for Mobile Ad Hoc Networks] analysis, it is found that AODV protocol should perform better in high mobility ad hoc network.

2.7.5.6 Hybrid routing protocols ZRP, DVB

Zone Routing Protocol (ZRP) is a hybrid routing protocol. It effectively combines the advantages of both proactive and reactive routing protocols. The key concept used in this protocol is to use a proactive routing within a zone in the r-hop neighborhood of every node and use a reactive routing for nodes outside this zone. The table driven scope is limited within a zone and when a destination is out of the table driven scope, on demand routing search is initiated. In this situation, control overhead is reduced, compared to both the route request flooding mechanism employed in on demand protocols and periodic flooding of routing information packet in table driven protocol.
In [DYNAMIC VIRTUAL BACKBONE ROUTING ROTOCOL: A HYBRID ROUTING PROTOCOL FOR ADHOC NETWORKS] the DVB hybrid protocol have been selected for the reasons:

- The convergence time for the reactive component of the protocol is the least compared to the other protocols considered
- The path obtained is the shortest
- The protocol can be extended to take care of traffic load balancing
- Local information is used to reconfigure/reconstruct the VB therefore it's a faster process.
- As there is no broadcasting method used in this protocol, it is more reliable
- Simpler representation of the whole network because a few VB nodes can represent the whole network
- Advantageous for the network with high call rates.

Dynamic virtual backbone routing method is a hierarchical routing protocol which is hybrid in nature. Virtual backbone routing (VBR/DVB) is influenced by the Zone routing protocol (ZRP) and hierarchical routing architecture. Hierarchical routing is a perfect example of selective representation. In this routing method nodes are grouped into super nodes based on some predefined metric, these super nodes are further grouped into super-super nodes and so on. The only way the nodes at lower level have information about faraway nodes is through the higher level nodes.

Selective representation of the network topology is advantageous as it reduces the control overhead of the routing protocol; because each node need not maintain the routing information about all the other nodes in the network, only nodes which are in its routing zone. Another advantage of selective representation is that the route query messages can be directed to different region of the network, instead of flooding throughout the network. Since VBR is a hybrid routing protocol it has two routing components, proactive component and reactive component. The proactive component is similar to the zone routing protocol, in which a routing zone of radius 'r' is defined around each node and each node has routing information about each other nodes in its routing zone, along with this functionality the proactive component in VBR is also responsible for the construction, maintenance and connectivity of the VB's. The reactive component of VBR is applied if the source does not find the destination in its local routing zone, in this case the route query is send to the nearest VB, which tries to resolve the query by looking in its routing zone, if the destination is still not found then the VB broadcasts the route query to all the VB through the multi-hop paths connecting the VB's. Some interesting facts about the VB nodes are that any node in the network can serve as a VB during its time as part of the network. The VB nodes are connected to each other through virtual links which span multiple regular nodes. VB nodes are also called as "database", because during the reactive routing phase the source nodes query the VB nearest to them. The VB are selected in such a way that all the nodes in the network are connected to at least one VB, in other words the VB represents the entire network.

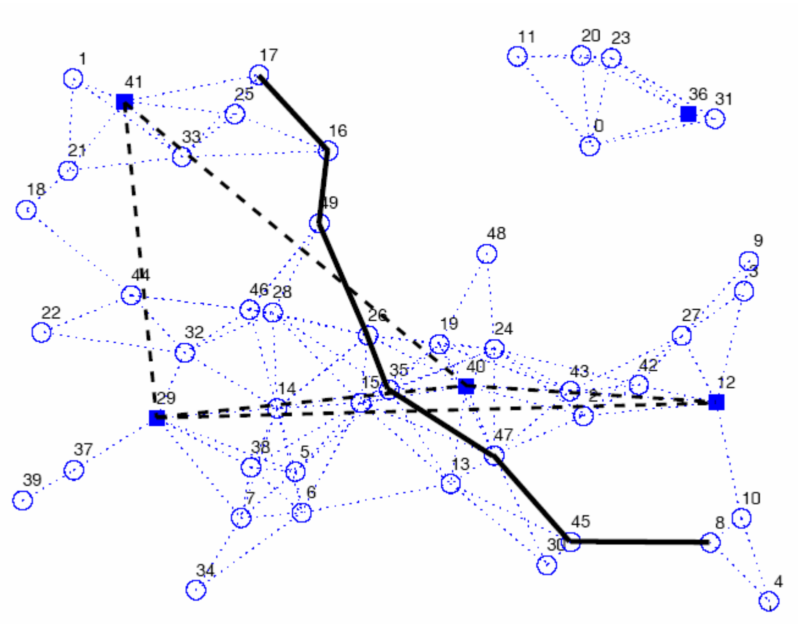


Figure 15: VBR with 50 nodes and r = 2. Circles: regular node; solid squares: VB nodes/databases; thin dotted lines: radio links; dashed lines: multi-hop virtual links; solid line: the actual data route between two nodes.

The advantage of this scheme is that the reactive route queries are always directed to certain location in the network, instead of flooding them throughout the network. Hence the overall control traffic in the network is reduced. Even though it’s a hierarchical protocol, the potential occurrence of congestion (hot-spots) is less as the routing paths need not always pass through the VB nodes. Node mobility might be a problem for this because the process of re-generation of the VB and stabilization of the network (means each node associating itself with at least one database) are time consuming operations and as mobility increases, more of these operation have to be performed very often. This approach provides multiple paths to the destination which can lead to proper load balancing. The reactive part involves a lot of control traffic transmissions.

2.7.6 Other Routing Strategies

2.7.6.1 Power-aware routing protocols

Power aware routing protocols aim at minimizing the consumption of an important resource in ad hoc network: battery power. To achieve that, the routing decisions are made based on the consumption of power. That is during the design of routing protocols, route cost less power will have more possibility to be chosen. Energy needed to transmit a signal is proportional to the square of the distance between transmitter and receiver. Transmitting a signal with half of the distance requires one fourth of the energy and if there is a node in the middle willing spend another fourth of its energy for the second half distance, data would be transmitted for half of the energy compared with the one going through a direct transmission. This however may introduce a delay, but this is the basic way that energy is saved. That is, to take energy consumption as the most important path selection metric.

Power saving and maximizing the use of power in order to transmit more traffic on the network are the aim.

2.7.6.2 Load-Aware Routing Protocols

In load aware routing protocols, load information is utilized as the path selection metric for routing in MANETs. There are multiple schemes which are load-aware, and the ways of calculating the load at nodes are different.

The first one is Dynamic Load-Aware Routing (DLAR) protocol. In DLAR, network load is defined as the number of ongoing traffic in its interface queue. The routing decisions will be made by comparing the queuing length of the node. That is, only the delay that caused by queuing is considered. In Load-Balanced Ad hoc Routing (LBAR) protocol, the network load is defined as the total number of routes passing through the node and its neighbors.

The Load-Sensitive Routing (LSR) protocol defines network load in a node as the summation of the number of packets being queued in the interface of the mobile host and its neighboring hosts. LSR protocol is more accurate than DLAR and LBAR since it considers both the traffic queuing at the current considering node and the traffic condition at the neighbor node, whereas it does not consider the effect of access contentions in the MAC layer (using IEEE 802.11). For example, we consider Node A and Node B each of which has one queued packet. Node A has three neighbors, and number of packet queued in each neighbors is one. Node B has only one neighbor having 3 queued packets. The total load for Node A and Node B are equal if calculated using LSR protocol, but Node A will have higher potential contention delay than Node B. By taking the contention delay into account, Delay-Oriented Shortest Path Routing (DOSPR) is proposed by analyzing the medium access delay in IEEE 802.11 wireless network.

The last load aware routing protocol discussed is Delay-based Load aware on demand routing (D-LAOR) protocol utilizes both the estimated total path delay and hop count as the route selection criterion. A time stamp is placed on each packet header. The average total node delay is calculated from the previous average total node delay and successful transmission delay of this time at this node. This averaged total delay includes the queuing, contention and transmission delays. The total path delay will be the summation of average total delay of the nodes that it went through. This D-LOAR is implemented on AODV protocol with some modifications during the path discovery. If the duplicated RREQ packet has a smaller total path delay and hop count than the previous one, the destination will send RREP packet again to the source node to change the route immediately, since the route with smaller delay should be chosen. In conclusion, nodes with smaller load which might cause lower delay are preferred to be chosen and this method also balances the load on the network.

2.7.6.3 QoS-aware routing protocols in Ad Hoc networks

Quality of Service [Quality of Service Aware Routing Protocols for Mobile Ad Hoc Networks] is the performance level of a service offered by the network to the user. In the originally used network model, traffic is transmitted only with best-effort. It means that there is no quality guarantee for each transmission. When with the real-time traffic is transmitted in the network, QoS becomes demanding. In addition, because of the limitation of network resources especially in wireless networks, real time traffic need to be given higher priority to ensure that the real time traffic arrive the destination on time.

QoS parameters differ from application to application. For example, for multimedia applications, the data rate and delay are the key factors, whereas, in military use, security and reliability become more important. If considering QoS required by emergency cases such as rescue, the key factor should be the availability. In sensor networks, battery life and energy conservation would be the prime QoS parameters. The QoS parameter considered here is aimed to real time applications. In real time applications, QoS requests can be expressed in term of many metrics in routing protocols. The most popular metrics are data rate and delay. To satisfy QoS requirements, the corresponding available data rate and delay that could be provided by the network of each route should be calculated in order to see which route could be used with satisfying QoS. QoS of a network can be considered at different layers. QoS considered in physical layer means the quality in terms of transmission performance. For example, through transmission power control both the stations that are near the sender or far away from the sender could hear the signal clearly with different transmission power. Power control is used both to ensure the quality of reception and to optimize the capacity.

QoS implemented in MAC layer is also important. It could provide high probability of access with low delay when stations with higher user priority want to access the wireless medium.

QoS implemented in the routing layer aims to find a route which provides the required quality. The metric which helps to choose the route is not only the number of needed hops along the route but also some other metrics like maximum delay and minimum data rate.

The existing QoS models can be classified into two types based on their fundamental operations. The QoS models are Integrated Service (IntServ) and Differentiated Services (DiffServ).

IntServ is a fine grained approach which provides QoS to individual applications or flows. It uses Resource Reservation Protocol (RSVP) to provide a circuit switched service in packet switched network. IntServ decides whether the desired service could be provided with the current available network resource. Admission control is performed to new flows. The admission of each new flow might cause interference to the already existing flows. One of the responsibilities of admission control is that the interference caused by adding a new flow should not make QoS of old flows get worse than it has required. The drawback of IntServ is the scalability problem. This is caused by the need of storing every flow state in the routes.

DiffServ provides QoS to large classes of data or aggregated traffic. It is a coarse grained approach. It maps flows into a set of service levels. In DiffServ, routers are divided into two types: edge routers and core routers. In edge routers, traffic will be classified, conditioned and assigned to different behavior aggregate when it traverse between different networks. The word different networks means for example networks that belong to different Internet Service Providers (ISP). Differentiated Services Code Points (DSCP) bits are reformatted which represent the Type of Service (ToS) in Internet Protocol (IP) header. Core routers forward packets based on this ToS field. In addition, core routes also need to follow the per-hope behavior (PHB) which takes charge of scheduling of packets. IntServ eliminated the need of keeping the flow state information somewhere in the network.

Mobile ad hoc networks differ from the traditional wired networks. They have certain unique characteristics which cause difficulties for providing QoS in such networks (dynamically varying network topology, lack of precise state information, shared radio channel, limited resource availability, hidden terminal problem and insecure medium).

The QoS metrics can be classified into three categories. They are additive metrics, concave metrics, and multiplicative metrics. Additive metrics is defined as sum of the value of the metric on all links along the path. Delay and jitter are additive metrics. Delay along the path is the sum of the delay at every link along the path. A concave metric means the minimum metric value over a path. Metric value on every link along the path is taken into account. The minimum metric value stands for the metric value of the whole path. Data rate is a concave metric. The minimum data rate of all the links along the path should be the data rate of this route. A multiplicative metric represents the product of the metric values on all links over a path. The criteria of reliability or availability of one link, e.g. link outage probability is a multiplicative metric.

Quality of Service for Ad hoc On-Demand Distance Vector Routing

The AODV routing protocol is intended to be used by mobile stations in ad hoc networks.

Considering QoS, extensions can be added to the messages used during the route discovery process. When QoS extensions are added to the AODV routing protocol, four items are added to each routing table. They are session ID, maximum delay, minimum available data rate, list of sources requesting delay guarantees, list of sources requesting data rate guarantees.

In AODV, routes are selected per destination, but it is not the case in a QoS based AODV protocol. Session ID is used to specify each flow, because two flows aimed at the same destination might have different QoS requirement.

This extension field maximum delay is used both in RREQ and RREP messages but with some differences. The maximum delay used in RREQ indicates the maximum time allowed to be used for a transmission from the current node to the destination. When a node receives RREQ, the maximum delay will be compared with the node traversal time (should include queuing delays, interrupt processing times and transfer times). If the maximum allowable delay is smaller than the node traversal time, the RREQ will not be broadcasted further, otherwise, the RREQ will be broadcasted and before the broadcasting, traversal time of this node must be subtracted from the maximum delay field in RREQ message. In RREP message, the maximum delay extension field will be filled with zeroes, and delays are cumulative from the destination along the route that the RREP come from to the source node, and the source node will choose the one satisfying the delay requirement.

When the node receives a RREQ, it will compare the available data rate with the one required in the minimum data rate extension field and only the one that satisfies certain condition could forward the RREQ further.

A QoS_Lost message is generated when the QoS could not be satisfied (node traversal time increase or available data rate decrease. It needs a list of sources whose traffic is affected by this change.

Comparing with AODV, in QoS based AODV, we have RREQ, including the QoS requirement, and only the node satisfying this request could broadcast the RREQ further.

In addition, the intermediate node cannot give the RREP even if it knows how to get to the destination, since data rate should be tested at each hop until to the destination.

What’s more, the network should have periodic mechanisms for checking the available data rate at each node. Because of the mobility of the node, the available data rate of one node is changing with the changing of its neighbors who are sharing data rate with it. If the available data rate of a node becomes smaller than zero, it means that the node cannot afford the current traffic any more. Source of the traffic going through this node should be noticed.

2.7.6.4 Position-based routing schemes

Location-based routing approaches eliminate some of the limitations of topology based routing through obtaining the node’s geographic locations via the GPS or some other types of positioning service. Recent research shows that location based routing can significantly reduce communication overhead compared with topology based methods. Location-based routing protocols are likely to be more scalable, since each node only needs to know the location of the destination and its neighbors' locations to make a forwarding decision.

There are two important issues in geographic ad hoc routing protocols. One is the existence of scalable location services, which manage the position information of any mobile node at any time throughout the entire network. The current position of a specific node can be learned in a deterministic manner with the help of a position service such as GPS. Each mobile node registers its current position with a location service. Another is the packet-forwarding strategies, via which a node makes the forwarding decision based on the position of a packet's destination and the positions of the node's immediate one hop neighbors. When a node does not have the position of its communication partner, it requires such information from a location service, and then embeds the position of destination into the header of data packets. However, the positions of the neighbors are typically learned through one-hop broadcasts. These beacons are broadcasted periodically by all mobile nodes and contain the position of the sending node. Then, a node forwards data packets to the destination node via packet-forwarding strategies. Examples of location-based routing protocols include DREAM, LAR, PSR, GLS and DLM. In fact, to design a scalable location management scheme is difficult because it would represent a functional deadlock problem; it is impossible to obtain the position information of a specific node without a location server, and without position information, a node cannot reach the location server.

We can distinguish three main packet-forwarding strategies for location-based routing, which are greedy forwarding, restricted directional flooding, and hierarchical approaches. In greedy forwarding, a node forwards a given packet to one of its neighbors that are located closer to the destination than the forwarding node itself. Generally, there are different optimization criteria a node can use to decide to which neighbor a given packet should be forwarded, such as most forward within transmission range (MFR), nearest with forward progress (NFP), and compass routing. In restricted directional flooding, a node forwards a given packet to some not only one one-hop neighbors. Unfortunately, both forwarding strategies may fail if there is no one-hop neighbor that is closer to the destination than the forwarding node itself. In order to cope with this kind of failure, recovery strategies have been proposed that the packet should be forwarded to the node with the least backward (negative) progress if no nodes can be found in the forward directions. Examples of recovery strategies are the face-2 algorithm and the Greedy Perimeter Stateless Routing Protocol (GPSR). Hierarchical approaches are promising methods to enhance the scalability of traditional networks by establishing some form of hierarchy. Hierarchical forwarding strategies use greedy forwarding for wide area routing and non-position-based approaches for local area routing.

At the top level, location services can be divided into flooding based and rendezvous-based approaches. Flooding-based protocols can be further divided into proactive and reactive approaches. In the proactive flooding-based approach, each (destination) node periodically floods its location to other nodes in the network each of which maintains a location table recording the most recent locations of other nodes. The interval and range of such flooding can be optimized according to the node’s mobility and the distance effect. For example, flooding should be more frequent for nodes with higher mobility, and flooding to faraway nodes can be less frequent than to nearby nodes. DREAM serves as a good example of proactive flooding-based location services. In reactive flooding-based approaches, if a node cannot find a recent location of a destination to which it is trying to send data packets, it floods a scoped query in the network in search of the destination.

In rendezvous-based protocols, all nodes (potential senders or receivers) in the network agree upon a mapping that maps each node’s unique identifier to one or more other nodes in the network. The mapped-to nodes are the location servers for that node. They will be the rendezvous nodes where periodical location updates will be stored and location queries will be looked up. Two different approaches of performing the mapping, quorum-based and hashing-based, have been proposed. In the quorum-based approach , each location update of a node is sent to a subset (update quorum) of available nodes, and a location query for that node is sent to a potentially different subset (query quorum). The two subsets are designed such that their intersection is non-empty, and thus the query will be satisfied by some node in the update quorum.

In hashing-based protocols, location servers are chosen via a hashing function, either in the node identifier space or in the location space. Hashing-based protocols can be further divided into hierarchical or flat, depending on whether a hierarchy of recursively defined subareas is used. In hierarchical hashing based protocols, the area in which nodes reside is recursively divided into a hierarchy of smaller and smaller grids. For each node, one or more nodes in each grid at each level of the hierarchy are chosen as its location servers. Location updates and queries traverse up and down the hierarchy. A major benefit of maintaining a hierarchy is that when the source and destination nodes are nearby, the query traversal is limited to the lower levels of the hierarchy.

The best example of hierarchical rendezvous based location service is GLS.

In flat hashing-based protocols, a well-known hash function is used to map each node's identifier to a home region consisting of one or more nodes within a fixed location in the network area. All nodes in the home region maintain the location information for the node and can reply to queries for that node; they serve as its location servers.

In GLS, each mobile node maintains its location information in a number of location servers in each sibling region with IDs closest to its own ID. The distribution density of a node's location servers is not even. Thus, it has the advantage that the distances traversed by a location query are proportional to data packet path lengths between destination nodes and source nodes. However, when nodes move arbitrarily, GLS becomes less efficient. Because in selection of location servers, a node has to possibly scan the entire region to obtain the one with the closest ID as the location server, and in location maintenance, with nodes entry or departure in a particular region, the scheme needs to check the entire particular region periodically and change the choices of location servers. All of these incur heavy signaling overhead. Another protocol proposed is Distributed Location Management (DLM). Within this scheme, a hash function is used to map a node's ID directly to a particular minimum partition; all nodes in this minimum partition are selected as the node's location server, which reduces the signaling overhead without the process of scanning the entire region compared with GLS. DLM introduces hierarchical addressing models based on logical network partitions. Positions of nodes can have different accuracy levels. Consequently, only a small set of location servers need to be updated when a node moves while different location servers may carry position information of different levels of accuracy. However, location servers distribute evenly throughout the whole network. The distance traversed by a location query is independent of data packet path length. If there are some void minimum partitions, DLM requires location query forwarding and hash function computation several times. It costs the available bandwidth of MANETs. Another protocol proposed and tested in [TOPOLOGY CONTROL, ROUTING PROTOCOLS AND PERFORMANCE EVALUATION FOR MOBILE WIRELESS AD HOC NETWORKS] is a novel location management service for mobile ad hoc networks, called Enhanced Scalable Location management Service (EnSLS), which combines the advantages of GLS and DLM. Firstly, EnSLS maintains the property: long distance queries are proportionally penalized. EnSLS maps a nodes unique ID to the central coordination of a specific minimum region via a hash function. It takes the GPSR as the underlying packet forwarding protocol, which allows that all packets destined for an arbitrary location (possibly unoccupied by a node) to be forwarded consistently to the same node in the neighborhood of that location. Then the node receiving the location maintenance packet acts as one of the location servers. In the process of location server's selection and update, it only requires control packet forwarding and hash function computation once. All of these can significantly reduce the end-to-end delays and total signaling traffic needed to manage the location information. Although each protocol is proposed along with some theoretical and/or simulation analysis, little is known about the relative performance of these protocols, especially in a realistic setting, i.e., a mobile environment.

2.7.6.5 Multi-Path Routing

Multipath routing has been explored in several different contexts [Multipath routing in mobile ad hoc networks: issues and challenges, Performance tools and applications to networked systems]. Traditional circuit switched telephone networks used a type of multipath routing called alternate path routing where each source node and destination node have a set of paths (or multipath) which consist of a primary path and one or more alternate paths. Alternate path routing was proposed in order to decrease the call blocking probability and increase overall network utilization. In alternate path routing, the shortest path between exchanges is typically one hop across the backbone network; the network core consists of a fully connected set of switches. When the shortest path for a particular source destination pair becomes unavailable (due to either link failure or full capacity), rather than blocking a connection, an alternate path, which is typically two hops, is used. Well known alternate path routing schemes is Dynamic Non hierarchical Routing and Dynamic Alternative Routing.

Multipath routing has also been addressed in data networks which are intended to support connection-oriented service with QoS. For instance, Asynchronous Transfer Mode (ATM) networks use a signaling protocol, Private Network-Node Interface (PNNI), to set up multiple paths between a source node and a destination node. The primary (or optimal) path is used until it either fails or becomes over-utilized, and then alternate paths are tried. Using a crank back process, the alternate routes are attempted until a connection is completed.

Alternate or multipath routing has typically lent itself to be of more obvious use to connection-oriented networks; call blocking probability is only relevant to connection oriented networks. However, in packet-oriented networks, like the Internet, multipath routing could be used to alleviate congestion by routing packets from highly utilized links to links which are less highly utilized. The drawback of this approach is that the cost of storing extra routes at each router usually precludes the use of multipath routing. However, multipath routing techniques have been proposed for Open Shortest Path First (OSPF), a widely used Internet routing protocol.

Multipath Routing for MANET

Most of MANET protocols use a single path. The topology of MANET may change rapidly and unpredictably over time, which can cause the single path to be easily broken. When the single path fails, these protocols need to process a potentially costly route discovery to locate an alternate route for the given destination. Therefore, single path routing protocols directly expose nodes to long network delays and excessive overhead when the path fails. Single path routing has not considered the load balancing problems. An unbalanced assignment of data traffic leads to power deletion on heavily loaded hosts. With more hosts powered down, the connectivity of the network will be reduced, which lead to the failure of communications due to network partition. Multi-path reactive protocols are designed to alleviate these problems by selecting a set of redundant paths between the source and the destination in a single route discovery attempt. It has been proven that the use of multiple paths could keep correct end-to-end transmission for a longer time than a single path. In other words, a new routing discovery process is required only when all paths fail, thus, the frequency of searching for new routes is much lower if a node maintains multiple paths to the destination. The lower frequency of the routing rediscovery process means lower route discovery latency and less routing overhead.

The multipath routing has a lot of advantages. Firstly, it promises load balancing. Multipath routing distributes traffic over multiple routes, which can alleviate congestion. Secondly, multipath routing provides fault-tolerance; when a link breaks, alternative routes still can be used to route the packets. A node disjoint route has a higher degree of fault-tolerance than a link disjoint and non-disjoint route, but it does not ensure transmission independence. Then, multipath routing has higher aggregate bandwidth: multiple paths can be used simultaneously to route data packets.

However, current multi-path protocols choose a set of multiple redundant paths through a single route selection parameter without considering the interplays of different selection parameters. For example, Split Multi-path Routing protocol (SMR) uses the number of intermediate hops as the selection parameter. SMR chooses the shortest routing path as the primary route, and then computes the maximum disjointed path as the secondary route. The Stability-Based Multi-path Routing algorithm (SBMR) also locates a set of totally independent multiple paths via the hops of the routes. Due to the special characteristics of MANET, the shortest path decided by the number of intermediate hops is not necessarily the most reliable or durable path since the topology of MANET is determined by many factors such as battery capacity, traffic pattern, link stability and nodal mobility. All of these factors are correlated. Thus, consideration of only one or two factors is not sufficient for choosing an optimal set of multiple paths. A multiple path selection protocol proposed by proposing and tested is Genetic Fuzzy Multi-path Routing Protocol (GFMRP), which considers the multiple correlated selection parameters, based on fuzzy set theory and evolutionary computing. The goal of GFMRP is for selecting a set of paths to maximize network lifetime and reliability. During routing discovery phase, the fuzzy inference system in the destination node takes four selection parameters (1) packet buffer occupancy rate at a node, (2) energy consumption rate at a node, (3) link stability between neighboring nodes, and (4) the number of intermediate hops in a route) (the information of these parameters are collected through RREQ packet) as input and outputs a crisp rank to evaluate every potential route path between the source node and the destination node. Then the route paths are classified into two categories according to their relative ranks. A path having a rank value above a certain threshold, α, belongs to a reliable set of multiple paths. Otherwise it belongs to an unreliable set. The routing paths in the reliable set are used by GFMRP as candidates for multi-path routing. GFMRP employs a simple load balancing method to distribute the traffic load, by spreading data packets over all the paths of the reliable multiple paths set in a round-robin fashion.

CHAPTER 3

3. Mars Surface Exploration Ad hoc network

3.1 The Exploration of Mars

Since 1960, the Russian and American space agencies have sent many spacecraft to Mars. Some have been a great success while others didn't even make it into space.
Mariner 4 was the first mission to make it successfully to Mars. It arrived in 1965. The Mariner 4, 6, 7, and 9 missions took tons of pictures of the Red Planet and its moons. Between 1971 and 1973, the USSR sent Mars 2, 3, 4, 5, and 6 to Mars. With varying success, these orbiters and landers sent back data about the Martian atmosphere, surface, gravity, and temperature. In the mid 70's, the Viking missions were very successful. In all, these orbiters and landers returned over 56,400 images of the Martian surface. After a quiet decade, Mars exploration took off again with the Mars Observer mission launched in 1992. Unfortunately, this spacecraft was lost due to explosion. NASA built upon this experience and created the successful Mars Surveyor Program which included the Mars Pathfinder and Mars Global Surveyor missions. In 1998, Japan also joined in Mars exploration with the Nozomi spacecraft. On April 7, 2001, the 2001 Mars Odyssey was launched. The 2001 Odyssey has been collecting data on what chemicals and minerals make up the Martian surface. It is also providing information about possible radiation hazards for future human explorers. The European Space Agency's Mars Express mission, including the Beagle 2 lander, arrived at Mars in December 2003. NASA launched twin Mars Exploration Rovers (MER) in June and July of 2003. They landed on the Red Planet in January 2004. Although they were only supposed to keep working for 90 days on the Martian surface, both are still going strong four years later NASA's Mars Reconnaissance Orbiter (MRO) was launched in August 2005 and went into orbit around Mars in March 2006. It is making very detailed maps of the Red Planet. Those maps will help scientists select landing sites for other future missions. The Phoenix Mars Lander was launched in August 2007. It landed near the North Pole of Mars in May 2008. Phoenix will use a robotic arm to dig up samples of Martian soil. Scientists hope Phoenix will find water ice just below the surface. About future missions, another rover, NASA's Mars Science Laboratory, is supposed to blast off in the fall of 2009. That rover will be even bigger than the Mars Exploration Rovers that have done such a great job. In the future we may also send airplanes or balloons to roam the Martian skies. We can send a drilling rig to search for water and maybe signs of life underground. Scientists also hope a there will be

sample return mission some day. It would bring Mars rocks back to Earth so scientists could study them in laboratories on Earth.

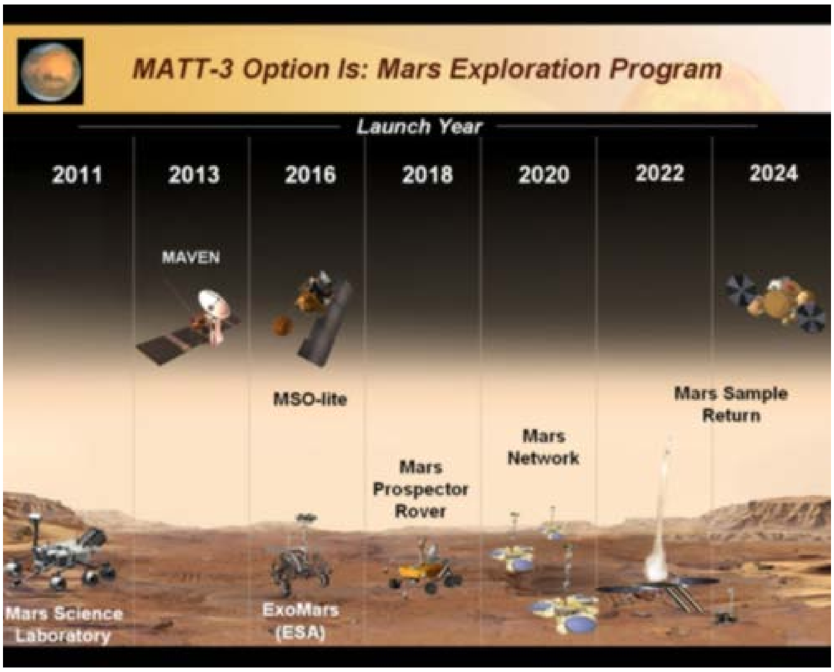


Fig. 16: Mars future programs

3.1.1 Planet Structure

The uniquely red global surface of Mars is marked by many interesting features - some like those on the Earth and others strangely different. The reddish color is caused by rust (iron oxide) in the soil. Some of these features are; volcanoes, canyon systems, river beds, cratered terrain, and dune fields. Of these features, the most interesting includes the apparently dead volcano Olympus Mons, which rises 23 km above the surrounding plains and is the highest known peak in the Solar System. Valles Marineris is a giant canyon system that runs about 2,500 miles across the surface of the planet and reaches depths of 6 km or 4 miles (for comparison, the Grand Canyon is not more than 1 mile deep).

3.1.2 Lower Atmosphere

The atmosphere of Mars is much thinner than that of Earth, with a surface pressure averaging 1/100th that at the surface of the Earth. Surface temperatures range from -113^{o}C at the winter pole to 0^{o}C on the dayside during summer. Although the length of the Martian day (24 hours and 37 minutes) and the tilt of its axis (25 degrees) are similar to those on Earth (24 hours and

23.5 degrees), the orbit of the planet about the Sun affects the lengths of the seasons the most. The atmosphere is composed mainly of carbon dioxide (95.3%), nitrogen (2.7%), and argon (1.6%), with small amounts of other gases. Oxygen, which is so important to us on earth, makes up only 0.13% of the atmosphere at Mars. There is only one-fourth as much water vapor in the atmosphere. Although small, this is thought to be enough to allow water ice to be frozen into the surface of the planet. With so little water, clouds are rarely seen in the Martian sky. The possible role in the past of liquid water in forming the dry river beds which we can see is still unknown, particularly because water ice is not plentiful on the surface of the planet.

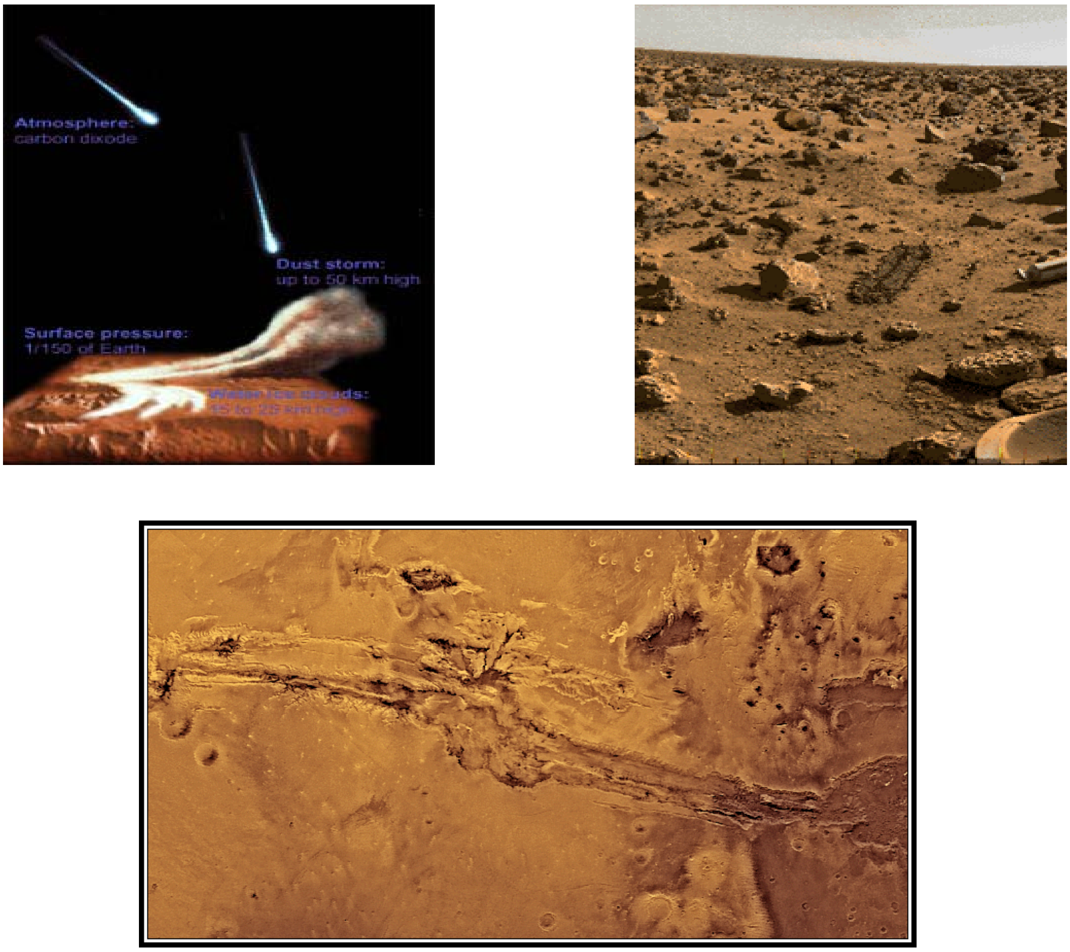


Fig. 17: Mars surface

3.2 Architecture of a Mars Surface Exploration Ad hoc network

Planetary landers, rovers and orbiters are typical of the devices that might participate in this class of networks. A renewable power source, such as solar or nuclear cells, and a larger storage capacity are the distinguishing characteristics of nodes. The mobility characteristics of these devices influence the requirements for the design of network solutions. Nodes can be either: Immobile, such as Landers, Self-mobile, such as Rover & Repeaters and Planetary orbits. As a result of these mobility characteristics, the potential for communications between two nodes may be very predictable, or may be difficult to predict. Because of the small number of nodes involved, the topologies of inter-vehicular proximity networks are fairly simple. These networks can easily be treated as a small collection of point-to-point links. In fact, in current and near term implementations, these networks are simply a single point-to-point link. For example, the Proximity-1 Space Link protocol implemented on the Odyssey Mars orbiter creates a point-to-point link between the orbiter and a lander, but does not provide a mechanism for routing traffic through intermediate nodes. The third element of the architecture on the Mars ground surface is mobile repeaters. In our architecture repeaters act a router between rover and lander. Repeaters have considerably higher power and storage when compared to rover. The main goals are to maximize the rover coverage, minimize the transmitting time, minimize the delaying time, and/or maximize the network throughput. The architecture provide rover advantage to extend its communication area, whereby rover can move away from lander and also able to explore some new interesting sites in Mars surface with same communication power.

Mars network are defined below:

Mars-Earth backbone network - This architectural element provides the long haul data links directly between Mars vehicles and the Earth as well as the Earth-based infrastructure elements. The remote assets include elements with long-haul capabilities (certain Mars surface vehicles, spacecraft in Mars orbit) and the in-space data relay network. The Earth-ground segments include the deep space network (DSN), NASA and other space agency Intranets and virtual private networks (VPN's), and the Internet. The critical technologies involve increasing the capacity and the autonomous operation of the remote long haul space link assets.

Mars vehicle proximity networks - These involve the wireless links with distances that are relatively near to the planet. There are three types of proximity links.

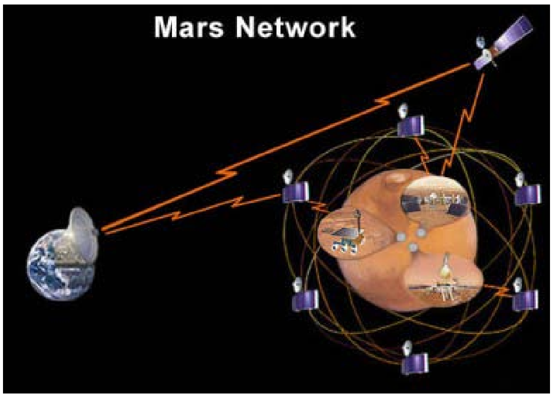
1. Orbiter to/from Surface (called access network): This architectural element provides the data links between Mars surface/airborne Vehicles and spacecraft in Mars Orbit. The orbiters will typically contain the long haul links to Earth and will host a gateway to the backbone

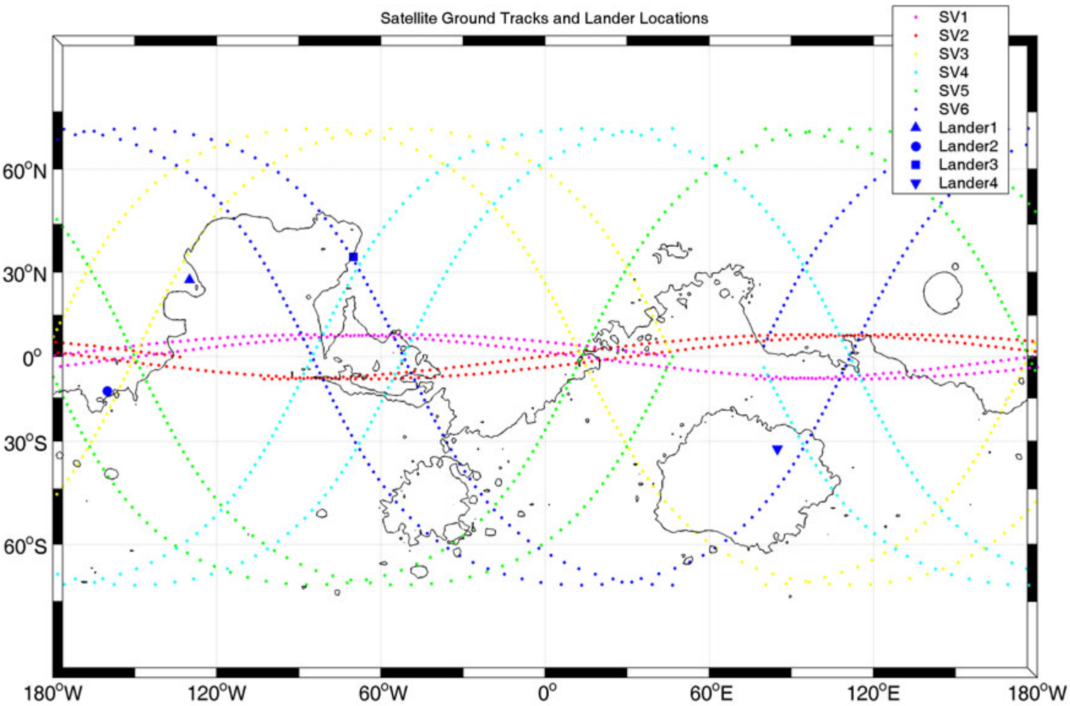
network to relay between the backbone elements and the mission spacecraft and/or vehicles, or will forward to another proximity vehicle.

2. Mars inter-spacecraft networks - This architectural element provides the data link between spacecraft flying in formation, clusters, or constellations in the Mars vicinity. It also includes the communication interfaces between approaching/departing spacecraft. In particular, spacecraft in a communication constellation that intercommunicate to relay the data to Earth are included.

3. Mars surface networks (called proximity network): This architectural element provides the data links between surface vehicles (rovers, airplanes, aerobots, landers, and sensors) spread out in an ad hoc network.

Local area networks (LAN's) on-board the Mars vehicles - This architectural element interconnects the various modules of the vehicle through an internal LAN consisting of one or more types of serial or parallel interconnected busses.





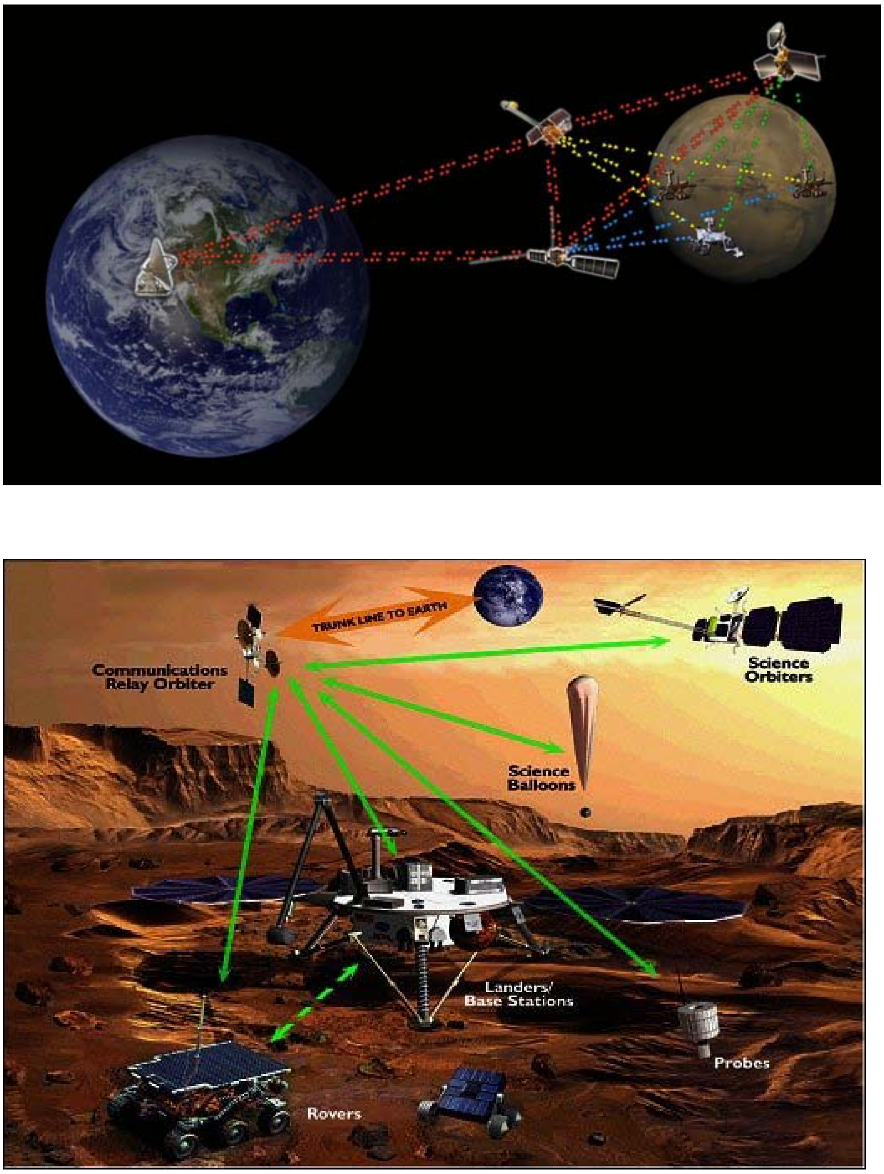


Fig. 18: Mars Network examples

Spacecraft have specific orbits, such as Low Earth Orbits (LEO), Medium Earth Orbits (MEO), or GEo-stationary Orbits (GEO) (see [Advanced Routing Protocols for Satellite and Space Networks] for details). Furthermore, the spacecraft deployed may either be relatively stationary, such as geo-stationary satellite systems (which are stand-alone spacecraft systems), or may be mobile, and join different constellations as they move around. In any case, these spacecraft have specific functions which allow them to be able to perform routing functions within a constellation (intra-constellation spacecraft) or between constellations (inter constellation spacecraft), and in some extreme cases, between constellations serving different celestial bodies (interplanetary). In networks with inter-planetary spacecraft, the network becomes a true IPAN, but intra-constellation and inter-constellation spacecraft are expected to use technologies very similar to inter-planetary spacecraft [Challenges in Interplanetary MANETs].

Stand-alone Spacecraft: A stand-alone spacecraft is deployed with a very specific orbit, and most probably is expected to have direct link(s) to a pre-determined earth station(s). Because of this reason, the routing issues in such spacecraft are very simplistic.

An example of such a satellite is GPS, which are deployed at MEO orbits, at about 20,000km above the earth's surface. Each of the 24 GPS satellites has a very well defined orbit, and the orbit can be computed as a function of time.

Commercial LEO satellites are another application of stand-alone spacecraft, where known-orbit, deterministic routing can be used to an advantage. Another application of stand-alone spacecraft is GEO satellites, whose location is always the same relative to earth. Because of this reason, they can benefit from extensive ground support. Typical one way latencies from earth to GEO satellites are around 0.25 seconds. Space stations are also examples of stand-alone spacecraft.

Intra-constellation Spacecraft: Typical uses for intra-constellation capability spacecraft are earth science missions and remote planet exploration. By allowing several of these spacecraft to be in orbit around earth or any other remote planet, certain useful services like water/air quality management, surveillance can be provided. For earth-science or remote planetary-science missions, a planet about the size of earth requires about 40 of these satellites for full coverage because these satellites are typically deployed at LEO orbits. Latencies from earth to satellite for LEO satellites are about 0.10 seconds. Intra-constellation spacecraft are deployed with the capability to perform routing between other spacecraft in the constellation apart from being able to communicate with a ground station or routing-facilitating agents such as Unmanned Airborne Vehicles (UAV). These spacecraft have no fixed location characteristics and float around in a controlled or uncontrolled manner in space. Such spacecraft, when within a constellation set up a MANET among them and perform the routing tasks autonomously. The spacecraft within a constellation are expected to be able to establish communications in an ad hoc manner when they are in the neighborhood of one another. Apart from this requirement, intra-constellation satellites may also be required to participate in routing activities with a hierarchical superior.

Inter-constellation Spacecraft: Inter-constellation spacecraft are very similar in function and purpose to intra-constellation spacecraft, but they also perform routing between adjacent constellations.

Inter-planetary Spacecraft: Inter-planetary spacecraft are deployed with a primary reason to be sent out in outer space for missions such as deep space exploration. Apart from that purpose, their role is also to facilitate IPAN routing between the local constellation in the distant planet and other constellations in neighboring planets (such as constellations orbiting earth). Thereby, they are expected to have more computing power and storage facilities, and are built to be very durable. One example of an inter-planetary spacecraft is the deep-space explorer Pioneer 10, which was launched on March 2, 1972 and is still functional.

Inter-planetary spacecraft are key players in the deployment of IPAN MANETs because they bridge the communication between smaller, less powerful spacecraft in orbit around neighboring planets. Planetary exploration has opened huge possibilities in the area of IPAN networking. For Mars we can use a constellation of spacecraft (which could be any combination of LEO/MEO/GEO satellites, ULDB and MEMS devices) which gather real-time data at Mars will need to relay the information to earth-based stations.

3.3 Challenges imposed by this type of environment

In the years 90's, NASA developed his first deep space communication network, DSN, to keep control of space probes that they sent in exploration missions to Mars and to the other bodies of the solar system. Recently, the agency used this powerful system to communicate with Mars rovers Spirit and Opportunity as well as with the Huygens lander that visited Titan.

3.3.1 Advantages and drawbacks

Signals transmitted on HF bands (1.8-30 MHz) at low incidence have difficulties to pass through ionosphere layers surrounding the Earth, whether they are transmitted from the ground or from space. Under some conditions, these regions of high electronic density act like mirrors for radio waves.

If it is a huge advantage for ground stations who wish to communicate with remote stations (by bouncing successively between the ionosphere and the ground, HF waves can reach DX stations), for the same stations wishing to communicate with satellites moving into space it is rather a big problem. Dust storms can begin clouding the Martian atmosphere with dust. The storms can put the rovers in the real possibility of system failure due to lack of power. The key problem caused by the dust storm is a dramatic reduction in solar power. There is so much dust in the atmosphere that it blocks 99 percent of direct sunlight to the rover. Normally the solar arrays are able to generate about 700 watt-hours of energy per day. During the storms, the power generated is greatly reduced. If the rovers get less than 150 watt-hours per day they have to start draining their batteries. If the batteries run dry, key electrical elements are likely to fail due to the intense cold. On July 18, 2007, the rover's solar-panel only generated 128 watt-hours, the lowest level ever. NASA responded by commanding Opportunity to only communicate with Earth once every three days, the first time that this had happened since the start of the mission. When a low-power fault is tripped, the rover's systems take the batteries off-line, putting the rover to sleep and then checking each sol to see if there is sufficient available energy to wake up and perform daily fault communications. If there is not sufficient energy, it will stay asleep. Depending on the weather conditions, the rover could stay asleep for days, weeks or even months, all the while trying to charge its batteries with whatever available sunlight there might be.

3.3.2 The cut-off frequency

[http://www.astrosurf.com/luxorion/qsl-mars-communication.htm]

The ionosphere is a mixture of plasma and atoms more or less ionized by the solar UV radiations. There is a cut-off frequency f_p for the ionosphere beyond which it loses its capacity to reflect shortwave. Depending the latitude, the season and the solar activity mainly, during the day this frequency is around 3-10 MHz and goes down to about 2-6 MHz during the night. Extrapolating these properties towards higher frequencies, we can conclude that signals transmitted in V/UHF and SHF bands easily pass through the ionosphere, excepting at grazing incidences.

Between 1-90 MHz the behavior of the ionosphere as mirror for shortwave is rather complex, with both reflections and attenuations of the signal quite important. But the ionosphere is not completely opaque to HF waves. Like a window, when the incidence angle is very wide (wide compared to the normal), the waves are completely reflected by the surface which acts like a mirror, while when the waves travel perpendicular to its surface, they are able to entirely pass through it.

But to work in this way is not without drawbacks, and at daytime transponders of these satellites are affected by the layer ionization during periods of high solar activity. HF is thus not always the best method to work by satellite.

3.3.3 Background noise

In the UHF bands radio waves carry much interference and various noises which are superimposed on signals. That starts in the long waves (LW or AM) with atmospheric noises (thunderstorms) that extend to the VHF, sensitive to meteors ionization trails and auroras, to attenuate towards approximately 500 MHz's. It is thus necessary to go up in frequency to cut off from these problems.

But the Sun and thunderstorms are not the only elements that scramble space communications and more one goes up in frequency more one records a background noise which is in fact related to the gamma radiations emitted by galactic and extragalactic sources (stars, supernovae). On the other hand, more one goes up in frequency and outside the gamma emission lines, the Sun influences on the wave propagation weakens.

3.3.4 Bandwidth

Beside these two problems of propagation and noise, using SHF bands (2.9-30 GHz) rather than HF for space communications, not only the noise level and interferences are much weaker there, but at these cent metric wavelengths we can use very directive antennas and broadband emitters to transmit high volumes of data. Indeed, to fasten transmissions of large volumes of data without error, it is necessary to work with bandwidth of several megahertz. It is thus impossible to work this way on HF (a signal uses 3 kHz in phone and 1 kHz in Morse) because the signal would occupy the entire HF spectrum and would make it unusable for any other application. Taking all factors together, if we want to pass through the ionosphere layers by the means of shortwave, we must use higher frequencies, which wavelength is insensitive to the electron density but also free of interferences and suited to carry high rates.

3.3.5 The X-band

For space communications the main frequencies bands that we can use are microwaves and particularly the S-, X- and K-bands at a few GHz, which represent a wavelength of a few tens of centimeters maximum. Beyond ~20 GHz other problems arise like the atmospheric absorption due to clouds and rain.

The S-band near 2 GHz is first of all used by ships at sea (radiolocation and radio navigation) as well as for GSM, Inmarsat and DCS communications.

The X-band is the favorite frequencies band used for many commercial and scientific transmissions: small and large dishes tuned on X-band between 10.7 and 20 GHz are used by satellite TV, in geodesy (TOPEX oceanographic applications), in meteorology (polar metric radar) or in radio astronomy. The X-band is so crowded because basically its main advantage is to be within the reach of everybody, and it can be used with small antennas, low power and relatively simple equipment compared to AM or HF installations for example. This band of frequencies is also less perturbed and allows the transfer of huge volumes of data. At last used with digital networks and high powers it is less sensitive to phase rotation. At first sight, many X-band applications work by satellite. Indeed, at a few GHz wave propagation does not exactly "work" like long waves (AM) or even like decametric or V/UHF wavelengths. Due to their short wavelengths, signals are quickly disturbed by usual obstacles like buildings or mountains. And they are also reflected by ionization clouds (auroras, meteor trails) and particles in suspension in the air (dust, rain, snow, hail).X-band antennas are thus first of all directed toward space, to artificial satellites, to the Moon and to the deep sky objects, including for emission purposes to probe the ground and the underground of the Earth, and to communicate with satellites. In most cases, professionals have upgraded the equipment for space telecommunications to powers of a few kilowatts and use parabolic antennas (dishes) of several meters in diameter. Today the X-band allows astronomers to make very accurate measurements on very far targets, which distance can exceed tens of millions of kilometers for planets. Such antennas are also capable to receive data from spacecrafts traveling at the borders of the solar system in the case of Pioneers and Voyager's, today quasi out of the solar system, or straight out from the deep sky when they study radiations emitted by quasars located at a few billion light-years away. NASA's Deep Space Network (DSN) is an international telecommunication network initiated in 1987. The first antenna, Deep Space Station 13 (DSS 13), was constructed between 1988 and 1989. It saw its "first light" in early 1990. The DSN consists of large antennas equipped with cryogenically cooled low-noise amplifiers (approx. 20 K noise temperature) that ensure communications between laboratories as well as with Mars Exploration spacecrafts placed on orbit or that landed on its surface. This network is of course used to contact most other spacecrafts from Ulysses to Voyagers. To permit constant control of spacecrafts while the Earth rotates on its own axis, the DSN use three deep-space communications facilities placed approximately 120° apart around the world. The DSN antennas are similar to radio telescope dishes: DSS 13 is 34 m (111.5 ft) and DSS 14 is 70 m (229.7 ft) in diameter. These sizes are necessary to communicate with spacecrafts located hundreds of millions km from Earth and which signals are very weak; the larger the antenna, the stronger the signal and greater the amount of information the antenna can send and receive. On their side each spacecraft, whether it is an orbiter or a lander, carries several antennas used for different phases of the mission and that permit it to communicate in different directions and with other spacecrafts at different transmission rates. The DSN communicates with nearly all spacecrafts exploring the solar system. Currently they are 28 spacecrafts cruising in space. DSN antennas are thus extremely busy trying to track all of these space missions at the same time. Therefore each spacecraft works in time sharing on the DSN antennas, each of them receiving a predefined timeslot to communicate with Earth and vice versa and a sophisticated scheduling system ensures that each mission's priorities are met. During critical mission events, such as landing on Mars, multiple antennas on Earth and the MGS orbiter track the signals from the maneuvering spacecraft to minimize risk of loss of communication. During the landing operations phase on the Martian surface, the Mars Exploration Rovers use the Multiple Spacecraft per Aperture (MSPA) capability of the DSN, which allows a single DSN antenna to receive downlink from up to two spacecrafts simultaneously. A Martian rover like Opportunity can use four different antennas: an UHF antenna or a low-, medium-, or high-gain antenna. This capability gives the mission team several different ways to send directly commands to rovers without relaying via the orbiter and to return data back to Earth. The rovers' downlink sessions (when rovers send information back to Earth) are generally limited to a couple of hours at a stretch, with a maximum of two downlink sessions per Martian day (sol) per rover. The MSPA technique allows only one spacecraft at a time to have the uplink and the rovers command early in each sol for roughly an hour each to provide instructions to execute during the day.

3.3.6 Doppler Effect

Grounds controllers that ensure communications with rovers know the frequency of the signal at rest that is transmitted by the spacecraft. But since the spacecraft is moving into space this frequency is being Doppler shifted a bit lower or higher in frequency. Knowing the speed of electromagnetic waves, a straightforward calculation allows finding the velocity of the spacecraft.

3.3.7 Ranging

Ranging allows knowing the distance to the spacecraft with accuracy. Engineers send a code to the spacecraft that immediately send it back using its own antenna to Earth. The delay between sending and receiving the code, minus the delay in turning the signal around on the spacecraft, is twice the light time to the spacecraft. So, in dividing by two and multiplied by the speed of light we can calculate the distance from the DSN station to the spacecraft. This distance is accurate to about 5-10 m, even though the spacecraft may be 200 millions km away.

3.3.8 Delta DOR

Delta DOR (Differential One-way Ranging), is similar to ranging, but it also takes in a third signal from a naturally occurring radio source in space, such as a quasar (a strong extragalactic radio source), and this additional source helps scientists and engineers gain a more accurate location of the spacecraft.

3.3.9 Special tones during entry, descent, and landing

During the phases of entry, descent, and landing (EDL), the Mars Odyssey spacecraft like all others encounter turbulent conditions. Navigators must safely maneuver Spirit and Opportunity to their precise atmospheric entry points to reach their landing targets on the surface of Mars. In only six minutes, the spacecraft must slow down from an incredible speed of 5.3 km/s or 19600 km/h to nearly zero, thus six times faster than the space shuttle. During these few minutes, the spacecraft experiences intense heating from friction caused by speeding into the atmosphere; it jostles when the parachute deploy; and the lander and rover bounce along the surface in the airbags before they come to a rest on the Martian surface.

The quick and intense movements caused by entry and landing make it difficult to accurately track the spacecraft during this phase, so the communications leads have set up a series of basic, special individual radio tones that will ring during different phases of the entry, descent, and landing process. In order for the engineers to know if the parachute deployed for example, a tone of a certain pitch will sound in the control room. Yet a different tone will ring when the airbags deploy. Engineers on Earth will track the EDL process by listening for 128 distinct tones (out of 256), all of which have individual meanings throughout the EDL process.

3.3.10 Response time (Delay) and interferences

At the perihelion opposition, at 56 millions km from the Earth, a distance as short as 0.37 AU (1 AU = 149.56x10^8 km), it takes 3 minutes and 7 seconds for a signal emitted by the DSN to reach Mars. But when Earth and Mars are the farthest apart at 2.52 AU, it takes 20 minutes and 57 seconds to transmit the same radio signal. The communication lasts seven times longer.

Due to these time delays it is impossible to communicate with and control the rover in real time.

When Earth and Mars are in conjunction (opposite sides of the Sun) at a distance of 2.49 AU another problem arises. This distance is not as much of a problem as having the Sun in the way, for it produces a lot of radio interference making communication almost impossible. Indeed, for distances of less than 10 solar radii around the Sun, the thermal noise contribution is quite severe and the use of amplifier at reception still increases this difficulty. Therefore it is very important than the spacecraft flying to Mars reaches the Red planet far before the conjunction so that engineers and scientists can gather data during a few months before the communications problems.

3.3.11 Noise

Apart the Sun and the problem of distance, two other noise sources interfere with telecommunications: cosmic rays and thermal noise generated by the receiver. The signal strength or noise level estimation, also known as the dB W, is a measurement of the absolute power expressed in watts, and no more a power ratio like could be the decibel. Knowing the signal power and the noise level at the source, at the distance of the Orbiter, we can estimate the signal-to-noise ratio (S/N) according to the bandwidth used. Like in radio astronomy, in space communications, engineers estimate that a noise level of -215 dB W/Hz at 10 GHz is acceptable for the large ears of the DSN network. But 8 dB means that the DSN can theoretically receive such a signal without using error correction protocols, DSP systems or any BPSK or alike mode (although it does). In such conditions the transmission rate is relatively fast, up to 21 KB/s (166 Kbit/s).

3.3.12 Local communications on Mars

Messages sent by rovers are first of all directed to Earth, but their power is so weak that even using the largest antenna from the DSN their location and recording is a complex and full-time task. They communicate daily directly to Earth using the HGA, but most of the time rovers mainly uplink their information to nearest spacecrafts orbiting Mars, utilizing the Mars Odyssey or Mars Global Surveyor orbiter for example as messengers to pass along news to Earth as soon as they are in light of sight of their antennas.

Conversely orbiters can also send messages received from the DSN stations to the attention of rovers. The benefits of using the orbiting spacecraft are first linked to the fact that the orbiters are much closer to the rovers than the DSN antennas on Earth, and then the orbiters have

Earth in their field of view for much longer time periods than the rovers on the ground that are subject to the Mars rotation.

This mode of communication is precious because orbiters are only 400 km above the surface of Mars and rovers don't have to use as much energy to send a message to orbiters as they do to communicate with DSN stations. The rovers communicate with Earth during a few months, using three different systems: a low-gain antenna (LGA), a UHF antenna, and a high-gain antenna (HGA). DSN antennas communicate with far-flung spacecrafts using microwaves S-, X- and K-bands at frequencies of 2.2, 8.4 and 32 GHz.

The rovers exploring Mars communicate with each other, with orbiter(s) and the DSN through X-band UHF antennas, which are close-range antennas used at low power. One UHF antenna is on the rover and one is on the lander to aid in gaining information during the critical landing event while the orbiter tracks the landing process.

When rovers communicate directly to Earth, they send messages via both the low-gain antenna (LGA) and the high-gain antenna (HGA), both being capable of receiving the 7.2 GHz uplink signal and transmitting the 8.4 GHz downlink signal. The LGA transmits signals at a low rate to the DSN antennas when the rover orientation is unknown. The Omni directional UHF antenna communicates through orbiters passing overhead, not only with the mother spacecraft (Mars Odyssey for the rover Opportunity) but also with other passing orbiters like Mars Global Surveyor.

The HGA is a steerable beam that points itself directly to any antenna on Earth. The benefit of having a steerable antenna is that the entire rover doesn't necessarily have to change positions to speak with the DSN. The rover can save energy by steering only its antenna in the right direction.

About half of all communications will go through the HGA, a dish 28 cm in diameter that beams data directly to DSN over the X-band (8 to 12 GHz). The downlink data rate to the orbiter is established with the HGA at 264 bytes/sec (1.85 kilobits/sec) while the uplink data rate is 125 bytes/sec (0.875 kilobits/sec) utilizing the LGA.

Several MANET routing schemes address the need for localizing route searches and flooding so as to utilize the bandwidth more efficiently. Given the fact that Omni directional beaming used in many terrestrial MANETs can be wasteful, we have to tackle the problem of improving routing efficiency by some form of location service (such as an inter-planetary GPS) so that resources on the spacecraft are used more efficiently. If we use directional antennas, by using precise tracking and location information, we can not only narrow our beams used for transmission and reception but we can also decrease collisions and increase efficiency of use of bandwidth. The velocities and range of the spacecraft used in IPAN related topologies range from zero to very high. For instance, Pioneer 10 launched by NASA in the early 70s is moving at a speed of 27,380 mph relative to the sun, at a distance of approximately 80 astronomical units from earth. This means tracking in an IPAN context is not an easy task. If tracking is possible, it has to be done very efficiently if we want reasonable usage of bandwidth. Even if the error in terms of angular location detection (azimuth and elevation details, if we use polar co-ordinates) is very low, the error in physical location of a target node can be drastically high. Another issue regarding antenna beam-forming is that we will have to deal with beam divergence even assuming no signal attenuation whatsoever. One solution proposes the use of lasers to circumvent divergence issues. If we use a MAC layer data structure like ZRP's Neighbor Discovery Mechanism's (NDM) neighbor table, then we can address the issue of neighbor location at the MAC layer by recording direction details of the neighbors in the neighbor table.

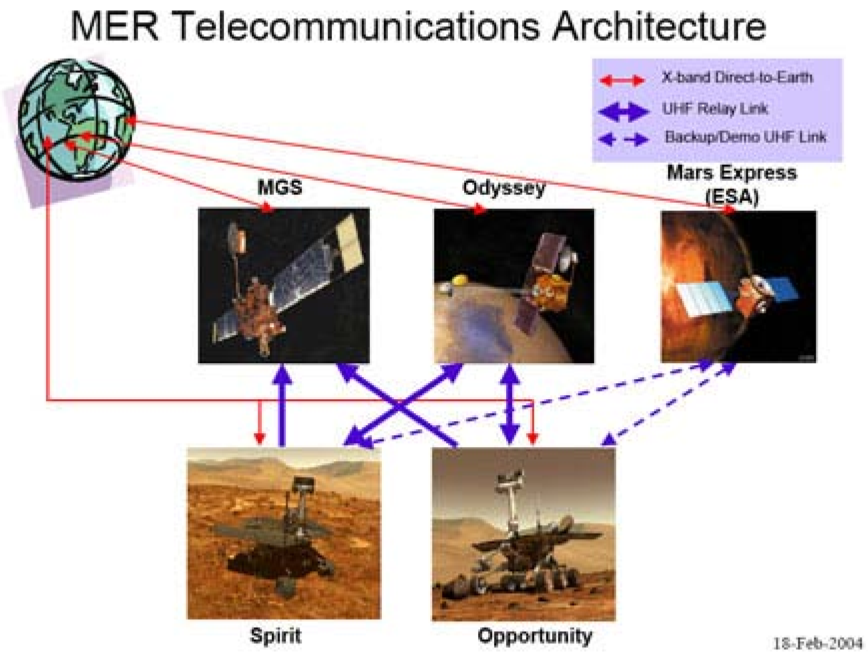


Fig. 19: MER Telecommunications architecture

3.3.13 Maximum usable distance

The rover telecom system was functionally tested up to a distance of 250 meters from its base emitter. It performed quite well under environmental conditions typical of a warm (35°C or 95°F) August day. Engineers increased the distance well over 700 meters, the interval showing abundance of multipath reflections. The radio modems performed also well under those conditions and did not lose communications.

However, under certain conditions, there is degradation in the quality of the communications link. In particular at the lowest acceptable operating temperature of -30°C (-22°F), the Bit Error Rate (BER), due to an operating frequency shift may cause a communications blackout problem. If the rover is kept in the line of sight of the lander and the radios are kept at a warmer operating temperature, the maximum usable distance of the rover telecommunications system should be at least 700 meters.

3.3.14 Deep space transponder

A transponder is a communication system placed on board a relay satellite, which receives and re-transmits the signal, often on different downlink and uplink frequencies. This system is activated when the ground stations are out of sight from the orbiting satellite or from the rover and the direct communication link is lost.

3.3.15 Telecommunication relay capabilities

Using the HGA, the downlink data rate directed to Earth varies from about 1.75 KB/s (12 kilobits/sec) to as small as 0.5 KB/sec (3.5 kilobits/sec), speeds roughly five times slower than a standard 56K modem. But the data rate to the orbiters is much faster, with a constant 18.2 KB/sec (128 kilobits/sec), twice faster as the mean data rate of a 56K modem.

An orbiter passing over the rover is in the vicinity of the sky to communicate with the rovers for about eight minutes at a time, per sol (Martian day). In that time, about 8.5 MB of data (about 1/100 of a CD-ROM) can be transmitted to the orbiter. That same volume of data sent directly by the rover would take between 1.5 and 5 hours to transmit directly to Earth. Rovers can only transmit direct-to-Earth for at most three hours a sol due to power and thermal limitations, even though Earth may be in view much longer.

But there is a "drawback" in this system. While Mars is rotating on its own axis it carries the rover with it and soon or late this latter is out of the field of view of Earth. Hopefully orbiters can see Earth for about 2/3 of each orbit, or about 16 hours a sol. They can thus send much more data direct-to-Earth than the rovers, not only because they can see Earth longer, but because they can operate their radio for much longer since their solar panels get light most of the time, and they have bigger antennas than rovers.

In all cases, the Orbiter provides navigation to assist in arrival at descent maneuvers as well as landed operations. Orbiting relays provide also a link to the night side of Mars that is hidden from the Earth's view.

This short surface-to-orbit relay link is far more efficient that a direct to Earth link. Space communications difficulty with Earth increases as the square of the distance. At worst the maximum range is 2.7 A.U. or 400 millions km. Compared to this, the in situ link ranges between 1000 and 6000 km only. In such circumstances the power loss to Earth exceed 266 dB (at 56 millions km) compared to the in situ link on Mars. Here your wattmeter should read 10^{-26} W at reception.

This loss must be compensated by the antenna gain, at both transmit and receive sites. This huge difference of power is difficult to manage even using the largest DSN antenna, DSS 14 of 70 m in diameter. The Orbiter offers thus a first major advantage in providing surface-to-orbiter relay link to Earth with a reduced amount of energy. The second advantage is that in the same timeslot the rate and volume of data transmitted have drastically increased. In future, orbiters will fly at higher altitudes to provide longer communications with the surface rover, what will extend the coverage from 2 hours to 6-12 hours. Only drawback of such high orbits is the longer slant range to users. It will be compensate by the use of a medium gain and steerable antenna.

Fig. 20: Mars Rover and Satellite

All spacecraft must utilize a high degree of autonomy due to their low-bandwidth, high latency communication channels to Earth. Traditional autonomous capabilities include responding to faults, pointing antennas, orientation control using star tracking, and onboard data storage and retransmission. Spacecraft used to explore space or orbit planetary bodies travel through a relatively simple environment and need only perform navigation corrections tens to hundreds of times during their multiyear missions. However, the rovers used to explore the surface of Mars must constantly interact with challenging and unknown terrain.

For example, in their first three years, the MER vehicles performed some 60,000 coordinated motions— powering either steering or drive motors continuously.

Autonomy software allows a rover to make decisions and command actuators based on its observations of the environment or sensor feedback. Without it, human operators must verify the state of the vehicle before and after it performs any action. High-bandwidth, low-latency communications make direct remote operation of a vehicle possible. Increased autonomy will be critical to the success of future missions, in which rovers will be expected to travel over long distances in a short period of time and handle dynamic processes such as taking a core sample from a rock while slipping on a slope. In addition to the communications limitations, exploring the surface of another planet presents many mobility and sensing challenges. A rover must sense the terrain in widely varying lighting conditions, interact with terrain that might not be fully characterized, and account for uncertainties in sensing or system faults.

Further, planetary vehicles have extremely limited computational resources due to the high radiation levels and large temperature changes of space. Sojourner had a 0.1-MHz Intel 80C85 CPU with 512 Kbytes of RAM and 176 Kbytes of flash memory, the MER vehicles have a 20-MHz RAD6000 CPU with 128 Mbytes of RAM and 256 Mbytes of flash memory, and the MSL vehicle will have a 200-MHz RAD750 PowerPC with 256 Mbytes of RAM and 512 Mbytes of flash memory. The MER and MSL vehicles use the VxWorks operating system and run many parallel tasks continuously, leaving autonomy software with less than 75 percent of the available CPU time.

3.3.16 Routing Issues

There are a lot of networking concerns that are characteristic of such missions. Among the bigger difficulties involved in setting up IPANs are the amount of time involved for setting up the infrastructure in space. Using the bandwidth-delay product efficiently is another issue.

One way of addressing the networking needs of IPANs and related topologies such as those described in the previous section is by fitting the MANET framework to the topology and addressing the issues that arise in doing so.

The protocols studied by the working groups are not immediately applicable to the IPAN networks situation because of the extreme needs of space networks. In IPANs and IPAN related topologies using ad hoc space networks, high speed networks in space may need to track other satellites in a constellation or adjacent constellations to maintain a viable network structure. In addition, such networks may consist of spacecraft deployed at different orbits, possibly together with geo-stationary spacecraft, to support multi-tier ad hoc structures. Routing may be performed within the same tier or across tiers based on different latency and bandwidth constraints. Some issues regarding routing in an IPAN-related topology are discussed in [Interrogation-Based Relay Routing in Ad hoc Satellite Networks]. Table-driven routing fails in such networks owing to multiple reasons – there are many network nodes that do not have a deterministic route pattern, the connectivity is intermittent, and the network is very sparse. Further, table-driven routing is not scalable. Proactive approaches are also not very effective in the IPAN context because proactively maintained routes change too soon. Reactive schemes such as AODV and DSR seem preferable because they initiate the route discovery process in an on-demand basis, but reactive protocols suffer from longer delay and excessive control traffic owing to on-demand route discovery process. Besides, in a sparse network, route discovery will not work. Relay-based schemes are an option in such situations.

In IPANs using the interrogation-based relay routing scheme, a next hop may not be available immediately for the current node to forward the data. The node will need to buffer the data until the node gets an opportunity to forward the data. The underlying assumption of the relay routing scheme is that it does not rely on any pre-calculated routes as in the static routing scheme, or any periodic route computation as in the dynamic routing scheme. Relay-based routing is essentially an on-the-fly store and forward technique. The nodes must be capable of storing data temporarily for a considerably longer duration to cater to the situation when there is a network partition.

An issue closely tied with routing in IPAN topologies is addressing. As with other MANETs, the hierarchical address allocation scheme currently used in the Internet may not be optimal for IPAN networks because of the fact that the nodes need not carry stable hierarchy/topological information. Besides, given the vastness of space, the current addressing scheme may fall short of our future addressing needs.

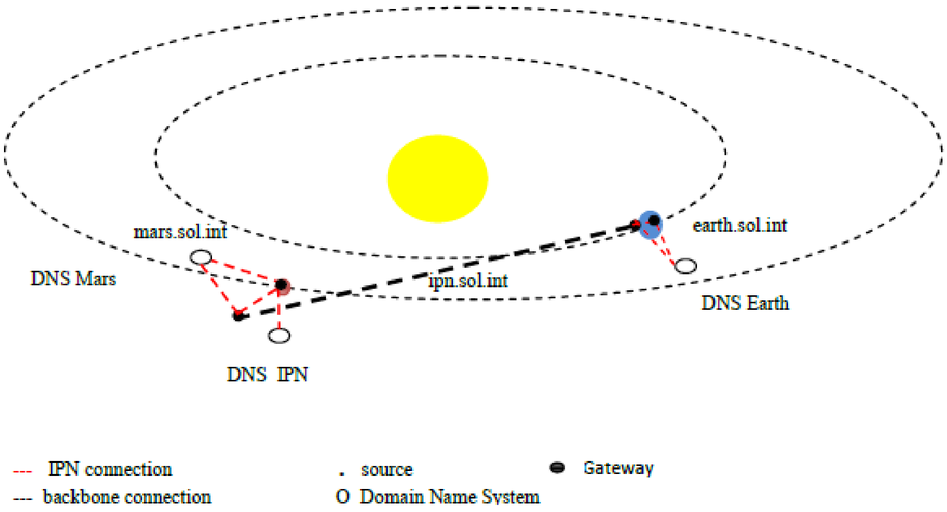


Fig. 21: Earth-Mars Network addressing

Due to the potential impairment of free space wireless connections between spacecraft and between spacecraft and ground stations, managers and cluster heads may not be in constant contact with managed nodes, as is required in the hierarchical approach. Besides, a mobile agent-based, adaptive, autonomous approach to management is required to address the number of events that are generated in IPAN contexts.

In the autonomous approach, we envision the IPAN network to be self-organized and self-managed with the collaboration of autonomous nodes. The nodes range in heterogeneous functionality and capability, from geostationary and LEO satellites to in-situ and ocean buoys. Therefore, not all nodes will be able to contribute equally to the management tasks. Nodes with greater capability will possess management intelligence (management software modules and states) and function as nomadic managers, while other nodes are managed nodes. For scalability purposes, nodes are adaptively clustered into groups with at least one nomadic manager in each group. The nomadic managers collaborate autonomously to manage the entire ad hoc space network without the help of any external entity. Nodes may also leave a group and join another when changing their positions. Corresponding nomadic managers cooperate to facilitate the “hand-off” process. Due to the dynamic nature of ad hoc space networks, the role of nomadic managers may change according to topology, node density, and other attributes. When node density increases, a nomadic manager may decide to clone its management intelligence to another node that is capable of hosting nomadic modules in order to share the management load.

Chapter 4

4. Proposed Solutions

4.1 DTN (Delay Tolerant Networking)

In a physical ad hoc network, the assumption that there is a contemporaneous end-to-end path between any source and destination pair may not be true, as illustrated below. In MANETs, when nodes are in motion, links can be obstructed by intervening objects. When nodes must conserve power, links are shut down periodically. These events result in intermittent connectivity. At any given time, when no path exists between source and destination, network partition is said to occur. Thus, it is perfectly possible that two nodes may never be part of the same connected portion of the network.

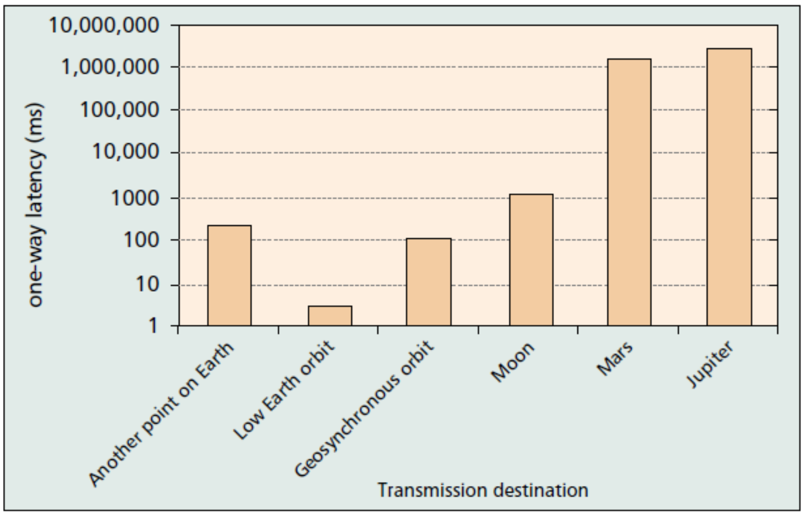


Fig. 22: Maximum latency (transmit from Earth)

Applications must tolerate delays beyond conventional IP forwarding delays, and these networks are referred to as delay/disruption tolerant networks (DTN). Routing protocols such AODV and OLSR do not work properly in DTNs, since under these protocols, when packets arrive and no contemporaneous end-to-end paths for their destinations can be found, these packets are simply dropped. New routing protocols and system architectures should be developed for DTNs. There are many potential applications in DTNs, such as inter-planetary network (IPN), Zebranet, DataMule, and village networks. IPN consists of both terrestrial and interplanetary links, which suffers from long delays and episodic connectivity. In the DataMule project [http://nesl.ee.ucla.edu/project/show/19] for example, DataMules randomly move and collect data from low power sensors. There are many different terminologies used for DTNs in the literature, such as eventual connectivity, space-time routing, partially connected, transient connection, opportunistic networking, extreme networks, and end-to-end communication.

The data unit in DTNs can be a message, a packet, or bundle, which is defined as a number of messages to be delivered together. For simplicity, we use bundles, messages, and packets interchangeably. The characteristics of DTNs are very different from the traditional Internet in that the latter implicitly has some well-known assumptions: continuous connectivity, very low packet loss rate, reasonably low propagation delay, low or constant transmission latency, low error rate, low congestion, high transmission rate, symmetrical data rates, common name or address expression syntax or semantics or data arrival in transmission order.

DTNs do not satisfy all of these assumptions, and sometimes none. The challenges in designing efficient protocols in DTNs are:

• Very large signal propagation latencies (on the order of minutes)

• Relatively low data rates (typically 8–256 kb/s)

• Possibly time-disjoint periods of reception and transmission, due to orbital mechanics and/or spacecraft operational policy

• Intermittent scheduled connectivity

• Centrally managed access to the communication channel with essentially no potential for congestion

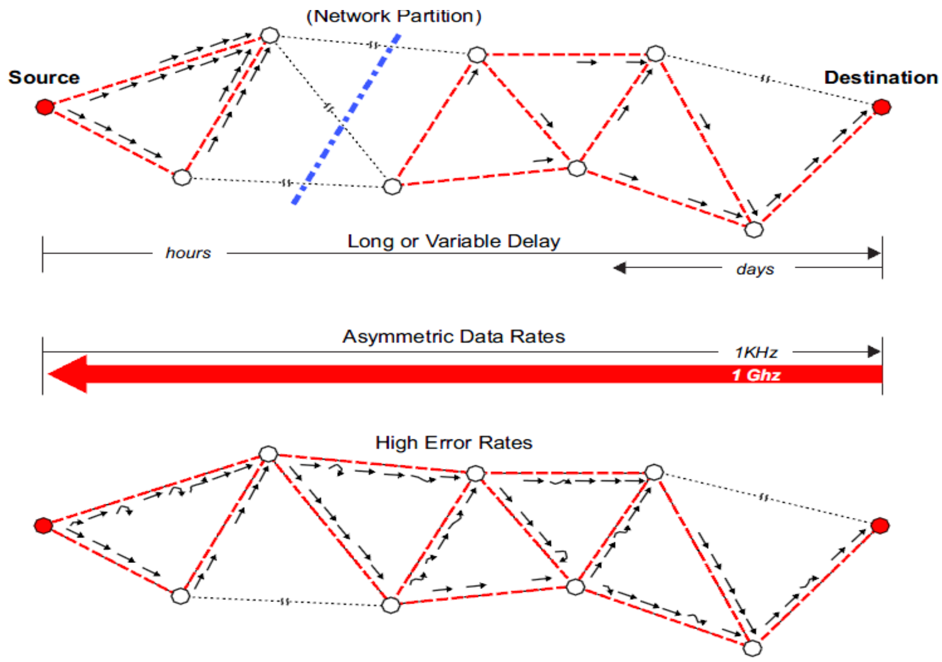


Fig. 23: DTN properties

Working from these considerations, fundamental principles of delay tolerant networking (DTN) architecture were identified.

•A postal model of communications. Because transmission latency can be arbitrarily long, reliance on negotiation, query/response, or any other sort of timely conversational interchange is inadvisable, in both the overlay network protocol and the applications themselves. Insofar as it is possible, the data transmitted through the network should constitute self-contained atomic units of work. Applications should issue messages asynchronously not wait for the response to one message before sending the next.

For example, a delay-tolerant request for transmission of a file would not initiate a dialog as in FTP. It would instead bundle together into a single message not only the name of the requested file, but also — unprompted — all other metadata that might be needed in order to satisfy the request: the requesting user's name and password, encoding instructions, and so on. In recognition of this general model, the units of data exchanged via the DTN overlay network protocol are termed bundles (which are functionally similar to email messages); the protocol itself is named Bundling.

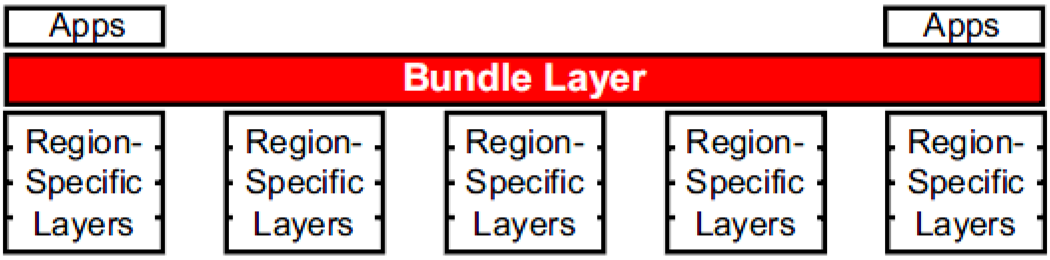


Fig. 24: Bundle layer

•Tiered functionality. The protocols designed for use within various environments already exploit whatever favorable circumstances the environments offer while operating within their constraints, so the DTN architecture relies on the capabilities of those protocols to the greatest extent possible. The Bundling protocol, one layer higher in the stack, performs any required additional functions that the local protocols typically cannot.

•Terseness. Bandwidth cannot be assumed to be cheap, so the DTN protocols are designed to be taciturn even at the cost of some processing complexity.

The main structural elements of DTN architecture, derived from these principles, are as follows. The combination of long signal propagation times and intermittent connectivity caused, for example, by the interposition of a planetary body between the sender and the receiver can result in round-trip communication delays measured not in milliseconds or even minutes but in hours or days. The Internet protocols do not behave well under these conditions. Yet a retransmission protocol of some sort is required to assure that every bit of the new executable module is correctly received. Forward error correction (FEC) can reduce data loss and corruption, but it consumes bandwidth whether data are lost or not, and it offers no protection against sustained interruption. Optimum utilization of meager links demands automated repeat request (ARQ) in addition to some level of FEC. What is needed on this part of the route is an ARQ system for efficient retransmission on the link.

The Internet protocols are in general poorly suited to operation on paths in which some of the links operate intermittently or over extremely long propagation delays. The principal problem is reliable transport, but the operations of the Internet's routing protocols would also raise troubling issues. While those issues don't seem insoluble, solutions would entail significant divergence from typical operations in the Internet.

Consequently, the existing protocols developed for the wired Internet are not able to handle data transmission efficiently in DTNs. In DTNs, end-to-end communication using the TCP/IP protocol may not work, as packets whose destinations cannot be found are usually dropped. If packet dropping is too severe, TCP eventually ends the session. UDP provides no reliable service and cannot "hold and forward." New protocols and algorithms need to be developed. There are several different types of DTNs due to their different characteristics. For instance, the satellite trajectories in example are predictable, while the movement of a soldier or tank may be random. Therefore, for different types of DTNs, different solutions may need to be proposed. Recently, the DTN research group under the Internet Research Task Force (IRTF) has proposed several research documents, including DTN architecture. This architecture addresses communication issues in extreme networks or networks encompassing a wide range of architectures. The architecture is a network of regional networks, and an overlay on top of the transport layer of these regional networks. It provides key services, such as in-network data storage and retransmission, interoperable naming, and authenticated forwarding. Some DTN solutions are concerned with message transport between infrastructures of disparate architectures by using gateways that handle "bundles" of messages between these infrastructures.

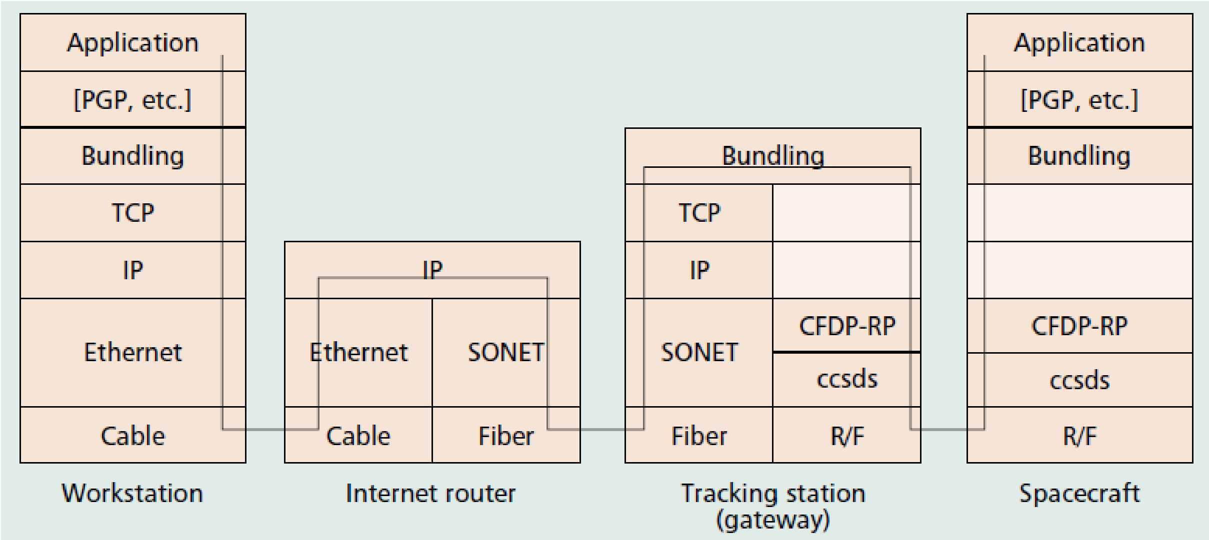


Fig. 25: Example of data flow

The DTN architecture addresses the issues of eventual connectivity and partitioned networks by the use of a store and forward mechanism, and handles the diverse addressing needs of the overlay architecture by using an addressing scheme that exploits the late binding of addresses — local addresses are not bound to nodes until the message is in the local area of the destination. This creates a hierarchical routing structure that makes routing across networks easier to implement. Routing in DTNs is one of the key components in the architecture document. Based on different types of DTNs, deterministic or stochastic, different routing protocols are required. Due to intermittent connectivity, it is likely that paths to some of the destinations may not exist from time to time. When a packet arrives and its destination cannot be found in the routing table, the packet is simply dropped under the routing protocols mentioned above (developed with the assumption that the network is connected).

Therefore, these routing protocols will not work efficiently in DTNs. To cope with intermittent connectivity, one natural approach is to extend the store-and-forward routing to store carry- forward (SCF) routing. In store-carry-forward routing, a next hop may not be immediately available for the current node to forward the data. The node will need to buffer the data until the node gets an opportunity to forward the data and must be capable of buffering the data for a considerable duration. The difficulty in designing a protocol for efficiently and successfully delivering messages to their destinations is to determine, for each message, the best nodes and time to forward. If a message cannot be delivered immediately due to network partition, the best carriers for a message are those that have the highest chance of successful delivery, with the highest delivery probabilities. As ad hoc networks could be very sparse, SCF routing could mean that the node may have to buffer data for a long period of time. This condition can become worse if the next hop is not selected properly. A bad forwarding decision may cause the packets to be delayed indefinitely. If messages must be stored somewhere, a buffer management scheme should be proposed.

If all the future topology of the network (as a time-evolving graph) is deterministic and known, or at least predictable, the transmission (when and where to forward packets) can be scheduled ahead of time so that some optimal objective can be achieved. If the time-evolving topology is stochastic, SCF routing performs routing by moving the message closer to the destination one hop at a time. If the nodes know nothing about the network states, then all the nodes can do is to randomly forward packets to their neighbors. Protocols in this category are referred to as epidemic. If one can estimate the forwarding probability of its neighbors, a better decision could be made. Protocols in this category are referred to as history based or estimation-based forwarding. Furthermore, if the mobility patterns can be used in the forwarding probability estimation, an even better decision may be made. Protocols in this category are referred to as model-based forwarding. In some cases, network efficiency can be achieved if the movements of certain nodes are controlled and these protocols are in the category of controlling node movements. Recently, coding- based routing protocols have also been proposed for DTNs.

Some of the routing protocols in DTNs [Routing in intermittently connected mobile Ad hoc networks and delay tolerant networks: overview and challenges]:

• Deterministic case

–Space time routing

–Tree approach

–Modified shortest path approaches

• Stochastic case

–Epidemic/random spray

–History or predication-based approach

- Per contact routing based on one-hop information
- Per contact routing based on end-to-end information

–Model-based

–Control movement

–Coding-based approaches

4.1.1 DETERMINISTIC ROUTING

In deterministic routing we assume that future movement and connections are completely known (that is, the entire network topology is known ahead of time). A notion of evolving graphs, which consists of formalizing time domain graphs, is introduced. Modeling time in mobile ad hoc networks gives rise to several different matrices that may serve as objective functions in routing strategies, such as “earliest time to reach one or all the destinations” or “minimum hop paths” (with or without the condition that packets arrive before a predefined time period). Algorithms selecting the path of message delivering are depending on the available knowledge about the motion of hosts. Three cases are considered. In the first case it assumes that global knowledge of the characteristic profiles with respect to space and time (that is, the characteristic profiles of the motion and availability of the hosts as functions of time) are completely known by all the hosts. Paths are selected by building a tree first (the tree approach). Under the tree approach, a tree is built from the source host by adding children nodes and the time associated with nodes. Each node records all the previous nodes the message has to travel and the earliest time to reach it. A final path can be selected from the tree by choosing the earliest time (or minimum hop) to reach the desired destination. In the second case, it assumes that characteristic profiles are initially unknown to hosts. Hosts gain this information through learning the future by letting neighbor hosts exchange the characteristic profiles available between them. Paths are selected based on this partial knowledge. In the third case, to enhance the algorithm in the second case, it also requires hosts to record the past, that is, it stores the sequence of hosts a message has transited within the message itself. For DTNs, several routing algorithms are proposed in depending on the amount of knowledge about the network topology characteristics and traffic demand. They define four knowledge oracles; each oracle represents certain knowledge of the network. The Contacts Summary Oracle contains information about aggregate statistics of the contacts (resulting in time-invariant information). A contact is defined as an opportunity to send data. The Contacts Oracle contains information about contacts between two nodes at any point in time. This is equivalent to knowing the time-varying networks. The Queuing Oracle gives information about instantaneous buffer occupancies (queuing) at any node at any time. The Traffic Demand Oracle contains information about the present or future traffic demand. Based on the assumption of which oracles are available, the authors present corresponding routing algorithms. For example, if all the oracles are known, a linear programming is formulated to find the best route. If only the Contact Oracle is available, modified Dijkstra with time-varying cost function based on waiting time is used to find the route. All of the algorithms developed (except the zero-knowledge case) in are for the deterministic case. Assuming the characteristic profile is known over an infinite time horizon may not be realistic in ad hoc networks. If we assume that the characteristic profile can be accurately predicted over the time interval of T, we can model the dynamic of the networks as a space-time graph. Routing algorithms in the constructed space-time graph are developed using dynamic programming and shortest path algorithm. The routing algorithm finds the best route for messages by looking ahead. The idea of the time layers in the space-time graph comes from time-expanded graphs. The time expanded graphs approach translates a problem of network flow over time to a classical "static" network flow, and standard tools of graph theory such as the Floyd-Warshall algorithm can be applied to compute the shortest path for a source destination pair. In all these approaches under the deterministic case, an end-to-end path (possibly time dependent) is determined before messages are actually transmitted. However, in certain cases the topology of the network may not be known ahead of time (stochastic or random networks).

4.1.2 STOCHASTIC OR DYNAMIC NETWORKS

In this section, we review some of the routing protocols when the network behavior is random and not known. These protocols depend on decisions regarding where and when to forward messages. The simplest decision is to forward to any contacts within range, while other decisions are based on history data, mobility patterns, or other information.

4.1.3 EPIDEMIC (OR PARTIAL) ROUTING-BASED APPROACH

In the epidemic routing category, packets received at intermediate nodes are forwarded to all or part of the nodes' neighbors (except the one who sends the packet) without using any predication of the link or path forwarding probability. Epidemic routing is a natural approach when no information can be determined about the movement patterns of nodes in the system.

Vahdat and Becker propose an epidemic routing protocol for intermittently connected networks. When a message arrives at an intermediate node, the node floods the message to all its neighbors. Hence, messages are quickly distributed through the connected portions of the network. Epidemic routing relies on carriers of messages coming into contact with another node through node mobility. When two nodes are within communication range, they exchange pair-wise messages that the other node has not seen yet. Their simulation results show that, in the special scenarios considered, epidemic routing is able to deliver nearly all transmitted messages, while existing ad hoc routing protocols fail to deliver any messages because of the limited node connectivity when the buffer capacity is sufficiently large. Another extreme is to let the source hold the message and deliver to the destination only when they are within communication range. This approach obviously has minimal overhead, but the delay could be very long. Grossglauser and Tse propose a 2-hop forwarding approach and have explored a theoretical framework where nodes with infinite buffer move independently around the network, and every node gets close to any other node for some short time period per time slot. Within this framework, a node s gives a message addressed to node t to another randomly chosen node one hop away in the network, called a “receiver.” When the receiver happens to be within the range of the destination node t, the receiver sends the message to the destination. Hence, a message will only make two hops and no message will be transmitted more than twice. They prove that a message is guaranteed to be delivered, even if its delivery time is averaged over many time slots. This result sets a theoretical bound, since it assumes a complete mixing of the trajectories so that every node can get close to another one. In the Infostation model users can connect to the network in the vicinity of Infostations, which are geographically distributed throughout the area of network coverage. Infostations provide strong radio signal quality to small disjoint geographical areas and, as a result, offer very high rates to users in these areas. However, due to the lack of continuous coverage, this high data rate comes at the expense of providing intermittent connectivity only. Since a node that wishes to transmit data may be located outside the Infostations’ coverage areas for an extended period of time and must always transmit to an Infostation directly, large delays may result. Upon arrival in a coverage zone, the node can transmit at very high bit-rates. Thus, Infostations trade connectivity for capacity, by exploiting the mobility of the nodes. It is assumed that the Infostations are connected. Another model proposed is Shared Wireless Infostation Model (SWIM), where SWIM is a marriage of the Infostations concept with the (epidemic) ad hoc networking model. (Propagation of information packets within SWIM is identical to the epidemic routing protocol). The only difference is that any one of the many Infostations could serve as a destination node, while in the other case there is only one destination node for a given packet. One of the benefits of SWIM, by allowing the packet to spread throughout the mobile nodes, is that the delay for the replicas to reach an Infostation can be significantly reduced. However, this comes at a price: spreading the packets to other nodes consumes network capacity. Again, there is a capacity-delay tradeoff, which can be controlled by limiting the parameters of the spread, for example, by controlling the probability of packet transmission between two adjacent nodes, the transmission range of each node, or the number and distribution of the Infostations. A relay-based approach to be used in conjunction with traditional ad hoc routing protocols can be proposed. This approach takes advantage of node mobility to disseminate messages to mobile nodes. The result is the Mobile Relay Protocol (MRP), which integrates message routing and storage in the network. The basic idea is that if a route to a destination is unavailable, a node performs a controlled local broadcast (a relay) to its immediate neighbors (that is the only time that broadcast is used in the protocol). All nodes that receive this packet store it and enter the relaying mode. In the relaying mode, the MRP first checks with the (traditional) routing protocols to see if a route of less than d hops exists to forward the packet. If so, it forwards the packet and the packet is delivered. If no valid route exists for the packet, it enters the storage phase, which consists of the following steps:

• If the packet is already stored in the n de’s buffer, then the older version of the packet is discarded.

• Otherwise, the node buffer is checked. If it is not full, then the packet is stored and the time-to-live parameter h in the MRP header of the packet is decremented by 1.

• If the buffer is full, then the least recent packet is removed from the buffer and it is relayed to a single random neighbor if h > 0. In a network with sufficient mobility, it is quite likely that one of the relay nodes to which the packet has been relayed will encounter a node that has a valid, short (conventional) route to the eventual destination, thereby increasing the likelihood that the message will be successfully delivered. To limit the amount of broadcasting to all its neighbors, the Spraying protocol restricts forwarding to a ray in the vicinity of the destination’s last known location. It is assumed that the destination’s last location is known and there is a separate location manager in the system. To deal with high mobility, spray routing multicasts traffic within the vicinity of the last known location of a session’s destination. The idea is that even though a highly mobile node may not be in the location last reported by the location tracking mechanism, it is likely to be in one of the surrounding locations. By “spraying” to the vicinity of the last-known location of the destination, the algorithm attempts to deliver packets to the destination even if it moves to a nearby location during the location tracking convergence time. A sprayed packet is first unicast to a node close to the destination, and then multicast to multiple nodes around the destination. The magnitude of the spraying depends on the mobility; the higher the mobility, the larger the vicinity. Upon a change in affiliation, a node sends a location update to its location manager. In order to communicate with a destination node, D, a source node sends a location subscribe to the location manager. The current location and changes (in location) thereafter are sent by the location manager to the source using a location information message. Note that it is possible that the destination receives duplicate packets, and it is assumed that there is an end-to-end duplicate detection mechanism that will discard such packets. Spray routing is an integrated location tracking and forwarding scheme. Both location managers and switches/routers participate in spray routing. Using epidemic propagation we can observe that nodes exhibit signs of regularity and affinity of contact. Furthermore, in many cases, success of message delivery from any source to a destination is not evenly distributed among the intermediate nodes. Thus, source nodes can potentially use this information for better routing decisions. During tests, power management is one of the critical issues.

Protocol	Control nodes trajectory	Node movement patterns known	Special nodes' movement if needed	Remarks	Year of Pub.
MV [35]	No	No	Controlled	Extra nodes needed	2005
Li&Rus-1 [34]	Yes, intermediate nodes change their trajectories	Completely known	Not needed	The network is assumed to be almost connected; MST is not a realistic assumption in Li&Rus-2	2003
Li&Rus-2 [34]	Yes, based on information learned from updates	No	Not needed; MST used for updates		2003
NIFM [36]	Yes, node trajectory controlled (move closer to MF)	No	Predetermined		2004
FIFM [36]	No	No	Based on user's request (via long-range radio)	Long range radio is not a realistic assumption?	2004
VMN [53]	No (nodes wait to get closer to VMN)	No	Predetermined		2004
Snake [38]	No	No	Special nodes form a snake		2001
Runners[39]	No	No	Random walk		2001

Fig. 26: Summary and comparison of the proactive approaches

Protocol	Buffer management (if buffer size is finite)	Estimation of link forwarding prob.	Use location or future information	Remarks	Complexity	Year of pub.
Epidemic [20]	Infinite	No	No	First work in DTN routing	Simple	2000
SWIM [22]	Infinite	No	No	Similar with [20], the probability to forward on a link can be controlled	Simple	2003
MRP [23]	Remove least recent (forwarding it randomly to one)	No	No	Broadcast once, then randomly forward to one node if buffer is full	Simple; integrate with existing routing protocols	2003
Spraying [24]	Infinite	No	Location	Forwarding once, then spraying	Simple; central location tracking,	2001
DLE [26]	DOA, DRA, DLE, DLR	Yes		Forwarding packets only upon ***request*** from next hop nodes	Exchange and computation	2001
PROPHET [27]	Infinite	Yes	No	Require two neighboring nodes exchanging routing information,	Exchange and computation	2003
MV [35]	Yes, delete those packets other nodes having higher probability	Yes, calculate the probability of delivering the bundle in h hops to destination		Need to estimate the likelihood two nodes meet	Exchange and computation	2005
SEPR [30]	Yes, remove those packets with smaller EPL	Yes, through measurement	No		Computation only	2003
CAR [28]	Infinite	Yes, Kilman filter	No	Integrate with DSDV	Computation only	2005
MBR [32]	Infinite	User profile assumed given	Yes, both	No details, only sketch of the algorithm presented	Exchange and computation; central location service	2001
IBRR [29]	Infinite	Yes, require two nodes exchange	Yes, both	Requires two-hop information exchange	Exchange and computation; location tracking, velocity, mobility	2002
Erasure coding [41]	Infinite	No	No	Source nodes enlarge the packets to be sent	Requires extra computation for decoding	2005

Fig. 27: Summary and comparison of the reactive approaches

4.2 Integrating DTN and MANET Routing

Mobile Ad-hoc Network (MANET) routing protocols aim at establishing end-to-end paths between communicating nodes and thus support end-to-end semantics of existing transports and applications. In contrast, DTN-based communication schemes imply asynchronous communication (and thus often require new applications) but achieve better accessibility, particularly in sparsely populated environments. A hybrid scheme [Integrating DTN and MANET Routing] that combines a MANET protocol (like AODV) and DTN-based routing and allows keeping the AODV advantage of maintaining end-to-end semantics whenever possible while, at the same time, also offering DTN-based communication options whenever available—leaving the choice to the application.

While the “classical” MANET protocols seek to establish an end-to-end path between the communicating peers and thus enable higher layer end-to-end protocols to operate largely unchanged, the need for operation in mobile ad-hoc networks with sparse node distribution has led to the development of specific application protocols that work using asynchronous communications and do not require an end-to-end path at any point in time. Different approaches have been studied for sensor networks, interpersonal communication, and Internet access in remote areas but also in support of mobile Internet access, among others. Rather than exchanging (small) packets end-to-end, DTN endpoints use messages of arbitrary size (bundles) forwarded by DTN routers hop-by-hop from the source to the destination. The DTN bundle protocol operates above the transport layer and may interconnect different internets with arbitrary underlying protocol stacks. Status reports may convey information about the delivery progress of a bundle from intermediate routers as well as receipts from the receiver to achieve end-to-end semantics. DTN endpoints and applications are identified by endpoint identifiers (EIDs) specified in a URI-style format: scheme: specific-address. The scheme defines the scope within which the specific address is interpreted.

DTN routing often relies on information replication for forwarding bundles to maximize delivery probability and/or minimize transit time. While the issue of node mobility and non-available end-to-end paths has been recognized for MANET environments and DTN routing schemes for mobile networks are being studied and improved, the present approaches have in common that they take either a pure end-to-end packet-based approach to communications or use exclusively asynchronous hop-by-hop information exchange. A solution is using schemes coupling DTN concepts and opportunities for end-to-end communications.

Even a DTN-capable application may prefer end-to-end communication whenever possible: for better performance, for preserving end-to-end semantics with immediate responses, or simply because some interactions cannot be performed using DTN. But asynchronous DTN-based communication may be considered a suitable backup in many cases. Applications should therefore be notified about the available communication options which can be determined when attempting to obtain a route to the target (via AODV). Such notifications should comprise possible paths for end-to-end communications (plus their hop counts) and information about intermediate DTN nodes (including hop counts and DTN routing metrics) so that the application can make a conscious choice (autonomously, based upon user policy, or after inquiring the user). End-to-end communication (if available) may be chosen if the number of wireless hops and the expected route stability are deemed suitable for the size of the message to be transmitted.

Otherwise, the application may opt for hop-by-hop communications (and provide hints about possible next hops to the DTN subsystem) or decide that the intended transaction is not suitable for asynchronous interaction, report failure, and retry later. The same applies to DTN routing (considering a DTN router such as dtnd as an application). Like an application trying to reach its peer, a DTN router attempts to find and contact peers for bundle forwarding. For this purpose, the DTN router may rely on link-specific adaptation layers or it may use the TCP convergence layer. While the former may limit interactions to directly adjacent nodes, the latter allows reaching a peer across an arbitrary number of IP hops. However, if a known peer is many wireless hops away, performance degradation and route instability may make direct ("end-to-end") communication infeasible. The hop count to the target peer (or its present non-reachable) can be learned from the AODV route search. This process may also yield closer available DTN routers (of which those more than one hop away would not be found by link layer interactions only) which can then be used by the DTN router for its forwarding decision. Both classes of applications can be supported by minor extensions to AODV (and possibly other ad-hoc routing protocols) and a minimal integration with DTN routing, in the DTN router and/or the application, as we present in the next section.

DTN and AODV fulfill complementary and independent functions and operate at different layers: AODV dynamically determines and maintains an end-to-end path between two peers wanting two communicate in MANETs while DTN is typically used when the existence of such a path is rather unlikely so that using asynchronous communication is advisable. When running DTN on top of IP, DTN communication exploits the multi hop end-to-end routing between two DTN nodes. However, as discussed, DTN nodes do not depend on underlying multi hop communication capabilities but may also perform the respective routing functions hop by- hop at the link layer. Nevertheless, the availability of multi hop routing from lower layers appears beneficial because it essentially enhances the view of the network for DTN nodes and thus their potential reach—which is likely to increase the number and diversity of possible next DTN hops for bundle forwarding.

Our goals to extend the view of DTN nodes and thus allow for greater flexibility in routing decisions, to enable communicating asynchronously with application peers otherwise out of reach, and to allow applications—but also DTN routers—to dynamically trade off immediate interaction for accessibility.

If an originating node S wants to communicate with a target node D. Ideally, such an interaction should be carried out end-to-end between S and D without involving intermediate nodes with more than plain IP routing. As noted, AODV will initiate route discovery on demand when a packet needs to be transmitted to a target IP node and no valid cached route is available. During the route discovery process, an AODV node performs an expanding ring search for the target node in its surrounding territory during which RREQs are processed and forwarded by all directly and indirectly neighboring nodes up to a maximum diameter or until the target node is found.

We augment this route search process to have each AODV node on the path note—in the request and later in the reply—whether it supports DTN routing and, optionally, report its DTN routing metrics for the target (and originator) node. Since this is part of the AODV route search, no additional messages are needed; also, caching and route invalidation can work for DTN-related routing information as they do for AODV paths. This essentially augments AODV route discovery to become an implicit service discovery mechanism for DTN routers. DTN-specific timeout handling ensures that route replies containing DTN routers are returned even if the final target cannot be reached. Caching and timeout mechanisms for the specific DTN extensions prevent the network from being flooded with repeated DTN route information. When the originator receives the results of the route discovery, it may or may not find AODV routes to the target as well as zero, one, or more available DTN routers in reach. The originator then decides whether to use end-to-end IP communication, to go for hop-by-hop DTN messaging, or to declare failure for this point in time (with the option of retrying later as deemed acceptable by the application). If DTN is used, the originator evaluates the obtained DTN routing information from the route discovery to determine a (set of) suitable next hop(s). Note that we assume that a DTN routing protocol is in place that is capable of taking the appropriate forwarding decisions based on the accessibility information supplied by our AODV extensions.

Some extensions are needed to be made for AODV. Three types of processing extensions are required to support DTN route(r) discovery in AODV: the route discovery needs to be expanded (the header fields as well as generation and processing of RREQs and RREPs); an additional timeout mechanism is required to deal with unsuccessful route requests; and specific error handling needs to be introduced to deal with route failures. Each AODV node has to maintain additional state information per AODV route during the search phase and after its establishment.

AODV can be used as a vehicle to locate DTN routers as possible optimization for communication or fallback in case no end-to-end path can be determined. In the following, we consider two cases that essentially differ in how much information from DTN routing is available to and communicated via AODV and how much the DTN routing algorithms can exploit the knowledge about underlying AODV routes. AODV can also be replaced in the future by other ad-hoc protocols for which similar extensions can be defined.

4.3 Optimized QOS Protocols for MANETs

The dynamic nature of the ad hoc networks makes Quality of Service (QoS) support in Mobile Ad hoc Networks (MANET) [Optimized QOS Protocols for MANETs] a challenging and difficult task where nodes may leave and join the network or move around anytime. A mechanism for resources allocation/reservation has to be implemented to provide the needed assurance for the desired level of quality. Due to the bandwidth constraint and dynamic topology of MANET, supporting QoS in MANETs is a challenging task.

A lot of work has been done in supporting QoS in the Internet, but unfortunately none of them can be directly used in MANETs because of their special constraints. To support QoS, the link state information such as delay, bandwidth, cost, loss rate, and error rate in the network should be available and manageable. However, getting and managing these in MANETs is very difficult because of the resource limitations and the complexity associated with the mobility of hosts.

First, a QoS model should specify an architecture in which some kinds of services could be provided in MANETs. All other QoS components such as QoS signaling, QoS routing, and QoS MAC must cooperate together to achieve this goal.

We must first find out what is feasible for supporting QoS in MANETs because it will influence the functionality of all other QoS components.

Second, QoS signaling in MANETs act as the control center in its support. It coordinates the behaviors of QoS routing, QoS MAC, and other components such as admission control and scheduling. The functionality of QoS signaling is determined by the QoS model.

Third, QoS routing searches for a path with enough resources but does not reserve resources.

It is the QoS signaling to reserve resources (if necessary in the QoS model) along the path determined by QoS routing or other routing protocols. Hence, QoS routing enhances the chance that enough resources can be guaranteed through QoS signaling. QoS signaling will work better if it coordinates with QoS routing.

However, since most QoS routing algorithms are complicated, we must balance the benefits against the cost of QoS routing in the bandwidth constrained MANETs.

Fourth, the QoS MAC protocol is an essential component in QoS support in MANETs. All upper-layer QoS components (QoS routing and QoS signaling) are dependent on it and coordinate with the QoS MAC protocol.

Finally, other QoS components in MANETs, such as scheduling and admission control, can be borrowed from other network architectures without or with few modifications.

FACTORS AFFECTING QOS MODELS FOR MANETS

The QoS Model specifies the architecture which covers services that operate better than the current "best effort" model that exists in MANETs. This architecture should take into consideration the challenges of Mobile Ad Hoc Networks. Therefore, the QoS model for MANETs should also consider the existing QoS models for the Internet: IntServ and DiffServ which are aimed for wired networks. The applicability of IntServ and DiffServ in MANETs therefore needs investigation. IntServ provides quantitative QoS for every flow. However, such per-flow provisioning results in a huge storage and processing overhead on routers. The precept of DiffServ is to use a relative-priority scheme to soften the requirements of hard QoS models like ATM and IntServ, thereby mitigating the scalability problem of the latter. DiffServ provides qualitative QoS support for aggregate flows.

Each node in MANETs has two-fold functions, i.e., host and router / switch. Hence, a MANET is similar to a backbone network in the sense of the functionalities of nodes, although the size of a MANET is not comparable with that of the Internet backbone. In MANETs, keeping the protocol lightweight in interior nodes is important since putting too heavy a load on a temporary forwarding node which is moving is unwise. Node mobility and non-infrastructure: Node mobility is the basic cause of the dynamics in MANETs. The MAC layer allocation of bandwidth to each node changes dynamically according to mobility scenarios. Consequently, the aggregate network capacity is also time varying and the feature of non-infrastructure makes the control of bandwidth difficult. The roles of nodes as a host or router change too together with node mobility and the dynamic topology. All these dynamics in MANETs require the QoS model and the supporting QoS mechanisms to be adaptive.

Power constraint: In MANETs, nodes processing capability is limited because of limited battery power.

This feature requires low processing overheads of nodes. The control algorithms and QoS mechanisms should use bandwidth and energy efficiently and the inter router information exchange to a minimum. Limited bandwidth and network size: As fast radios and efficient low bandwidth compression technology is developing rapidly, the emergence of high speed and large sized MANETs with plenty of applications is foreseeable.

Time-varying feature: In MANETs, the link capacity is time-varying due to the physical environment of nodes like fading and shadowing, the mobility of nodes and the dynamics of the network topology. This time varying feature makes the QoS mechanisms in MANETs more difficult than in wired fixed networks. This makes it hard to provide short timescale QoS by trying to keep up with the conditions.

4.3.1 QOS MODELS FOR MANETS

4.3.1.1 IntServ/RSVP on Wired Networks

The basic idea of the Integrated Service (IntServ) model is that the flow-specific states are kept in every IntServ-enabled router. A flow is an application session between a pair of end users. A flow-specific state should include bandwidth requirement, delay bound, and cost of the flow. In addition to Best Effort Service, IntServ proposes two service classes, Guaranteed Service and Controlled Load Service. The Guaranteed Service is provided for applications requiring fixed delay bound. The Controlled Load Service is for applications requiring reliable and enhanced best effort service. Because every router keeps the flow state information, the quantitative QoS provided by IntServ is for every individual flow. In an IntServ-enabled router, IntServ is implemented with four main components: the signaling protocol, the admission control routine, the classifier, and the packet scheduler. The Resource Reservation Protocol (RSVP) is used as the signaling protocol to reserve resources in IntServ. Applications with Guaranteed Service or Controlled-Load Service requirements use RSVP to reserve resources before transmission. Admission control is used to decide whether to accept the resource requirement. It is invoked at each router to make a local accept/reject decision at the time that a host requests a real-time service along some paths through the Internet. Admission control notifies the application through RSVP if the QoS requirement can be granted or not and then its data packets are transmitted. When a router receives a data packet, the classifier will perform a Multi-Field (MF) classification, which classifies a packet based on multiple fields such as source and destination addresses, source and destination port numbers, Type of Service (TOS) bits and protocol ID in the IP header. IntServ/RSVP model is not suitable for MANETs due to the following resource limitations in MANETs: 1) The amount of state information increases proportionally with the number of flows. Keeping flow state information will cost a huge storage and processing overhead for the mobile host whose storage and computing resources are scarce. 2) The RSVP reservation and maintenance process is a network consuming procedure. Thus RSVP signaling packets will struggle with the data packets for resources and more specifically for bandwidth. 3) In order to have a complete QoS Model mechanism, classification and scheduling must be provided. These mechanisms again require a respectable amount of network resources which are usually not available in MANETs.

4.3.1.2 DiffServ

Differentiated Service (DiffServ) is designed to overcome the difficulty of implementing and deploying IntServ and RSVP in the Internet backbone. DiffServ provides a limited number of aggregated classes in order to avoid the scalability problem of IntServ. DiffServ defines the layout of the Type of Service (TOS) bits in the IP header, called the DS field, and a base set of packet forwarding rules, called Per- Hop-Behavior (PHB). At the boundary of a network, the boundary routers control the traffic entering the network with classification, marking, policing, and shaping mechanisms. DiffServ may be a possible solution to the MANET QoS model because it is lightweight in interior routers. In addition, it provides Assured Service, which is a feasible service context in MANET. However, since DiffServ is designed for fixed wire networks, we still face some challenges to implement it in MANETs. First, it is ambiguous as to what the boundary routers in MANETs are. Intuitively, the source nodes play the role of boundary routers. Other nodes along the forwarding paths from sources to destinations are interior nodes. But every node should have the functionality as both boundary router and interior router because the source nodes cannot be predefined. This arouses a heavy storage cost in every host. Second, the concept of Service Level Agreement (SLA) in the Internet does not exist in MANETs. The SLA is a kind of contract between a customer and its Internet Service Provider (ISP) that specifies the forwarding services the customer should receive. The SLA is indispensable because it includes the whole or partial traffic conditioning rules, which are used to re-mark traffic streams, discard or shape packets according to the traffic characteristics such as rate and burst size.

DiffServ on the other hand is a lightweight model for the interior routers since individual state rows are aggregated into a set of rows. This makes routing easier in the core of the network.

Thus this model could be a potential optimized model for MANETs. In MANETs there is no clear definition of what is a core, ingress or egress router because of the dynamic topology of the network.

4.3.1.3 FQMM

It is labeled as a Flexible QoS Model for MANET (FQMM). It considers the characteristics of MANETs and tries to take advantage of both the per-flow service granularity in IntServ and the service differentiation in DiffServ. As in DiffServ, three kinds of nodes (ingress, interior, and egress) are defined in FQMM. An ingress node is a mobile node that sends data. Interior nodes are the nodes that forward data for other nodes. An egress node is a destination node. The role of a mobile node is adaptively changing based on its position and the network traffic. The provisioning in FQMM, which is used to determine and allocate the resources at various mobile nodes, is a hybrid scheme of per-flow provisioning as in InterServ and per-class provisioning as in DiffServ. FQMM tries to preserve the per-flow granularity for a small portion of traffic in MANET, given that a large amount of the traffic belongs to per aggregate of flows, that is, per-class granularity. FQMM is the first attempt at proposing a QoS model for MANETs. However, some problems still need to be solved. Without an explicit control on the number of services with per-flow granularity, the scalability problem still exists. Secondly, just as in DiffServ, the interior nodes forward packets according to a certain PHB that is labeled in the DS field. It is difficult to code the PHB in the DS field if the PHB includes per flow granularity considering the DS field which is at most 8 bits without extension. Finally, making a dynamically negotiated traffic profile is very difficult.

4.3.1.4 SWAN

SWAN (Service differentiation in wireless ad hoc networks) model for QoS in MANETs has been designed by the SWAN group. SWAN considers TCP traffic as best-effort traffic and UDP traffic as real-time traffic requiring QoS assurances. It tries to maintain delay and bandwidth requirements of UDP real time traffic by admission control of UDP traffic and rate control of both. Each node has an admission controller which estimates the availability of resources in the local neighborhood. To regulate the traffic and to maintain the bandwidth and delay requirements of the admitted real-time traffic, rate of the TCP best-effort traffic is controlled using MAC delay measurements as feedback. A classifier and a shaper situated between the IP layer and the MAC layer differentiate the best effort traffic from real time traffic, such that the real time traffic passes directly to the MAC layer for transmission, while the best effort traffic is shaped and delayed according to rate set by the rate controller. Dynamics of the network may however cause violations of the admitted real time traffic, thus degrading the quality of some of the flows. SWAN can provide only soft QoS assurances. An admitted real time flow is not guaranteed to meet its requirements for the entire duration of the connection.

4.3.1.5 INSIGNIA

INSIGNIA is an in-band signaling system that supports QoS in MANETs. It is claimed to be the first signaling protocol for MANETs. The signaling control information is carried in the IP option of every IP data packet called the INSIGNIA option. Like RSVP, the service granularity supported by INSIGNIA is per-flow management. Each flow state information is established, restored, adapted and removed over an end-to-end session in response to topology change and end-to-end QoS condition. The packet forwarding module classifies the incoming packets and forwards them to the appropriate modules. If a received IP packet includes an INSIGNIA option, the control information is forwarded to and processed by the INSIGNIA module. In the meantime, the received packet is delivered to a local application or forwarded to the packet scheduling module according to the destination address in the IP head. If the mobile host is the destination of the packet, the packet is processed by a local application. Otherwise it will forward the packet to the next hop determined by the MANET routing protocol. Before the packets are sent through the MAC component, a packet scheduling module is used to schedule the output of the flows in order to fairly allocate the resource to different flows. In INSIGNIA, a Weighted Round-Robin (WRR) discipline that takes location dependent channel conditions into account is implemented. The INSIGNIA module is responsible for fast flow reservation, restoration and adaptation algorithms that are specifically designed to deliver adaptive real time service in MANETs. INSIGNIA uses QoS reports to inform the source node of the status of the real-time flows. The destination node actively monitors the received flows and calculates QoS statistical results such as loss rate, delay, and throughput etc. The QoS reports are periodically sent to the source node. Through this kind of feedback information, the source node can take corresponding actions to adapt the flows to observed network conditions. As a whole, INSIGNIA is an effective signaling protocol for MANETs. Coordinating with other network components INSIGNIA can efficiently deliver adaptive real-time flows in MANETs. However, since the flow state information should be kept in the mobile hosts, the scalability problem may hinder its deployment in the future.

4.3.1.6 MRSVP

Since RSVP does not adequately support nodes mobility, a new protocol MRSVP (Mobile RSVP) was developed to address host mobility in mobile networks. MRSVP takes nodes mobility into account, where MANET nodes can provide information about their movement directions to the protocol, the protocol then uses this information to reserve resources in those locations where the mobile node is expected to visit with its active flow. Although MRSVP provides better performance than RSVP in the MANET environment, the protocol still shares with RSVP the scalability problem and high processing power requirement in each node. However, MRSVP provides a better chance (higher probability, but no guarantee) in meeting the QoS requirement as specified by the application at the allocation time. MRSVP fits that type of MANET applications where node mobility and movement directions can be predetermined or estimated in advance.

4.3.1.7 DRSVP

In DRSVP (Dynamic RSVP), the dynamic nature of the network, rather than the mobility of the nodes are taken into consideration. A new resource reservation approach based on expanding the semantic of the reservation is developed. Instead of the semantic being a single value indicating the level of service needed by the application, it becomes a range of service levels. This provides high flexibility in reserving resources while network conditions are changing with time. This technique makes better use of the resources available in each node by allowing resources to be redistributed dynamically. Instead of forcing network to make a binary decision (admit/fail) for each flow, the network provides feedback to applications with the current available resource limits, the application can then adapt its behavior to run with the level the network can support. Like MRSVP, DRSVP provides better performance than RSVP in the MANET environment, like RSVP it is not well scalable and requires high processing power requirement in each node. Yet, DRSVP provides a better chance in meeting the flexible QoS requirement as specified by the application. Unlike MRSVP, that takes only nodes movements into consideration, DRSVP takes into consideration the overall dynamic nature of MANET (variation in wireless links and node mobility). Therefore, DRSVP fits that type of MANET applications where nodes mobility is relatively moderate and their movement patterns are random by nature.

4.3.1.8 DSRRSVP

This is a very simple form of RSVP, which uses the DSR (Dynamic Source Routing) protocol. The advantage of DSR is that with a slight modification, it can provide the back track functions to RSVP by discovering all the possible paths between any given pair of source and destination. With a slight modification to DSR, it can provide RSVP with the back track function. In this approach bandwidth reservation can be done at each candidate node.

Chapter 5

5. Simulations on DTN

5.1 The Opportunistic Network Environment simulator

To make complex DTN simulations more feasible and understandable, we use The Opportunistic Network Environment simulator (ONE) [Opportunistic Network Environment simulator] that combines movement modeling, routing simulation, visualization and reporting in one program. Movement modeling can be done either on-demand using the integrated movement models or movement data can be imported from an external source. The node position data that the movement models provide is then used for determining if two nodes can communicate and exchange messages. This information can be exported for routing simulation in external simulators or it can be given to the internal routing modules which implement several different DTN routing algorithms. The internal routing modules perform operations on the messages on their own, but they can also be commanded using event generator modules or external traces. The movement modeling and routing simulation is interactively observable in the simulator's GUI and report modules can gather data of the simulation for further analysis or interaction with other programs. The core of the ONE is an agent-based discrete event simulator.

The simulations can contain any number of different types of agents, for example, wireless nodes. The nodes are grouped in node groups and a one group shares a set of common parameters such as message buffer size, radio range and mobility model. Since different groups can have different configurations, creating a simulation with pedestrians, cars and public transportation is possible. All movement models, report modules, routing algorithms and event generators are dynamically loaded into the simulator so extending and configuring the simulator with different type of plug-in is made easy for users and developers: just creating a new class and defining its name in the configuration file is usually enough.

5.1.1 Mobility modeling

Mobility models dictate how the nodes move during the simulation. Three different types of mobility models were initially implemented for ONE. For reference purposes, even despite of its shortcomings, ONE includes the basic Random Waypoint movement model. For more realistic mobility scenarios ONE provides variety of map-based movement models which constrain the node movement to predetermined paths. Finally, ONE also supports importing of mobility data from external sources. Map-based movement models accept map data that is described using a subset of the Well Known Text (WKT) format. In map-based models, the nodes move using only the roads and walkways of the map area. Different node groups can be configured to use only certain parts of the maps which can prevent, e.g., cars driving on pedestrian paths or inside buildings (if paths in buildings are also defined). The simplest Map-Based Model, MBM, places nodes randomly in the map area and moves them forward on the path segments until they hit the end of the road and turn back or end up in an intersection. In intersections, nodes using MBM select randomly a new direction but do not head back where they came from. When a node has traveled a configurable distance, it stops for a (configurable) while, and then continues its journey.

The more advanced version of MBM, Shortest Path Map-Based movement Model, SPMBM, so initially places the nodes in random places but selects a certain destination in the map for all nodes and uses Dijkstra's shortest path algorithm to find the shortest path to the destination. When the destination is reached, the node waits for a while and selects a new destination. Normally all the places in the map have equal probability of being chosen as the next destination, but the map data can also contain Points of Interest (POIs). POIs are grouped in POI groups and every node group can have configurable probability for choosing POI in certain group as the next destination. Some nodes can also have predetermined routes that they travel on the map. This kind of Route-Based Models, RBMs, are good for simulating e.g., bus and tram routes. Routes consist of map points that model the stops on the route and nodes wait on every stop for some time before continuing, using the shortest path, to the next stop.

5.1.2 Routing simulation

While the mobility models decide where the nodes should move next, the routing modules get to decide where the messages, or bundles, end up. The ONE has six implementations of different well known routing algorithms and also a passive routing module that can be used for interaction with external DTN routing simulators.

The active routing modules included in the ONE are: First Contact, Direct Delivery, Spray and Wait (normal and binary), Epidemic, PRoPHET and MaxProp. When two (or more) nodes meet and there is a chance to exchange messages, all of the routing modules first check if they have any messages that are destined for the other node and try to send them. If the message was already earlier received by the node, it declines receiving it and other messages can be tried. After all, if any, such messages are exchanged, the behavior with rest of the messages depends on the routing algorithm.

The simplest routing modules are the Direct Delivery, Epidemic and First Contact. The Direct Delivery module does not start any further transactions after exchanging the deliverable messages since it will send messages only if it is in contact with the final recipient. This saves buffer space and bandwidth but is not obviously an optimal approach in many cases if a high message delivery probability is the goal. The Epidemic routing module uses quite different approach since after two nodes have exchanged the deliverable messages, it tries to exchange also all the other messages until both nodes have the same set of messages or the connection breaks. If we had unlimited buffer space and bandwidth, this would result in maximal spreading of the messages throughout the nodes and therefore also to maximal delivery probabilities. However, if the buffer space and/or bandwidth are limited, Epidemic routing is also likely to waste a lot of resources. For example, a message that was delivered shortly after it was created could be transmitted and stored by all the nodes for a long time even if there is no longer use for it. This way the message is using resources that could be better used for the messages that are not yet delivered. While also the First Contact module sends as many messages to the other node as it has time, it removes the local copy of the message after a successful transfer. This results in only a single copy of every message in the network. To prevent two nodes who stay in contact for a long time exchanging the same messages back and forth, the receiving node accepts a message only if the message has not passed through it before. Unfortunately, there are no guarantees that the first node that is met is a better candidate than the previous node carrying the message, so First Contact is not likely to achieve very high delivery probabilities either. The Spray and Wait works a bit like the Epidemic but is a bit more complex routing module since it restricts the amount of copies that are spread in the network. This is done by letting each created message to replicate only a certain amount of times. A node that has more than one copy of the message left, can give either a one copy to another node (the normal mode) or half of the copies (the binary mode). If the node has only a single copy of the message left, it is transmitted only to the final recipient. By using a different amount of initial copies, Spray and Wait can balance between high diffusion of messages and excess use of resources. PRoPHET and MaxProp take the complexity a bit further since they keep track of which node has been in contact with which nodes. This information can be used to reason if a certain node is a good candidate for delivering a message to another node by assuming that if two nodes have met before, they are more likely to meet (soon) again. While PRoPHET Checks if another node is more likely to meet the final recipient, MaxProp takes this idea further and uses Dijkstra's algorithm to calculate whole paths from node to node using the meeting probabilities. MaxProp also uses acknowledgments of delivered messages that help flushing redundant messages from the network. The passive routing module can be used as an interface to other DTN routing simulators.

5.1.3 GUI and Reports

In the GUI mode the simulation is visualized in real time. The largest part of the GUI is taken by the play field view which contains a view of the geographical simulation area. Node locations, their radio range, current paths, amount of messages etc. are visualized on the play field view. If the current movement model is map based, also the map path segments are drawn in the view. Additionally, a background image, such as an aerial photograph or a raster map, can be displayed under the other graphics. The play field view can be centered by clicking with a mouse button and zoomed with the mouse wheel. The simulation speed can be adjusted with the controls in the upper part of the GUI and any node can be selected for inspection with the node selector panel on the right side. When a node is selected for inspection, its current location and amount of messages is displayed. Any of the messages the node is carrying can be taken for further inspection and also routing module specific information can be retrieved. The GUI also keeps track of notable events and displays them in the event log panel. By clicking a node or message name in the log panel, more information will be shown about that node or message. The event log controls panel can be used to adjust which events are shown in the log and the simulation can also be made automatically pause in case of some type of event.

When a new movement model or routing module is created, the GUI can give an intuitive view on how it works and whether the operations make much sense. Since node locations are shown on the map, by following their movement for a while, the movement model's creator can see if the nodes move like they were supposed to move and end up in places where they were expected to go. This also applies to the routing modules: the simulation can be automatically paused when, e.g., a message is transferred between two nodes. Then, the routing module's state can be inspected to see if the message was really supposed to be transferred or not. During the development of the ONE, these methods were extensively used to test and debug the simulator.

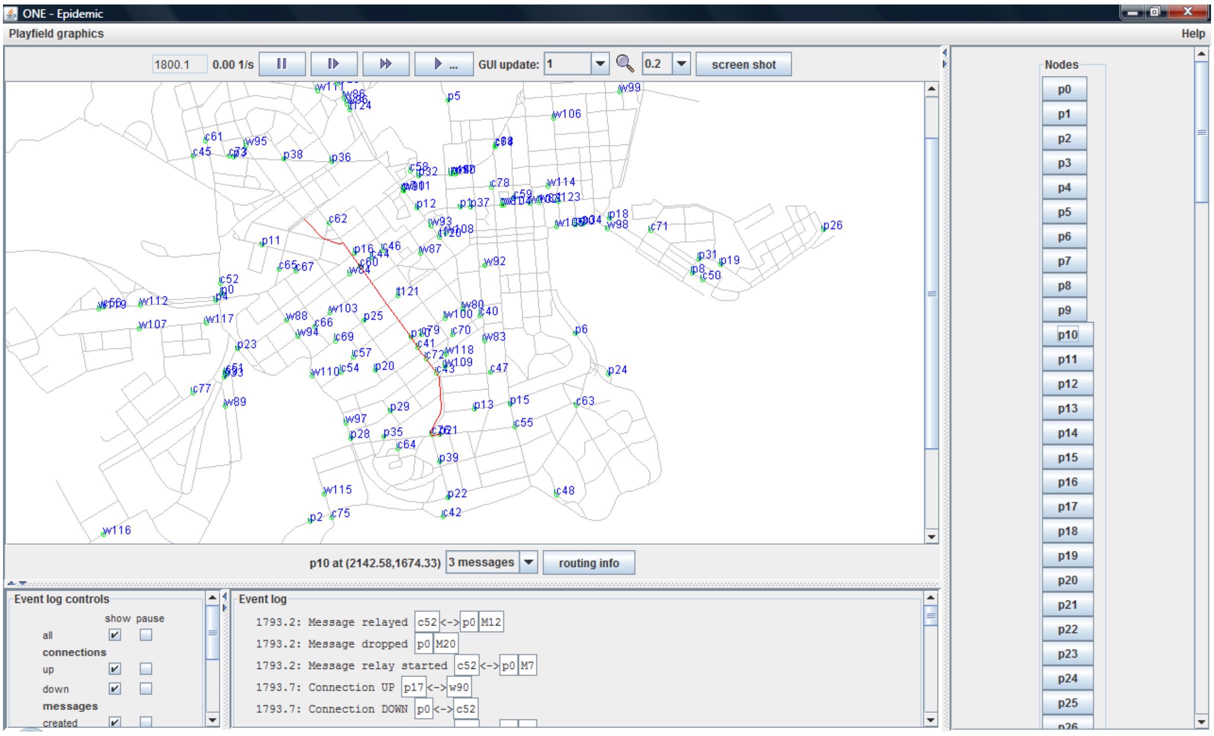
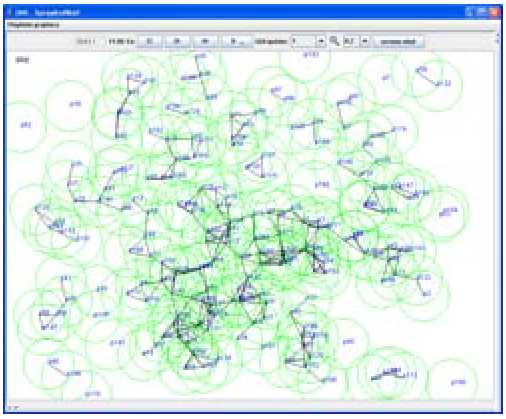


Fig. 28: Simulator running



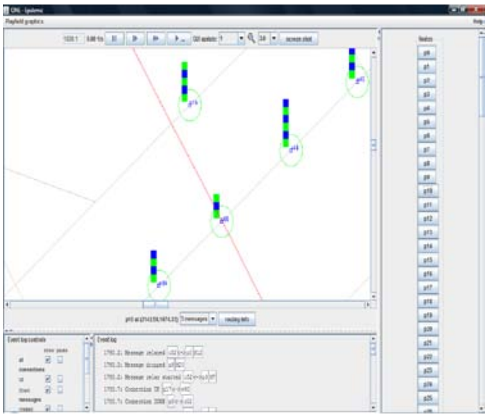


Fig. 29: 200 nodes, 250m radio range Fig 30: Zoomed view of simulation

While the external events provide input for the simulation, the most important output is generated by the report modules. Report modules can register to connection, message, and/or movement related events. Whenever such an event happens in the simulation, the related method is called for all registered report modules and a reference to the relevant objects (nodes and messages) is given to the report module. The report modules usually either write information about the event to a report output file if it was relevant or just store the information to an internal data structure for creating a summary when the simulation is done.

The ONE has already numerous reporting modules, for example, for message statistics (such as delivery probability and round-trip times), node contact and inter-contact times and message delivery delays and distances. Other interesting report modules are the ones used for communication with other programs. When a set of report files have been created, either with GUI or batch mode, they can be further post-processed with suitable programs. Two of the ONE's ready-made report modules produce files that are directly suitable for GRAPHVIZ input. The adjacency graphviz report module creates node adjacency graphs. In the graph the nodes that have been in contact more than once have an edge drawn between them and nodes that have been more often in contact are drawn closer to each other. Another Graphviz suitable module is the message graphviz report module which produces directed graphs of the delivered messages' paths.

The ONE includes also a set of post-processing scripts that can automatically create, for example, bar graphs of message delivery probabilities or plot cumulative distribution of inter-contact times using GNUPLOT.

5.1.3.1 AdjacencyGraphvizReport

AdjacencyGraphvizReport generates Graphviz compatible graph from connections.

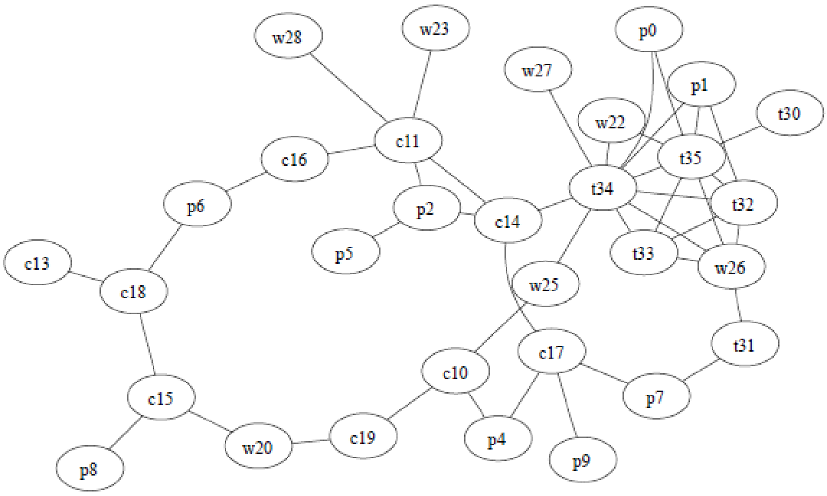


Fig. 31: Example of a node adjacency graph

5.1.3.2 ConnectivityONEReport

Link connectivity report generator for ONE StandardEventsReader input. Creates and returns a String presentation of the connection where the node with the lower network address is first parameter h1, the other node of the connection parameter h2 and the other node of the connection return String presentation of the connection.

1.00 CONN 15 35 up

3.20 CONN 24 36 up

3.20 CONN 15 35 down

5.1.3.3 ContactsDuringAnICTReport

The number of contacts during an inter-contact time metric is similar to the inter-contact times metric, except that instead of measuring the time until a node meets again, we count the number of other nodes both of the nodes meet separately. In contrast to the inter-contact times, the number of contacts during an inter-contact is not symmetric, i.e. during an inter-contact both nodes wait the exact same time but will meet a different number of nodes.

5.1.3.4 ContactTimesReport

ContactTimesReport reports the node contact time (how long they were in the range of each other) distribution. Report file contains the count of connections that lasted for certain amount of time.

5.1.3.5 DeliveredMessagesReport

DeliveredMessagesReport reports information about all delivered messages.

#time ID size hop-count delivery-Time remainingTtl isResponse

1578.6000 M9 587686 2 1313.6000 n/a N

5.1.3.6 DistanceDelayReport

Report for how far apart the nodes were when the message was sent and how long time & how many hops it took to deliver it. If message is not delivered, its delivery time & hop count are reported as -1

#distance at msg send, delivery time, hop count, MSG_ID

1717.6842 1313.6000 2 M9

2030.0926 -1.0000 -1 M8

1079.9724 -1.0000 -1 M13

5.1.3.7 EncountersVSUniqueEncountersReport

EncountersVSUniqueEncountersReport count the total- vs. the unique encounters for each node.

5.1.3.8 InterContactTimesReport

InterContactTimesReport reports the inter-contact time (i.e., the time between the end of previous contact and the beginning of a new contact between two hosts) distribution.

5.1.3.9 MessageDelayReport

Reports delivered messages' delays (one line per delivered message) and cumulative delivery probability sorted by message delays.

messageDelay cumulativeProbability

1313.6000 0.0161

5.1.3.10 MessageDeliveryReport

Report for of amount of messages delivered vs. time. A new report line is created every time when either a message is created or delivered.

#time created delivered delivered/created

32.0000 1 0 0.0000

58.0000 2 0 0.0000

…………………..

1748.0000 61 1 0.0164

1777.0000 62 1 0.0161

5.1.3.11 MessageGraphvizReport

This report creates a graphviz compatible graph of messages that were passed.

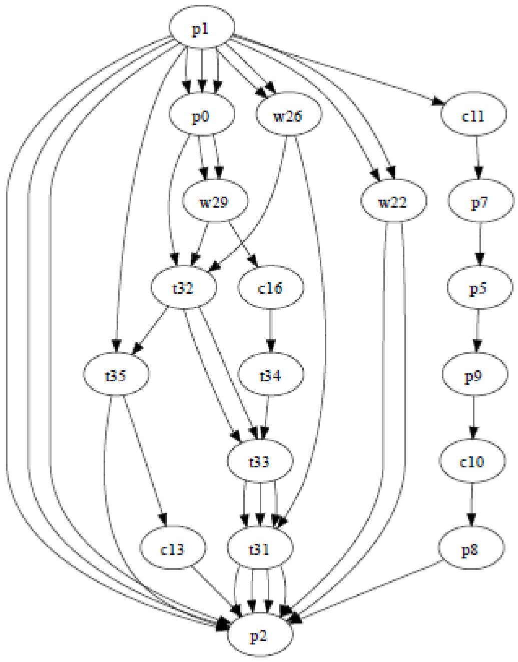


Fig. 32: Example of a message graph

5.1.3.12 MessageStatsReport

MessageStatsReport is a report for generating different kind of total statistics about message relaying performance. Messages that were created during the warm up period are ignored. If some statistics could not be created (overhead ratio if no messages were delivered) "NaN" is reported for double values and zero for integer median(s).

Message stats for scenario

sim_time: 1800.1000

created: 62

started: 10

relayed: 3

aborted: 7

dropped: 0

removed: 0

delivered: 1

delivery_prob: 0.0161

response_prob: 0.0000

overhead_ratio: 2.0000

latency_avg: 129.3000

latency_med: 129.3000

hopcount_avg: 1.0000

hopcount_med: 1

buffertime_avg: NaN

buffertime_med: NaN

rtt_avg: NaN

rtt_med: NaN

5.1.3.13 TotalEncountersReport

A report of the distribution of how many encounters (contacts) a node has had.

5.1.3.14 UniqueEncountersReport

UniqueEncountersReport class creates a report of the distribution of how many unique of the other nodes a node has encountered.

5.2 Simulation scenarios: Epidemic vs. Spray and Wait

We validated our model by comparing it to real user traces in terms of inter-contact times or contact durations. The inter-contact times and contact durations are commonly used in the literature for characterizing connectivity in DTNs.

We had 3 different scenarios (configuration files) regarding the nodes moving on a map (of the Helsinki centre area with the surrounding districts) with the size of roughly 7000 m^2.

The first scenario included 2 groups of nodes (Scenario.nrofHostGroups = 2) every group containing 20 hosts (Group.nrofHosts = 20). The first group was a pedestrian group (Group1.groupID = p, Group.transmitRange = 10, Group.speed = 0.5, 1.5) and the second one was a cars group (Group2.groupID = c, # cars can drive only on roads, Group2.okMaps = 1, # 10-50 km/h, Group2.speed = 2.7, 13.9). The group movement model was Shortest Path Map Based Movement (Group.movementModel = ShortestPathMapBasedMovement) and the scenario time was 1800s (Scenario.endTime = 1800).

We simulated the two routing scenarios Epidemic and Spray and Wait (Group.router = EpidemicRouter, SprayAndWaitRouter).

The second scenario contained 6 groups, each group with 40 hosts (except tram groups: 2 trams each Group6.nrofHosts = 2). The first group was pedestrians, the second cars, third another group of pedestrians (ShortestPathMapBasedMovement) and then the tram groups, having different map route movement and same speed (Group4.movementModel = MapRouteMovement, Group4.routeFile = data/tram3.wkt, Group6.speed = 7, 10). The scenario simulation time was also 1800s.

The third scenario was the most complex having a simulation time duration of 43k (12 hours) and the same number of groups and hosts as the previous.

In all scenarios we used a warm up period of 800s (MovementModel.warmup = 800), the same message creation parameters (# How many event generators Events.nrof = 1 # Class of the first event generator Events1.class = MessageEventGenerator # Creation interval in seconds (one new message every 25 to 35 seconds) Events1.interval = 25,35 # Message sizes (500kB - 1MB) Events1.size = 500k, 1M # range of message source/destination addresses Events1.hosts = 0,125 # Message ID prefix Events1.prefix = M) and very important the MovementModel.rngSeed = 1;the seed for all movement models' random number generator. If the seed and all the movement model related settings are kept the same, all nodes should move the same way in different simulations (same destinations and speed & wait time values are used).

In the Spray and Wait scenario we also have to define: SprayAndWaitRouter.nrofCopies = 6 SprayAndWaitRouter.binaryMode = true.

5.2.1 Simulation results

Using the reports generated we can make a comparison between the different scenarios we implemented. For example using the results of MessageStatsReport (Epidemic/Spray And Wait) we can make a comparison between the number of messages created, relayed, aborted, delivered or their delivery probability, latency, overhead ratio.

Scenario 1 (2 groups, 1800s)

Message stats for scenario Epidemic	Message stats for scenario SpayAndWait
sim_time: 1800.1000	sim_time: 1800.1000
Created: 62	Created: 62
Started: 10	Started: 126
Relayed: 3	Relayed: 41
Aborted: 7	Aborted: 85
Dropped: 0	Dropped: 1
Removed: 0	Removed: 0
Delivered: 1	Delivered: 1
delivery_prob: 0.0161	delivery_prob: 0.0161
response_prob: 0.0000	response_prob: 0.0000
overhead_ratio: 2.0000	overhead_ratio: 40.0000
latency_avg: 129.3000	latency_avg: 1313.6000
latency_med: 129.3000	latency_med: 1313.6000
hopcount_avg: 1.0000	hopcount_avg: 2.0000
hopcount_med: 1	hopcount_med: 2
buffertime_avg: NaN	buffertime_avg: 1704.1000
buffertime_med: NaN	buffertime_med: 1704.1000
rtt_avg: NaN	rtt_avg: NaN
rtt_med: NaN	rtt_med: NaN

The reports are showing that creating the same number of messages the two scenarios have the same delivered messages, but SNW has a bigger overhead ratio and latency.

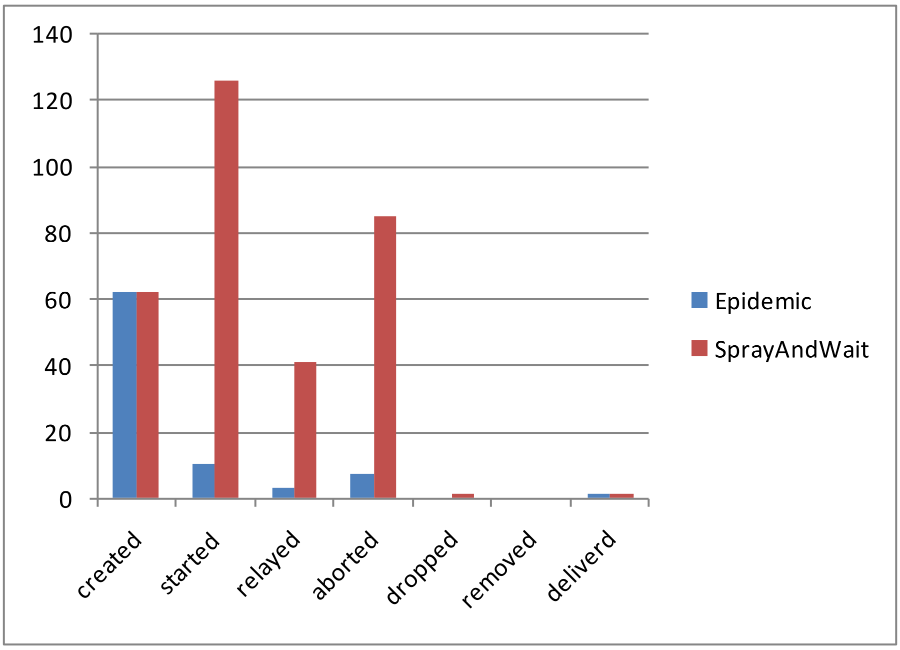


Fig. 33: Message Statistics (2 groups)

Scenario 2 (6 groups, 1800s)

Message stats for scenario Epidemic

sim_time: 1800.1000

Created: 59

Started: 1334

Relayed: 637

Aborted: 692

Dropped: 145

Removed: 0

Delivered: 7

delivery_prob: 0.1186

response_prob: 0.0000

overhead_ratio: 90.0000

latency_avg: 1001.6571

latency_med: 1192.7000

hopcount_avg: 2.8571

hopcount_med: 2

buffertime_avg: 480.5097

buffertime_med: 435.6000

rtt_avg: NaN

Message stats for scenario SprayAndWait

sim_time: 1800.1000

Created: 59

Started: 453

Relayed: 189

Aborted: 262

Dropped: 0

Removed: 0

Delivered: 5

delivery_prob: 0.0847

response_prob: 0.0000

overhead_ratio: 36.8000

latency_avg: 875.6200

latency_med: 722.5000

hopcount_avg: 2.4000

hopcount_med: 2

buffertime_avg: NaN

buffertime_med: NaN

rtt_avg: NaN

The second scenario's reports present almost the same delivered messages (7 for Epidemic, 5 SNW delivery probabilities 0, 118 vs. 0,084) but more latency and overhead for the Epidemic case.

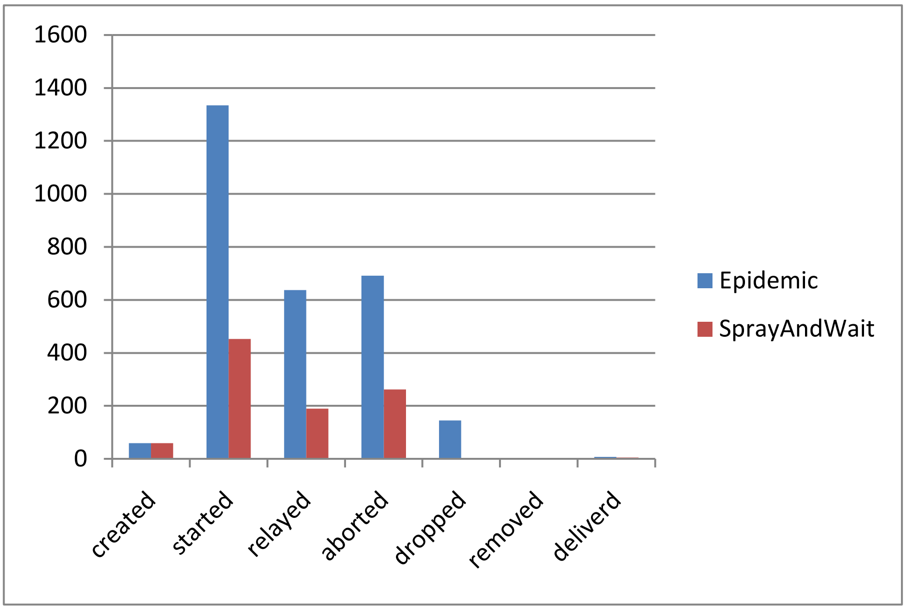


Fig. 34: Message Statistics (6 groups)

Scenario 3 (6 groups, 12 hours)

Message stats for scenario Epidemic

sim_time: 43000.1000

Created: 1454

Started: 60062

Relayed: 32238

Aborted: 27823

Dropped: 32060

Removed: 0

Delivered: 339

delivery_prob: 0.2331

response_prob: 0.0000

overhead_ratio: 94.0973

latency_avg: 4762.4018

latency_med: 3301.3000

hopcount_avg: 4.4897

hopcount_med: 4

buffertime_avg: 1462.2695

buffertime_med: 914.7000

Message stats for scenario SprayAndWait

sim_time: 43000.1000

Created: 1454

Started: 20910

Relayed: 7619

Aborted: 13290

Dropped: 7365

Removed: 0

Delivered: 696

delivery_prob: 0.4787

response_prob: 0.0000

overhead_ratio: 9.9468

latency_avg: 3492.6078

latency_med: 2432.0000

hopcount_avg: 2.5000

hopcount_med: 3

buffertime_avg: 3643.8662

buffertime_med: 3238.8000

This scenario presents the most complex situation and the longest. The results indicate that Spray and Wait scenario has a greater delivery probability (0,478 vs. 0233) with 696 delivered messages from 1454 created, when Epidemic can only deliver 339 messages. Epidemic scenario also has a bigger latency and overhead ratio, but the buffer time average is lower, as we could have predicted.

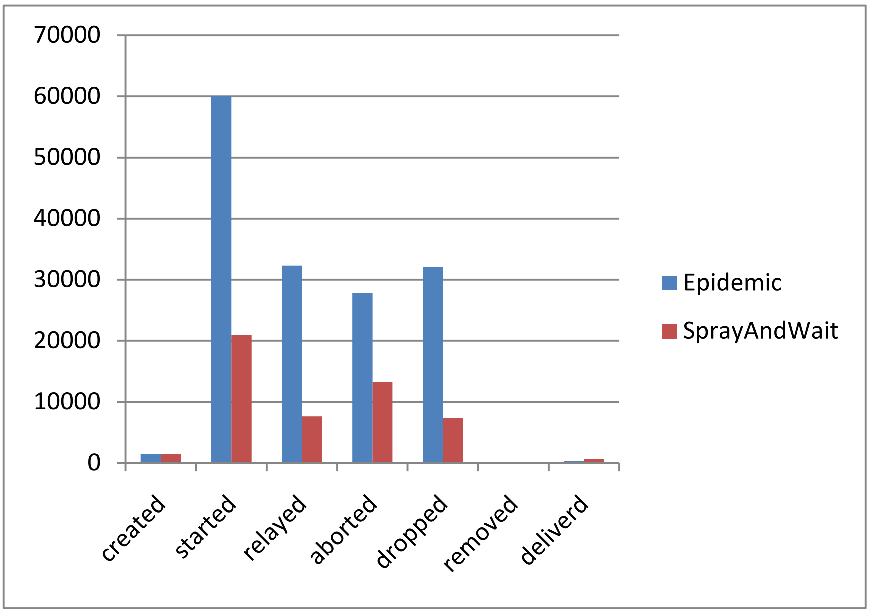


Fig. 35: Message Statistics (12 hours)

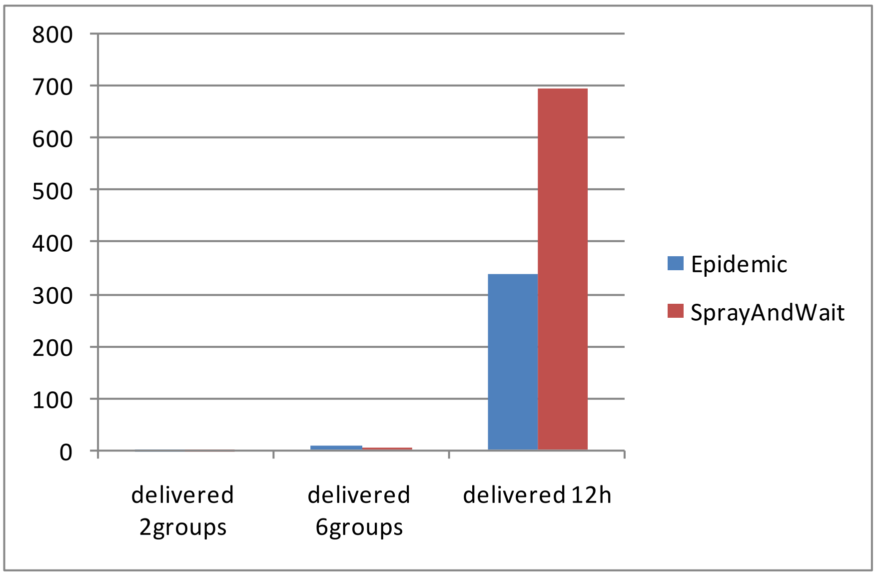
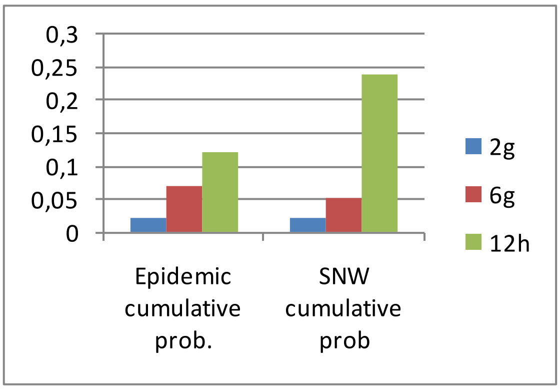


Fig. 36: Comparison between scenarios delivered messages



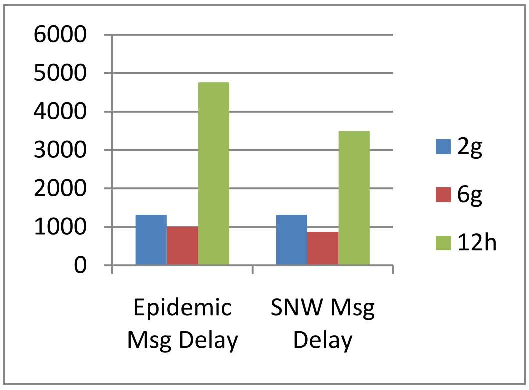


Fig. 37: Comparison between scenarios cumulative probability/message delay

The other reports (ConnectivityONEReport, ContactTimesReport, DeliveredMessagesReport etc) are also very valuable and used to create plots or graphical views (using programs like Gnuplot or Graphviz).

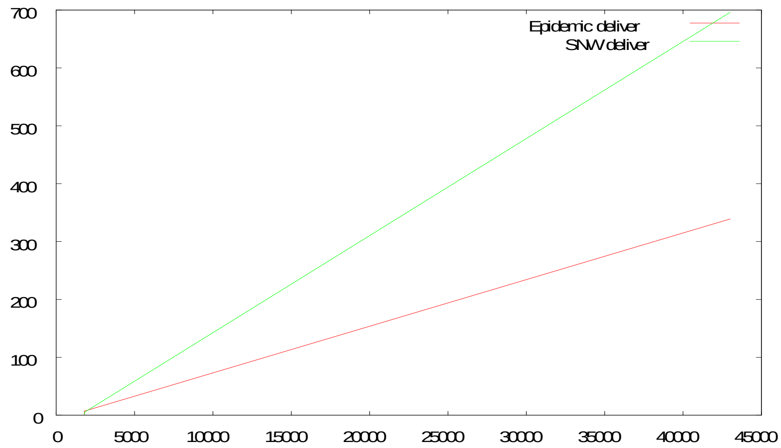


Fig. 38: Deliver plot

Fig. 38 shows the raport between the Epidemic and Spray and Wait number of delivered messages in a 43k second (12 hours) simulation.

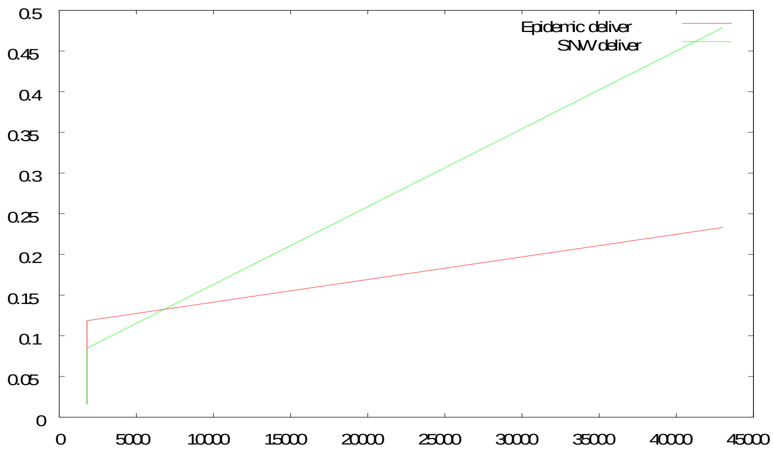


Fig. 39: Delivery probability plot

Fig. 39 shows that even thought at the beginning SNW scenario has a smaller delivery probability, for longer periods of time it overcomes the Epidemic result and has a better delivery rate.

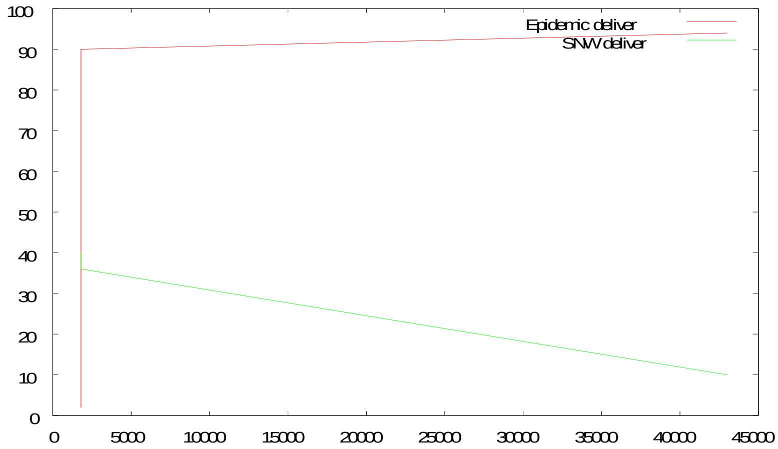


Fig. 40: Overhead ratio plot

The plot confirms that the overhead ration for the SNW scenario is much smaller and is decreasing over the time (even thought at the beginning, for small periods of time the ratio is higher than for the Epidemic case).

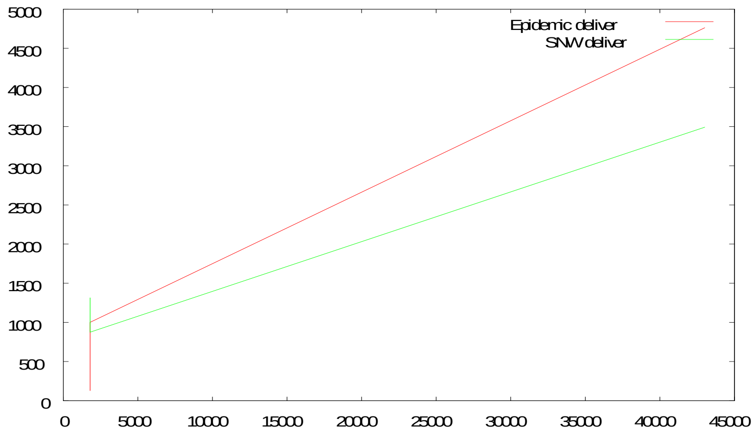


Fig. 41: Latency plot

The same result occurs for the latency and hop counts. For a 1800s simulation scenario SNW introduces a higher latency, but than for a 12 hours simulation the Epidemic scenario overcomes SNW in latency.

The Epidemic case has more hop counts than SNW for a longer duration. In the 1800 s simulation the result is the same as in latency and overhead situation.

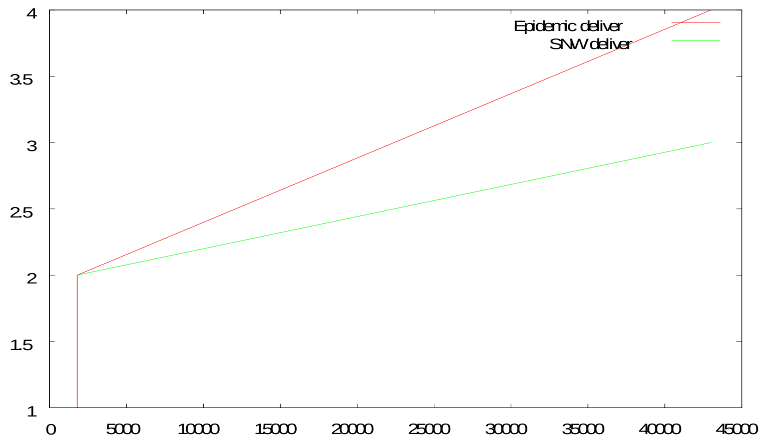


Fig. 42: Hopcount plot

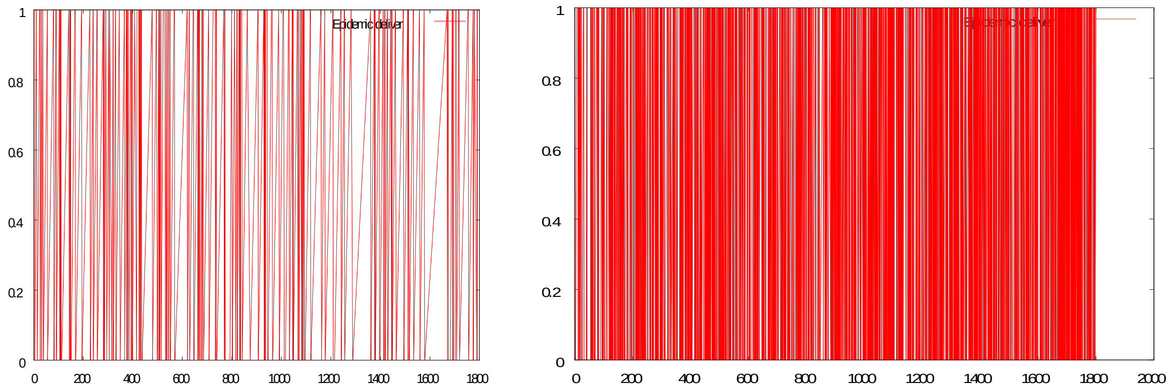


Fig. 43: Connectivity Epidemic 2 groups / 6 groups

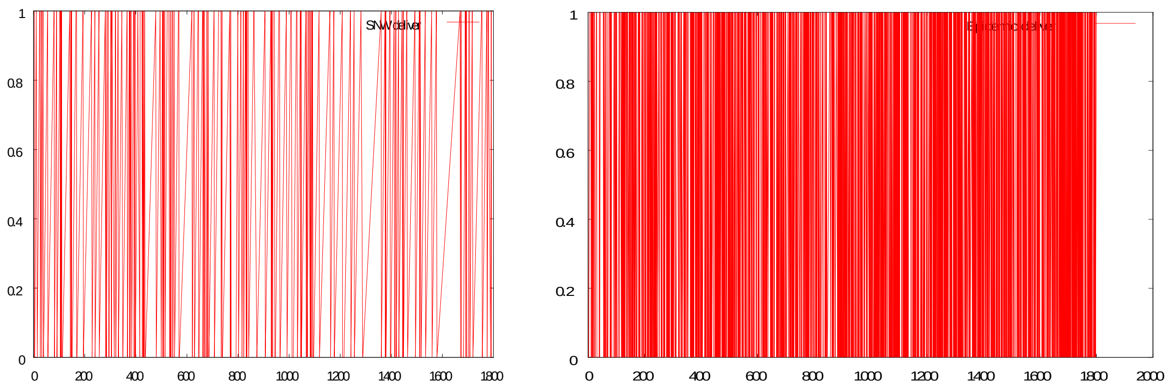


Fig. 44: Connectivity SNW 2 groups / 6 groups

The connectivity report present how often the nodes communicate between them and form a connection link with a rate. The differences between the two scenarios are not very obvious. Instead the differences between the deliveries time of the two scenarios can be compared from the plots above. In the 12 hours simulation SNW delivers in the same period of time more messages (700 messages SNW/ 350 Epidemic) and in the 1800 s case it is exactly the opposite, but the difference is smaller (5 messages SNW/ 7 Epidemic).

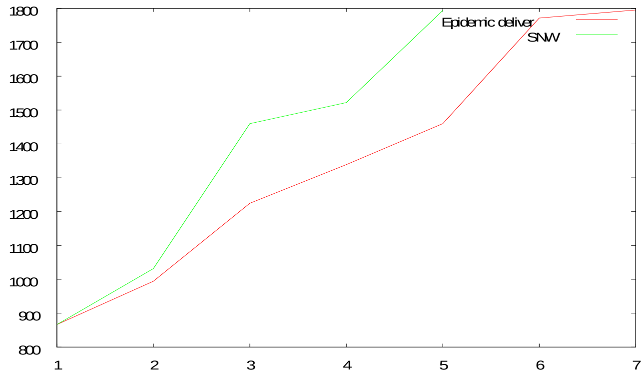
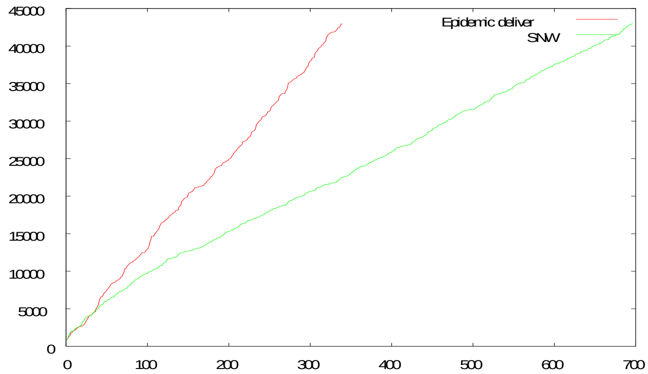


Fig. 45 and 46: Time of delivery (1800 s/12h)



TotalEncountersReport and UniqueEncountersReport provide the necessary information to plot the total and unique number of encounters between the nodes (40 nodes for 2 groups and 120 for the 6 groups scenarios).

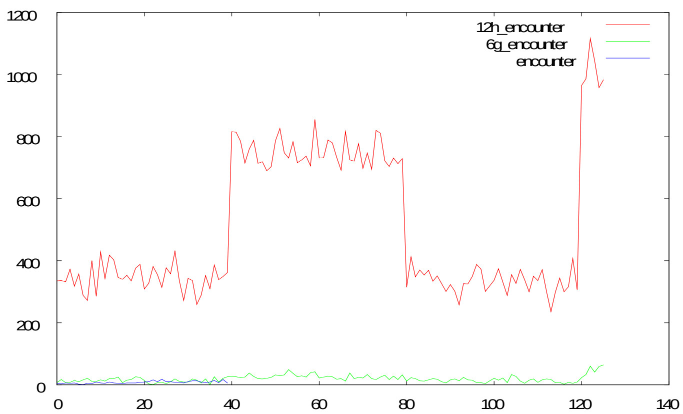


Fig. 47: Encounters plot

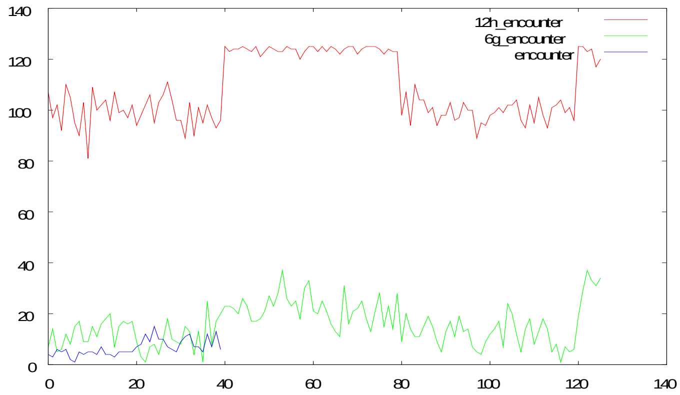


Fig. 48: Unique encounter plot

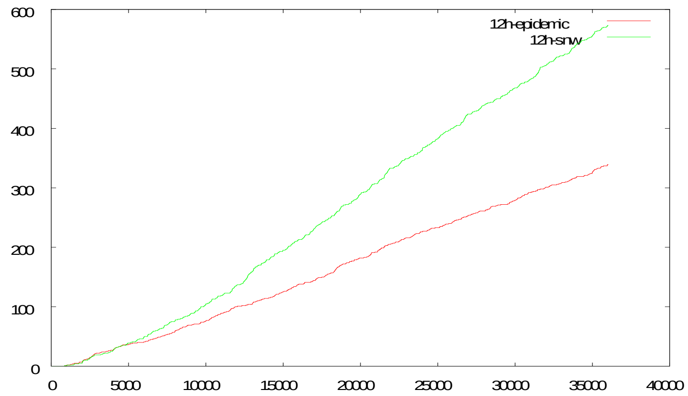
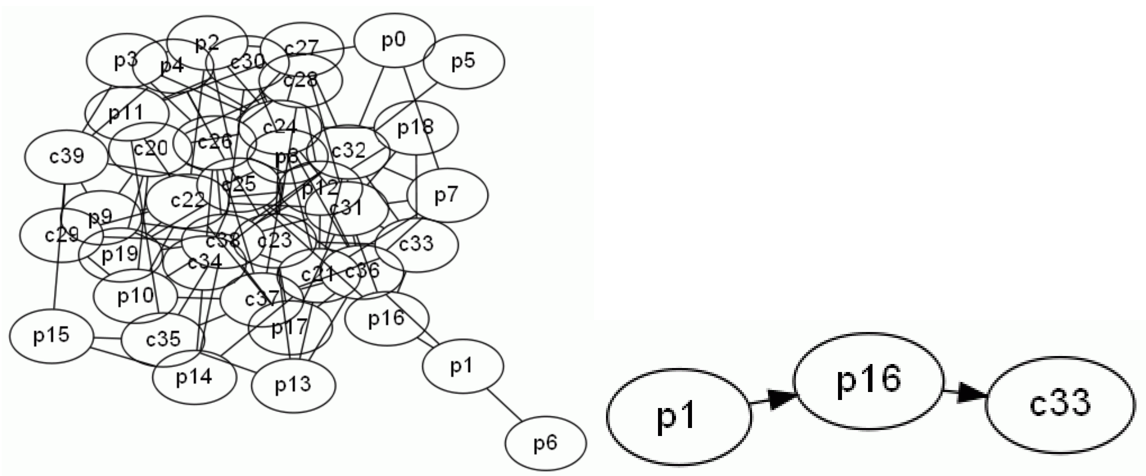


Fig. 49: Time deliver (12h) plot

We also used the AdjacencyGraphvizReport and MessageGraphvizReport reports with Graphviz and we visualized for both scenarios Epidemic and Spray and Wait the connections and messages that were passed graphs.



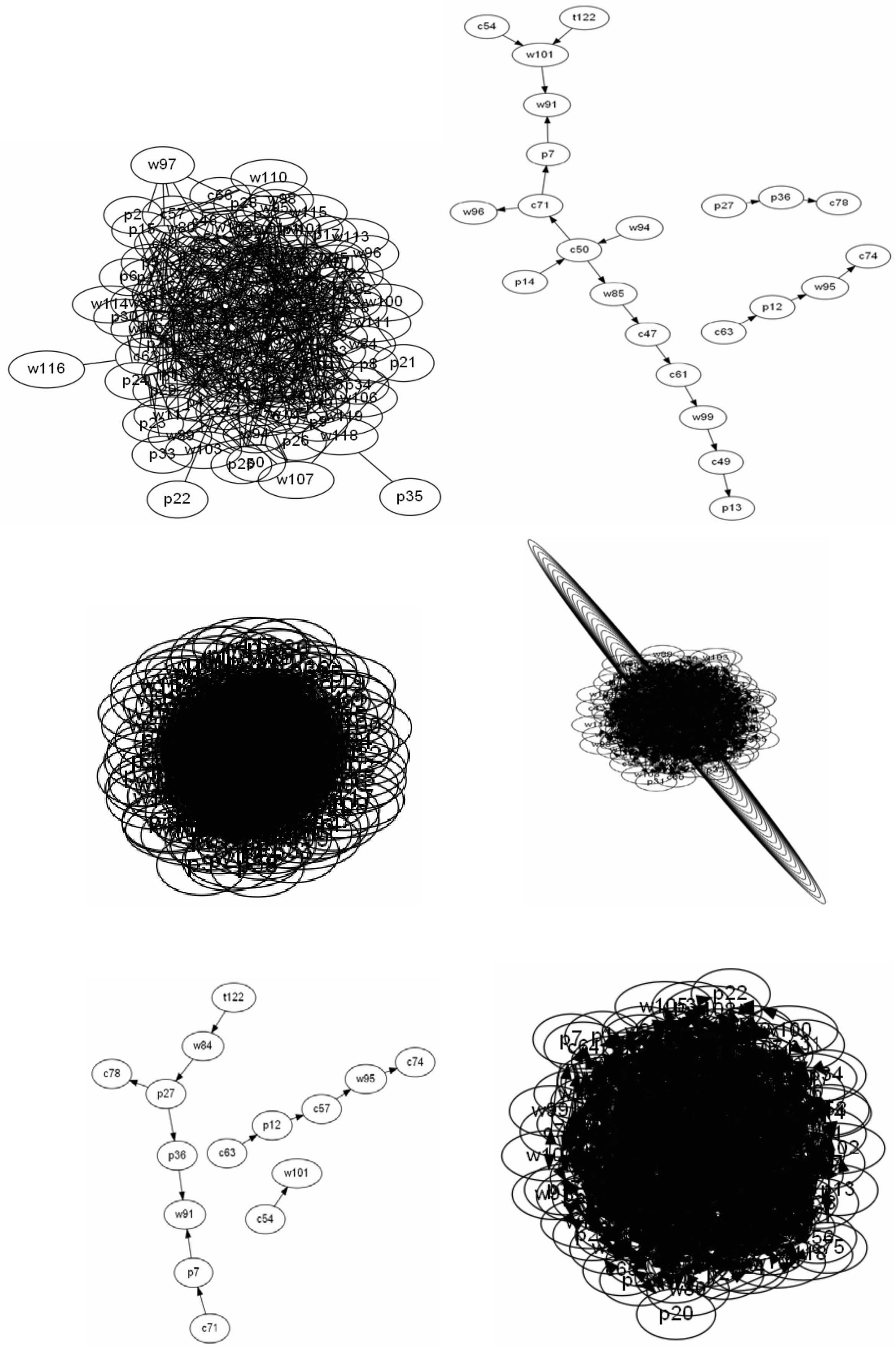


Fig. 50: Adjacency and Message graphs for Epidemic and Spray and Wait scenario

5.3 Mars DTN Spray and Wait Simulation

We can use for this simulation 5 groups of nodes: Rovers (10), Landers (4), Sensors (50), Humans (6), and Satellites (4).

A rover speed is 5cm/s (Group.speed = 0.05), we assume landers and sensors are stationary humans move with 0.5 m/s (Group.speed = 0.5) and satellites with 3,359 km/h ($v = \sqrt{\frac{M*G}{R+h}}$ where M is the mass of Mars ($0{,}64*10^{24}$), G is a constant = $6{,}673 *10^{-11}$ Nm^2/kg^2, R is the Mars radius (3386 km) and h (400km) is the distance between the planet surface and the satellite).

Group.movementModel = MapRouteMovement

Group.router = EpidemicRouter / SprayAndWaitRouter

Group.bufferSize = 5M

Group.transmitRange = 10

transmit speed of 2 Mbps = 250kBps

Group.transmitSpeed = 250k

Group.waitTime = 0, 120

walking speeds

Group.speed = 0.5, 1.5

#Group.msgTtl = 60

Group.nrofHosts = 10

group1 (humans) specific settings

Group1.groupID = p

Group1.nrofHosts = 6

Group.transmitRange = 50

group2 (landers) specific settings

Group2.groupID = c

Group2.speed = 0

Group.nrofHosts = 4

Group.transmitRange = 1500

#group3 (rovers)

Group3.groupID = r

Group2.speed = 0

Group3.speed = 0.05

Group.transmitRange = 250

Group.nrofHosts = 10

#the sensors groups

Group4.groupID = s

Group4.bufferSize = 5M

Group.nrofHosts = 50

#the Satellites groups

Group5.groupID = s

Group5.bufferSize = 50M

Group5.movementModel = MapRouteMovement

Group5.routeFile = data/Satellites.wkt

Group5.routeType = 1

Group5.waitTime = 10, 30

Group5.speed = 3359

Group5.nrofHosts = 4

Chapter 6

6. Summary and Conclusions

The thesis started with a short introduction about MANET. We discussed the principles, the applications, the most important qualities and major problems and then we looked at the most representative MANET's routing protocols. We classified them in different categories with different specifications and we discovered the strategies utilized for these routing protocols and again we focused on the challenges and advantages.

After this introduction about MANET we proposed to take a look at the red planet and begun with its surface and atmosphere. Then we listed the past missions over the years and we also imagined the Mars future. The thesis describes then some Mars architectures, the most important network types and the challenges imposed by this type of environment. Solutions for these challenges were also discovered and the thesis presents them.

One proposed solution is DTN (Delay Tolerant Networking) and chapter 4 presents the characteristics and types of this approach. Other discussed solutions are integrating MANET with DTN and the optimized QOS protocols.

Chapter 5 is dedicated to DTN simulations using the Opportunistic Network Environment simulator. Different scenarios, like Epidemic and Spray and Wait routing are presented and then comparisons between them are being made. The particular case of a Mars DTN is considered after presenting the plots and graphics.

Based on the simulations it is found that the Spray and Wait scenario is working better in simulations with higher numbers of nodes and for longer periods of time. Because of the limitation of time and resources, in this thesis work, only limited scenarios are simulated, and each result is taken from an average reports results. Thus, if more accurate performance results are expected, more models and scenarios should be taken.

Even though the ONE is a quite capable simulator, it naturally also has its limitations. To make simulations feasible, many real world aspects have been abstracted or even completely discarded. Some of these features are due to the innate nature of the simulator but others are possibly some day solved by future versions of the simulator.

Present technology and solutions to challenges made recent exploration of Mars possible and has revealed evidence of several defining moments in its history. Still this domain is at the beginning and more is to be done and study.

References

[1] A Mobile Ad Hoc Network with Mobile Satellite Earth-stations, Song Han, Guochang Gu1, and Jun Ni School of Computer Science and Technology, Harbin Engineering University, 2007 IEEE

[2] A Novel Solution for Global Connectivity in MANET, Vinh Dien HOANG, Zhenhai SHAO, Masayuki FUJISE, Wireless Communications Laboratory, National Institute of Information and Communications Technology, Singapore Office, Singapore, 2004 IEE

[3] Adaptive Routing for Intermittently Connected Mobile Ad Hoc Networks, Mirco Musolesi, Stephen Hailes, Cecilia Mascolo, Department of Computer Science, University College London , Proceedings of the Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM'05)

[4] Autonomy for Mars Rovers, Max Bajracharya, Mark W. Maimone, and Daniel Helmick, Jet Propulsion Laboratory, California Institute of Technology, 2008 IEEE

[5] Buffer Management Policy for Mars Intelligent Proximity Network, Mahendiran Prathaban and Josephine Kohlenberg, Dept - Reseaux et Services des Telecommunications, Institut National des Telecommunications, 2008 IEEE

[6] Challenges in Interplanetary MANETs, Chien-Chung Shen Sundaram Rajagopalan Computer and Information Sciences, University of Delaware Newark, DE 19716

[7] Delay Tolerant Networking, Scott Burleigh, Adrian Hooke, and Leigh Torgerson, Jet Propulsion Laboratory Kevin Fall, Intel Corporation Vint Cerf, MCI Bob Durst and Keith Scott, The MITRE Corporation Howard Weiss, SPARTA, Inc., IEEE Communications Magazine • June 2003

[8] Experience with delay-tolerant networking from orbit, W. Ivancic, W. M. Eddy, D. Stewart, L. Wood, J. Northam and C. Jackson, a draft International Journal of Satellite Communications and Networking, Advanced Satellite Mobile Systems Conference (ASMS 2008).

[9] Integrating DTN and MANET Routing, J¨org Ott, Helsinki University of Technology Networking Laboratory SIGCOMM'06 Workshops September 1115, 2006, Pisa, Italy.

[10] Load Balanced, Energy-Aware Communications for Mars Sensor Networks, Xiaoyan Hong, Mario Gerla and Hanbiao Wang, Computer Science Department, University of California at Los Angeles, supported by NASA JPL under IPN-ISD Technology Task
Project: “Ad-hoc Network Protocols”, 0-7803-7231-X/01/$10.00/ c2002 IEEE IEEEAC paper #194, Updated Dec 2, 2001

[11] Optimized QOS Protocols for Small-Sized Manets, Mohammad Inayatullah, Sheeraz Ahmad, Abdus Salam, NWFP University of Engineering & Technology, Pakistan, IEEE—ICET 2006 2nd International Conference on Emerging Technologies Peshawar, Pakistan 13-14 November 2006

[12] ORBIT Mobility Framework and Orbit Based, Joy Ghosh, Sumesh J. Philip, Chunming Qiao, Department of Computer Science and Engineering University at Buffalo CSE TECHNICAL REPORT

[13] Proximity Networks Technology Assessment, Timothy J. Salo, Barry A. Trent, and Timothy Hartley, Architecture Technology Corporation, Eden Prairie, Minnesota, NASA/CR—2003-212623, October 2003

[14] Routing for Data Delivery in Dynamic Networks, Padma Mundur, Sookyoung Lee, and Matthew Seligman

[15] Stigmergic Learning for Self-Organizing Mobile Ad-Hoc Networks (MANET’s), H. Van Dyke Parunak, Sven A. Brueckner, Altarum Institute, AAMAS'04, July 19-23, 2004, New York.

[16] Routing in intermittently connected mobile ad hoc networks and delay tolerant networks: overview and challenges, ZHENSHENG ZHANG, SAN DIEGO RESEARCH CENTER, IEEE Communications Surveys & Tutorials 1st Quarter 2006

[17] Dynamic virtual backbone routing protocol for ad hoc networks, Melvin John, University of Adelaide, November 2006

[18] Quality of Service Aware Routing Protocols for Mobile Ad Hoc Networks, Shan Gong, Espoo, 2006

[19] Anomaly Detection for Wireless Ad-Hoc Routing Protocols, Yi-an Huang, North Carolina State University, Raleigh 2001

[20] Performance Comparison of MANET Routing Protocols in Ad-hoc and Hybrid Networks, Thomas Staub, University of Berne, Switzerland, Februar 2004

[21] Topology control, routing protocols and performance evaluation for mobile wireless ad hoc networks, HUI LIU, College of Arts and Sciences Georgia State University, 2005

[22] Advanced Routing Protocols for Satellite and Space Networks, Chao Chen, School of Electrical and Computer Engineering, Georgia Institute of Technology, May 2005

[23] Metric of Trust for Mobile Ad hoc Networks, Denise Umuhoza, University of the Western Cape, May 2006

[24] http://www.dtnrg.org/wiki/Docs

[25] http://www.sciencedirect.com

[26] http://www.netlab.tkk.fi/tutkimus/dtn/theone/#contrib

[27] http://www.astrosurf.com/luxorion/qsl-mars-communication.htm

[28] Advanced Communication and Networking Technologies for Mars Exploration, Kul Bhasin and Jeff Hayden, Jonathan R. Agre, Loren P. Clare, and Tsun-Yee Yan, prepared for the 19th International Communications Satellite Systems Conference and Exhibit sponsored by the American Institute of Aeronautics and Astronautics, Toulouse, France, April 17-20, 2001

[29] Interrogation-Based Relay Routing in Ad hoc Satellite Networks

[30] http://nesl.ee.ucla.edu/project/show/19

[31] Opportunistic Network Environment simulator, Ari Keranen, Helsinki University of Technology, Department of Communications and Networking, Special assignment, May 29, 2008

[32] Working Day Movement Model, Frans Ekman, Ari Keränen, Jouni Karvo and Jörg Ott Helsinki University of Technology TKK, Dept. of Communications and Networking

[33] http://www.gnuplot.info/

[34] http://www.esa.int/

[35] Delay-Tolerant Networks Tutorial, Vinton Cerf, Scott Burleigh, Adrian Hooke,DTN research group Internet Draft, March 2003