Design Considerations for Sensor Networks with Gateways

Lotfi Benmohamed, Phil Chimento, Bharat Doshi, and I-Jeng Wang

Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723

ABSTRACT

In a large class of sensor network deployments, a small subset of the sensors covering the sensor field is equipped with special communications capability to communicate with operators outside the sensor field. These sensors play the role of gateways for off-field communication in the sense that all communications to- or out of the field is through these nodes, and the other non-gateway nodes are only capable of sensor-to-sensor communication. This design achieves a lower cost by concentrating expensive communication devices in a small subset of nodes. An important problem in designing such gateway-based sensor networks is determining the number of gateway nodes needed, their location in the sensor field, and the automation of the sensor-to-gateway association for off-field communication. These design considerations are addressed in this paper.

In determining the number of gateways the tradeoff is between performance and cost. As the number of gateways increases, less traffic load is placed on each gateway and its surrounding nodes, resulting in longer network lifetime and larger off-field aggregate transmission capacity. However, with a larger number of gateways the network may be too costly to deploy as gateway nodes are more expensive than non-gateway sensor nodes. We develop and analyze models that allow us to determine the optimal number of gateways and their location in the sensor field. We also provide initial results with respect to determining the needed number of fusion nodes. While the presence of multiple gateways offers a higher degree of off-field communication reliability, a sensor will need to select one of the gateways at a time for off-field communication. In this paper, we also propose a dynamic sensor-to-gateway association protocol. Based on current energy levels, the distributed protocol dynamically assigns sensors to gateways in such a way that the overall transmission load is balanced among the different gateway regions over the lifetime of the sensor field.

Keywords: Distributed sensor field, sensor networks, gateway nodes

1. INTRODUCTION

The confluence of many technological advances, emerging needs, and innovation in architecture is driving towards viability and desirability of large and distributed sensor fields in a wide variety of applications. The applications may involve: short term and medium term weather prediction; understanding phenomena like tornados and hurricanes; detection and tracking of hidden and/or moving objects on ground, at sea, under sea, or in air; object tracking and targeting; detection of biological and chemical agents; etc. All these applications involve a large number of sensors, each sensor collecting "information" that it can obtain, and use of the collective information to achieve the goals of detection, tracking, targeting, etc. However, the applications differ in major ways. In particular, underlying physics of sensing, remote controllability of sensors, autonomous and uncontrollable movement of sensors, degree of collaboration needed and possible, possible ways to communicate results within and outside the field, energy constraints, and expected duration of use vary significantly among applications. Also, the applications differ significantly in the desirable trade-offs between false alarms and missed detection, in the penalty of the field itself getting detected by others, in vulnerability to being fed wrong information, and in the value of the speed of detection. Thus, the design of sensor fields have to carefully consider the above factors associated with underlying application. Our research is aimed at such a comprehensive study and design methods for one specific category of sensor fields, one used to detect and track objects moving under water (for example, submarines). Advent of low cost, low power, and small sized sensing and communication devices make

For further information send correspondence to L. Benmohamed, E-mail: Lotfi.Benmohamed@jhuapl.edu

Digital Wireless Communications VII and Space Communication Technologies, edited by
Raghuveer M. Rao, Sohail A. Dianat, Michael D. Zoltowski, Rabindra Singh, Susan P. Miller,
Proc. of SPIE Vol. 5819 (SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15 · doi: 10.1117/12.603745

it economical to deploy large sensor fields for this application. On the other hand, the advances in quieting and hiding technologies and increasingly "noisy" surroundings in littoral scenarios make it necessary to deploy a large number of closely spaced sensors to maximize detection probability while keeping the false alarm rate to a manageable level.

Much work has been done in developing sensing technologies for this application. Work has also been done in understanding the impact of some of the factors mentioned above. However, the collective influence of these factors on the design of topology and topology controls, design of communication architecture and protocols within the field and between the field and external world, design of fusion mechanisms within and outside the field have not been studied before. Our research is directed at this multi-dimensional problem. In a companion paper, we describe the overall problem and the myriad of factors involved in the overall design. In this paper, we focus on one problem in the overall design. This problem involves determining the number and locations of gateway nodes (nodes that provide gateway capability between the sensor field and the "users" outside the field). The problem also deals with determining how the gateways are used by other nodes (homing and routing from a sensor node to external world).

The sensor field problem we study involves 100 to 10,000 sensors in a square, rectangular, or circular/elliptical formation. Sensors within the field behave like mobile ad hoc networks. The communication level is small in absence of potential detection. However, it may increase significantly after a suspected detection. Thus, false alarms may result in significant communication and drain on the battery life. In many applications, batteries cannot be replenished and thus the false alarms create major reduction in the field lifetime. In order to minimize the probability of the field getting detected and to minimize the drain on energy store in sensors, the sensors operate and communicate at low power and hence with short range. Thus, most pairs of sensors will need many hops to communicate with each other. Of course, some information needs to be communicated outside of the field to shipboard, airborne, or space based platforms, where "users" can collect and analyze the information and decide on a course of action. The "users" may also send queries to sensors for more details. This external communication requires higher power and more expensive equipment, and hence the design of the field should use a limited number of nodes with external communication capability. We call these nodes 'gateways'. Besides gateways, the field may have other specialized nodes: fusion nodes and cluster heads. Their roles and relationships with gateways have been discussed in.[1]

As mentioned earlier, this paper focuses on the problems related to the design of the gateway subsystem and its relationship with other nodes and external platforms.

In Section 2, we describe the model in some detail, with an emphasis on gateway subsystem design problem. Section 3 discusses network traffic pattern and how it is affected by the number and location of gateways. In Section 4, we use the insight from Section 3 to discuss approaches to node planning. Given the number and locations of gateways, the next questions are homing and routing from other nodes. These are discussed in Section 5.

While this paper and our work are focused on a specific set of applications of distributed sensor field with fusion, a subset of the analysis and results will be applicable broadly.

2. MODEL

The sensor field we consider typically involve $N = 100$ to 10,000 sensor nodes. The configuration depends on the area to be covered for detection. Common situations are rectangular or elliptical planar arrangements of sensors covering a few tens to over a hundred square miles. In some applications, the sensors may be arranged in concentric ellipses surrounding a platform. For simple analysis, we will use the model involving square planar grid. Simulation models will consider more complicated geographies.

Detection may be based on passive signal monitoring, that is, sensors listening to the signals created by active equipment in the target. Increasingly effective quieting techniques and high noise level in littorals make it hard to use this passive monitoring effectively. Detection may also be based on proactive methods. One such method uses reflection phenomenon. That is, a sensor sends a burst of radiation (ping) and all sensors listen. An object of interest will reflect and some of the receivers will receive reflected energy and may consider this as possible detection. Others may not receive reflected signal even when an object is present. Finally, some may receive

high energy direct ping and may not be able to identify reflection. The physics suggests a geographical area around the target where the receivers may receive reflected signal that can be suspected to indicate the presence of a target.[2] There is an ellipse of uncertainly and the parameters of the ellipse are governed by the object to be detected and by the distances between ping source and target. Of course, the signal received may be affected by many other things going on under water as well as by things happening on the water surface. The noise may be stronger near the shore than at deep sea. Fusion within the field may be used to improve "signal to noise ratio" and reduce false alarm probability while increasing probability of detection. Fusion within the field also helps reduce the suspected detection traffic going far and reducing the life of the field. Sensors consume energy in pinging and listening. Thus, the number of pings during a sensor's lifetime is limited. The frequency and order of pings may be preprogrammed, decided by external platform and communicated to sensors after sensors are deployed, or it may be decided in a distributed fashion by rules based on recent pings in the neighborhood. The selection of the mechanism and algorithm for frequency and order of pinging is an interesting and important challenge.

If detection hypothesis is further supported by fusion within the field, that information needs to be sent to external "user" which can fuse information from all fusion nodes indicating potential detection and take action. The information is carried from the fusion node to the gateway node and then communicated out to the 'user'. Thus, the traffic flows from sensor nodes to fusion nodes close to the sensor nodes and then from fusion nodes to the gateway node, and finally to external "users". The opposite happens in reverse direction.

A simple analysis of the situation suggests that nodes near a gateway carry more traffic than the nodes away from gateway and may thus become limiting factor in the life of the field. Also, the traffic at a gateway node reduces when the number of hops in "fusion node to gateway node" communication reduces. This happens when the number of gateway nodes increases. Thus, larger number of gateway nodes implies a reduction in the load per sensor node and hence longer life of the field. Having larger number of gateways also allows higher overall capacity for communication between the field and external "user" and provides redundancy. Of course, gateway nodes are more expensive and hence we need to analyze the trade-off between longer life and higher capacity on one hand and lower cost on the other. That analysis is one of the goals of this paper. Analysis to determine locations of the gateways and mapping from sensor nodes to gateway nodes indicating traffic flow is another key goal.

We assume that the sensor-sensor communication uses power P, giving a nominal range R and nominal data link rate C. Of course, these parameters change depending on the environmental noise, fading, and other impairment. We assume that the normal sensor nodes have a total supply, E, of energy available. There are $N_g < N$ gateway nodes. These nodes have normal sensor functionality plus additional energy supply of E_g in the equipment designed for external communication (to ships, planes, satellites, etc.). Both the number N_g and the location of these N_g nodes are design variables for our problem.

As mentioned earlier, our sensor field may involve two more types of special nodes (nodes that do more than basic sensing and communication with other sensor nodes) in addition to the gateway nodes:

The second set is fusion nodes. These nodes get sensing results from nearby sensor nodes and apply fusion algorithms to derive composite picture. The results, if of interest, are typically sent to the nearby gateway node for communication to external node where further fusion can be accomplished and action can be taken. Number of fusion nodes, N_f, and their locations are design variables. If N_f is zero, we have no fusion inside the field and fusion, if any, is done in external nodes. In this paper, we provide some insight into the tradeoffs involved in choosing N_f, mainly from the communication cost and lifetime perspectives.

The third type of special nodes is cluster heads. For large fields (large N) and small range (R), some sensor-sensor communication will involve a large number of hops using many sensor nodes acting as relays. If we have a small number of gateways, then communication from a sensor to a nearby gateway may use many hops and hence use energy from many sensor nodes. The most impacted are the ones near the gateways. Clustering is one way to mitigate the imbalance in loading created by large number of hops in communication. Effectively, a hierarchy is created among sensor nodes. Some nodes are able to communicate longer distances among themselves with higher power (and possibly at a different frequency). These nodes will also communicate with normal nodes at normal power. These nodes are called cluster heads. Sensor-sensor communication and sensor-gateway communication

may involve sensor-sensor-cluster head-cluster head-gateway type routing. Cluster heads may carry extra energy supply of E_c. There is also a possibility of designing the field so every node can become cluster head. In that case, cluster heads configuration can be changed dynamically at regular intervals or based on the field status. In this paper, we will not consider such a dynamic cluster head arrangement. We do not use clustering in this paper. However, we do analyze the impact of dynamic topology control, that is a mechanism by which communication can bypass some nodes and hence reduce the number of hops.

3. NETWORK TRAFFIC

Sensor network traffic both affects and is affected by the choices of protocols for communication within the field, as well as by the design of application processing. In this section, we describe briefly the different types of traffic that the communication subsystem in the sensor field must handle. We can classify traffic sources as follows: Sensor (application) sources; query sources and responses; and network control sources. In part, the traffic generation patterns of the field will depend upon the types of surveillance targets; surface targets will present different signal characteristics to the sensors than underwater targets, as well as have different movement patterns and speeds. In general, though, we can expect low signal-to-noise ratios at the sensors and consequently a high false positive rate. We assume that initial sensor contact messages will be fairly frequent, but because we also assume significant in-field processing, we expect this traffic to be distributed only locally among nearest neighbors (i.e., 1 or 2 hops away) and not sent to the gateways. This local traffic may be directed toward fusion nodes that have additional processing capabilities, or may trigger local distributed processing in order to provide initial false positive filtering.

We envision that the application-driven traffic patterns will evolve in the following way: As a result of an active probe (for example a "ping") or as a result of passive monitoring, the sensors on an individual node determine a signal of interest. To determine whether the signal is truly significant, the sensor node shares its information with neighboring nodes which, in general, should also have sensed a signal. (Note that environmental phenomena may cause signals to be "heard" far beyond their normal propagation range.) The sensor nodes collaborate locally and determine whether the information from the neighborhood of sensors requires further processing and fusion. At this point, one or more of the sensors may forward this information to fusion nodes. Note that an alternate design is to skip the local processing and always simply forward information to fusion nodes for processing.

The fusion nodes, having more processing power and information from multiple sensor nodes, determine whether the event sensed by individual sensors should be sent off-field. In this case, the fusion nodes (or the individual sensor nodes) forward their information to gateway nodes for off-field transmission. Note that in this system model, we have $N \geq N_f \gg N_g$ where N is the number of sensor nodes, N_f is the number of fusion nodes and N_g is the number of gateway nodes.

The types of traffic patterns that emerge from this behavior are: scoped local broadcast (for initial information sharing) and many-to-one unicast (for sensor node to fusion node communication and for fusion node to gateway node communication). We expect the scoped local broadcast to have the frequency of false positives, and the frequency of the many-to-one unicast communication to be somewhat less; that is, some fraction of the false positive rate corresponding to the reduction in false positives expected from local processing. We also expect that on average, the length of the paths to the fusion nodes will be much shorter than the length of the paths to the gateway nodes.

In the case where active sensors are employed, the traffic patterns to support fusion are more predictable. In response to each ping originated from a sensor in the field, only a subset of the sensor nodes can receive the reflection from the target with enough signal strength for a detection. Therefore, the information that is relevant for fusion can only come from this subset of nodes. Given the specific location of the pinging node, the set of possible receiving nodes can be determined based on the node locations and the active sonar equation with appropriate parameters.[2] The fusion node where responses to an active ping should be transmitted can be selected to minimize the average number of hops from the set of possible receiving nodes to the fusion node to optimize energy efficiency. Furthermore, the design of the pinging strategies (the dynamic scheduling of pinging from the sensors) can take into account both the coverage for detection performance and the energy consumption as a function of the anticipated traffic.

Another type of traffic that the sensor field will have to handle is query/response traffic. This traffic is initiated from outside the sensor field, possibly in response to fused information sent from the field through the gateways. We assume that queries of individual nodes, as well as groups of nodes is possible, and further that these queries may produce relatively large quantities of information in response, perhaps even megabytes if the query is for recent historical (i.e. time-series) data from the sensors.

Queries may target individual fusion nodes, groups of fusion nodes, or geographical neighborhoods of sensor nodes. These different query types produce (respectively) unicast from the gateway to a fusion node and back; multicast from the gateway to fusion nodes and many-to-one unicast patterns back from the fusion nodes to the gateways; limited multicast from the gateway to sensor neighborhood and many-to-one unicast patterns back to the gateway. We assume that the query messages are short, but the unicast responses are relatively large and comprise multiple messages. This type of communication holds the greatest potential for congestion.

By "network control sources" we mean traffic from the protocols that maintain the communication infrastructure of the sensor field. This includes especially routing protocols and gateway- and fusion-node election protocols in addition to clustering protocols. This traffic is most sensitive to the choice of protocol and to mobility in the sensor field.

We expect that as the sensor field is deployed for underwater surveillance, significant drift will occur through the lifetime of the field. As nodes drift into or out of radio range of one another, the communication topology changes and consequently routing information and information about the nearest specialty node (i.e. gateway, fusion or cluster head node) and cluster identity will change. Because the topology is dynamic, path determination may need to be performed as late as possible before information is sent, which would argue for reactive rather than proactive routing protocols.

4. GATEWAY AND FUSION NODE PLANNING

As mentioned in the Section 2, we assume that a small subset of the sensors covering the sensor field (N_g out of N) is equipped with special communications capability to communicate with nodes outside of the sensor field. These sensors play the role of gateways for off-field communication in the sense that all communications into- or out of the field is through these nodes, and the other non-gateway nodes are only capable of sensor-to-sensor communication. Figure 1(a) shows a sensor field with three gateways using a satellite for off-field communication. This allows for a lower cost design by concentrating expensive communication devices in a small subset of the nodes. This section discusses an important problem in designing such gateway-based sensor networks. That is, determining the number of gateway nodes needed and their location in the sensor field. We will show that, in determining the number of gateways, the tradeoff is between performance and cost. As more gateways are deployed, less traffic load is placed on each gateway and its surrounding nodes, resulting in longer network lifetime. However, with a larger number of gateways the network may be too costly to deploy since gateway nodes are more expensive than non-gateway sensor nodes.

Consider the square planar grid shown in Figure 1(b) made up of one gateway and n sensors. In this grid it is assumed that the radio range R is such that each non-boundary sensor has four neighbors within range. It can be easily shown that in such grid the number of sensors that are h hops away from the gateway is $n_h = 4h$ and $n = 2H(H+1)$ where H is the largest hop count (associated with the boundary nodes) . If each node generates sensor data destined for the gateway at a rate of r messages per unit time, it can be shown that a node h hops away from the gateway will need to transmit messages at a rate

$$R_h = \left[\frac{H(H+1)}{2h} - \frac{h-1}{2}\right] r = \left[\frac{n}{n_h} - \frac{h-1}{2}\right] r. \tag{1}$$

Figure 2(a) is a plot of R_h/r as a function of h for different values of H. This figure shows that most of the message forwarding load is placed on the first-hop sensors (sensors that are located one hop from the gateway). Since radio communication is the main source of energy consumption, the network lifetime will be determined by the lifetime of these first-hop sensors, and we will proceed with a first-hop analysis to determine the number of gateways needed to achieve a given lifetime for the sensor field.

The forwarding rate R_1 at the first-hop nodes is made up of two components

$$R_1 = \left(\frac{n}{n_1}\right) r = \left(\frac{n}{n_1} - 1\right) r + r, \tag{2}$$

with the first term corresponding to relayed traffic (received from two-hop nodes) and the second term corresponding to traffic generated from local sensing. Let e_t denote the energy consumed to transmit a message and e_r the energy consumed when relaying a message, which includes receiving the message from a neighbor and transmitting it to a next-hop neighbor. Obviously, $e_r > e_t$, with $e_r - e_t$ reflecting the energy dissipated during the message receive phase. With E being the initial energy supply at each sensor (at time $t = 0$), the lifetime of a first-hop sensor corresponds to the time T when the total dissipated energy is equal to E:

$$\left(\frac{n}{n_1} - 1\right) rTe_r + rTe_t = E$$

and is given by

$$T = \frac{\frac{E}{r}}{\left(\frac{n}{n_1} - 1\right) e_r + e_t}. \tag{3}$$

In order to derive the sensor network lifetime as a function of the number of gateways N_g, consider the following:

- As we discuss below, the sensor field will be organized into a number of gateway areas (one per gateway) and each sensor will be associated with a gateway in such way that the overall communication load is balanced among the gateway areas, which is achieved when $n = N/N_g$. Consequently, first-hop sensors associated with different gateways are expected to have similar lifetimes.
- Even if each gateway area is not a square grid as in the model of Figure 1(b), the first-hop analysis remains valid as long as the loads on each of the first-hop nodes in an area are comparable, and even when the number n_1 of such nodes is different from the regular grid case where $n_1 = 4$.

Substituting $n = N/N_g$ in the above equation we obtain

$$T = \frac{\frac{E}{r} N_g}{e_r \frac{N}{n_1} - (e_r - e_t) N_g}, \tag{4}$$

which is shown graphically in Figure 2(b) for the case $N/n_1 = 10^3$, $e_r/e_t = 1.2$, $E/e_t = 10^6$, and $r = 0.1$ message/sec. For small values of N_g, which is the case in practical situations (the operational range in Figure 2(b) is limited to $N_g < N/n_1$), T is essentially linear in N_g. Given a target lifetime T, the number of gateways can be determined from equation 4 as

$$N_g = \frac{\frac{e_r}{e_t} \frac{N}{n_1} T}{\frac{1}{r} \frac{E}{e_t} + \left(\frac{e_r}{e_t} - 1\right) T} \tag{5}$$

Note that so far we assumed that the selection of the number of gateways is driven more by network lifetime considerations than by network capacity considerations. Obviously, the aggregate bandwidth that can be achieved out of the sensor field increases with the number of gateways, and if capacity could be limiting then a similar analysis can be done to determine the minimum number of gateways needed. However, with enough fusion being performed inside the network, it is more likely that the number of gateways will be driven by lifetime and fault tolerance rather than bandwidth (capacity) considerations. Since gateway nodes are more expensive than sensor nodes, network cost could also play a role in the sense that the number of gateways could be constrained to less than some limit N'_g. In this case, whether the lifetime objective is met depends on whether the lifetime corresponding to N'_g is at least as large as the target one. If not, then the achievable lifetime will be limited to $T(N'_g)$.

Once the number of gateways is determined, the next step in planning the sensor network deployment is to determine where gateways need to be placed. Since the initial energy allocation is the same among the sensors,

the gateways need to be placed in such a way that the load is balanced among the different gateway areas, resulting in first-hop sensors in all areas running out of energy at the same time. Under uniform density, the sensor field is partitioned into areas of similar size with one gateway at the center of each area as shown in Figure 1(a) for an example with three areas. As we discuss in the next section, to enforce this partition during network operation, a protocol for automating the sensor-to-gateway association is needed.

In Section 2, in addition to sensor and gateway nodes we also mentioned another set of special nodes, namely fusion nodes. However, the presence of these nodes and fusion mechanism used also have an impact on the traffic pattern and this effect will play a role in the design of gateway and cluster topologies. In particular, the fusion process consolidates information from many sensors and reduces the message rate sent beyond the fusion node towards the gateway. This will help the nodes between fusion node and the gateway in reducing energy consumption, in particular the nodes that are one hop from the gateway (gateway first-hop nodes), and may even create a more balanced energy consumption pattern. However, it also increases energy consumption on the nodes that become one hop nodes with respect to a fusion node (fusion first-hop nodes) as a result of adding a fusion node.

In order to develop some insights into these issues, consider a sensor network with a gateway area shown in Figure 3(a). Without fusion nodes, the gateway first-hop nodes see their energy drained at a rate essentially equal to $(n/n_1)re_r$. With fusion nodes deployed, the new drain rate D_g is smaller and is dominated by the term $(n_f/n_1)r_f e_r$, where n_f is the number of fusion nodes in the gateway area (N_f is the total number throughout the sensor field) and r_f is the message rate for the fused information. This rate corresponds to fused sensor information from n/n_f sensors, each with rate r (it is expected that the fused rate $r_f << (n/n_f)r$). The drain rate for the fusion first-hop nodes D_f is dominated by $((n/n_f)/n_1)re_r$ due to n/n_f sensors sending at rate r to the fusion node through the n_1 first-hop nodes. Note that we are assuming that routing to the fusion or gateway nodes is such that the total load is balanced among the first-hop nodes.

Since D_g, as a function of n_f, is increasing whereas D_f is decreasing, the optimal value of n_f is the one that results in $D_g = D_f$ so that the network lifetime is maximized by having energy depleted at both types of first-hop nodes at the same time. This optimal value is of the order of $\sqrt{n}$ and is given by $n_f = \sqrt{r/r_f}\sqrt{n}$. Figure 3(d) shows D_g/D_f as a function of n_f when n is constant, such as when n_f is increased from Figure 3(a) to Figure 3(b). However, when n is allowed to scale linearly with n_f ($n = Kn_f$) such as from Figure 3(b) to Figure 3(c), then D_g/D_f is linear in n_f as in Figure 3(e), with $n_f = Kr/r_f$ as the optimal number of fusion nodes in this case.

Note that the above model for fusion traffic is a simplified one, and in order to help reduce the false alarm rate the set of sensors that get fused typically is not static but is a function of target activity. Models based on the specifics of the fusion process and its impact on the network traffic and overall topology will be discussed in a future paper.

5. TOPOLOGY CONTROL, ROUTING, & GATEWAY ASSOCIATION

In a dense sensor network, it is more energy efficient to select a sparser topology for radio communication. Given the set of sensors V, the transmission topology is determined by the graph $G = (V, E)$ where E is the set of edges between any two nodes in V that can communicate when using maximum power. The role of topology control is to conserve energy by computing and maintaining a connected topology T which is a sub-graph of G. Some of the approaches used include power control and relay node control.

With power control, all nodes keep their radio modules operating but with different power levels determined by topology control. By lowering power, some links in E are removed resulting in $T = (V, E')$ where E' is a subset of E. There are a number of studies in this area; they typically use some radio propagation model and assign node power to generate a topology with some bound on the maximum node degree or on the energy stretch factor, defined as the maximum over all node pairs u and v of $E_G(u,v)/E_T(u,v)$ where $E_G(u,v)$ (respectively $E_T(u,v)$) denote the energy of the minimum-energy path between u and v in G (respectively in T).[13,14] Under relay node control, a subset of the nodes are identified as relay nodes and form a connected topology T among them ($T = (V', E')$ where $V' \in V$ and $E' \in E$), the non-relay ones are considered redundant for the purpose of packet forwarding and can be placed in a communications sleep mode (by turning off their radios). A dynamic

topology control protocol is used so that nodes can alternate going into this sleep mode in order to balance energy usage among them. Cluster-based protocols use active neighbor discovery so that nodes can group themselves into clusters identified by a cluster-head with packet forwarding handled by these cluster-heads, and nodes take turns being cluster-heads.[15] Cluster heads may use different frequencies than sensors for communication.

The sparser topology T (with a smaller number of nodes and/or links) identified by topology control is the one used by the routing protocol for path selection (in the extreme case when T is a strict spanning tree no path selection is needed). There are a number of proposals for routing in wireless ad-hoc networks (see[16] for a survey of routing in sensor networks and[17,18] for general discussion of issues that could affect routing). They belong to one of two main categories: proactive or on-demand routing. Proactive routing attempts to keep routing information for all the nodes up to date by advertising topology changes. Each node maintains routes to all reachable destinations at all times, whether or not there is currently any need to deliver packets to those destinations. However the size of the network and the mobility of the nodes present a big challenge. In contrast to proactive algorithms, reactive routing protocols discover routes only when they are needed by flooding route-request messages in the network, these routes are cached and updated on-demand. Hybrid protocols have been proposed that maintain up-to-date routing information about destinations within some neighborhood and use on-demand routing for distant nodes.

As discussed in Section 4, the existence of multiple gateways in the field and the potential difference in target detection activity in different gateway areas (and the resulting uneven energy consumption) calls for a dynamic sensor-to-gateway association protocol (DS-GAP). The DS-GAP is in addition to any topology control and routing mechanisms that may be in place, but these three components can be closely related and intertwined.

While the presence of multiple gateways offers extended lifetime and a higher degree of off-field communication reliability, a sensor will need to select one of the gateways at a time for off-field communication. Based on attributes that include current available energy levels, the distributed protocol dynamically assigns sensors to gateways in such a way that the overall load is balanced among the different gateway areas over the lifetime of the sensor field. The gateways will need to advertise themselves; and sensors choose one of the gateways as their destination gateway. In the uniform density case, choosing the closest gateway (as measured by the minimum hop path to the gateway) should result in the desired partition into gateway areas of similar size. However, this is not most efficient because (1) the sensor field density may not be uniform, and (2) over time the available energy will not be the same in different areas due to different target detection activities. A DS-GAP is needed to reassign sensors among areas when needed to equalize the available energy in the different areas, resulting in extended network lifetime. With such a protocol, the partition into gateway areas will change over time with area sizes expanding or shrinking as required for load balancing. Gateways will need to advertise available energy, which is dominated by the available energy in their first-hop sensors. First-hop sensors will have to inform the gateway of their available energy, this could be done systematically by all nodes as part of their periodic neighbor-to-neighbor hello message exchange. Gateways will then periodically flood a message to all sensors indicating among other things the available energy level. As these messages are relayed they can also record path information such as hop count. A sensor receiving these messages from different gateways can select the most appropriate gateway taking into account advertised available energy and possibly other attributes such as hop count. Flooding of these gateway messages can take advantage of an underlying topology control mechanism where only a subset of the nodes will need to relay the messages. Once a sensor selects a particular gateway, it can either use an existing routing protocol to route its messages to the selected gateway, or it can forward messages along the reverse path that the gateway advertisement message was forwarded on. Among these two approches we now provide more details about the DS-GAP operation under the first approach.

We propose a method for sensor nodes in a field to choose a gateway in order to balance and minimize energy consumption in the network. This method uses several different mechanisms in order to accomplish its task, including application level communication and modifications to routing protocols.

Since we assume that not every node has gateway capability, the first task is to announce which nodes have the capability. Because the sensor field could be quite large (on the order of 1000-10000 nodes) a general broadcast is not suitable. Instead, the gateway announcement broadcast is a scoped, low-duty-cycle broadcast. We use a standard flooding broadcast algorithm for the announcement. Gateway announcements are small messages that contain an address of the gateway, a scope, and a timestamp (i.e. an increasing sequence number). For

the receivers of the announcement, a message with a larger sequence number always replaces stored information. Messages with a sequence number smaller than or equal to that of the stored information are discarded. If the information from the message is fresh, the scope field of the message is decreased and the message is propagated to all neighbors of the receiver. When the scope field expires, the message is discarded.

Two key parameters in this protocol are the scope and the frequency of announcement messages. The frequency is directly related to node mobility in the sensor field. If the sensors and gateways are not mobile, then the announcement frequency can be set when the sensor field is deployed to be very infrequent. Announcements only have to be made when a gateway loses power and passes its responsibilities to a backup. On the other hand, if the sensors and gateways are mobile, then the frequency of announcement depends upon node density in the area covered by the gateway and the speed with which the gateway and the sensor nodes move relative to one another. If the sensor field is dense and the gateway and sensor nodes are moving fast relative to one another, then announcements should be more frequent since many new nodes will be coming into range in a short period of time. If the sensor field is sparse and the nodes are not moving fast, then announcements should wait until the gateway has traveled roughly the distance of the radius of its radio reach. Additionally the scope parameter limits the extent of the announcement broadcast and depends on the density of the nodes in the sensor field. We can compute these parameters given a particular sensor field topology, density and a model of node mobility.

As an example of how a MANET routing protocol can be modified to provide the information necessary for sensor nodes to choose a gateway, we provide a brief analysis of the Ad-Hoc On-Demand Distance Vector (AODV) routing protocol.[19,20] We chose AODV for this example because it is a much-studied reactive routing protocol. Although we chose AODV for this example, almost any of the routing protocols currently considered in the IETF MANET working group could be modified to produce the same information.

The AODV protocol operates by using a route discovery broadcast (Route Request) followed by intermediate nodes processing Route Reply packets in order to build forward and reverse paths through the network to the destination. As the broadcast Route Request moves from the source to the destination, the intermediate nodes build a reverse path to the source of the request. When replies are generated (either from the destination node itself or from intermediate nodes having a fresh route to the destination), they are unicast back to the source of the original route request. During this process, if a node receives multiple replies, it drops the replies when it has a better route, and forwards replies that have a better route than the intermediate node.

AODV messages already provide hop counts, which is useful information for gateway association. We would propose that the route replies carry also a list of node IDs and energy levels of the nodes that the reply passes through. When an intermediate node receives a route reply traveling upstream, it modifies the behavior currently specified by AODV in that instead of dropping route replies that have the same sequence number and hop count as locally stored information, the intermediate node checks the list of nodal energy levels and if different from what it stored but with different hop count, it stores also the new route reply and the ID of the neighbor that it came from. Note that for the sake of load balancing the energy consumption, a longer path (in terms of hop count) may be preferred over a shorter one when more energy is available along the longer path. The intermediate node also adds its node ID and energy level to the list in the route reply. In this way, there are a number of alternate paths, with differing energy levels that branch at a given intermediate node and that offer alternate paths to the source with different energy use characteristics. The source node that sent the original route request then receives a somewhat pruned set of paths, but with the information of the remaining energy levels of the nodes in each path.

If an intermediate node chooses to reply to a route request, instead of forwarding it (implying that the intermediate node has a relatively fresh route to the destination), then the nodal energy information stored along the route to the destination should be reduced before the route reply is sent upstream to the source. The reduction should be dependent on the time that the route has been retained, and the number of hops from the gateway. The time that the route has been retained can be coupled with the average reporting rate from nodes to the gateway, and the node density in the field. This discounted energy information can be replaced when fresher route information is obtained from another route request.

On receiving a new gateway announcement, sensor nodes acquire information about the path to the gateway by trying to find a route to the gateway. In dense sensor networks especially, it is better to avoid situations

where large numbers of nodes broadcast packets simultaneously or nearly so. One solution to this problem is that when sensor nodes receive an announcement, they choose a random time to wait before they probe for a path to the new gateway. This technique should also reduce the total traffic resulting from a gateway announcement because as later nodes request routes to the gateway, routes from earlier requests should alreday exist and will be returned by intermediate nodes.

Once a given source node has the routes and nodal energy information, the next phase is for the sensor node to select a gateway. Several different strategies are possible, and here we give several examples. One strategy is to choose the gateway where the path(s) to the gateway have the largest average energy reserve. In other words, where e_i is the remaining energy at node n_i, $\{g_k\}$ is the set of gateway nodes, s is the sensor node, p_k is the path from s to g_k and N_k is the length of p_k:

$$G^* = \max_{g_k}\{\frac{1}{N_k}\sum_{n_i \in p_k} e_i\}.$$

Another strategy is to choose the gateway with the largest average nodal energy one hop away from the gateway. This would require that the source examine all the paths to a given gateway and extract the one-hop nodal energy levels. That is, where $h_k = \{n_i | n_i \text{ is one hop from } g_k\}$, and $M_k = |h_k|$:

$$G^* = \max_{g_k}\{\frac{1}{M_k}\sum_{n_i \in h_k} e_i\}.$$

Yet another strategy is to select a gateway where the path(s) have the smallest average difference in energy levels between nodes on the path. Let $d_k = \{\delta = e_i - e_j | e_i, e_j \in p_k \wedge i \neq j\}$ and $P_k = |d_k|$:

$$G^* = \min_{g_k}\{\frac{1}{P_k}\sum_{i=1}^{P_k} \delta_i\}.$$

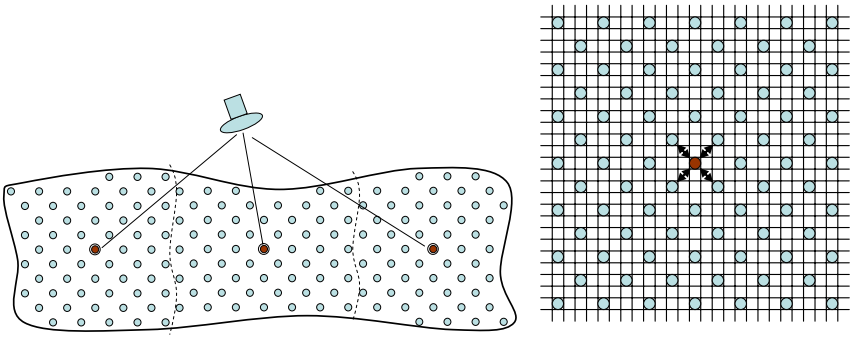
The strategy used to choose a gateway should also be incorporated into the routing protocol to affect how the routing decision is made hop-by-hop. Normally, the MANET routing protocols make routing decisions based on the least number of hops to the destination. In the AODV example, assuming that multiple routes are cached at each node, the routing decision would have to be modified according to the optimization criteria chosen. Sensor-to-gateway traffic could in fact be treated differently than other traffic (the type of traffic in the sensor field could be indicated by a field in the packet header; in the case of IP, the DSCP bits could be used to differentiate traffic for the routing algorithm).

In summary, by simple modifications to existing MANET routing protocols, additional information can be gathered that can greatly enhance the choice of gateway for sensor nodes in the field. In a subsequent paper, we will report on the analysis of these ideas.

REFERENCES

1. B. Doshi, L. Benmohamed, P. Chimento, and I-J. Wang, "Sensor fusion for coastal waters surveillance," *Proceedings of SPIE Multisensor & Multisource Information Fusion Conference*, Orlando, FL, March 2005.
2. T. M. Higgins, A. E. Turriff, and D. M. Patrone, "Simulation-Based Undersea Warfare Assessment," *Johns Hopkins APL Technical Digest*, 2002, vol. 23, no. 4, pp. 396–402.
3. F. E. White, "A model for data fusion," in *Proceedings of the 1st National Symposium on Sensor Fusion*, 1988.
4. A. Steinberg, C. Bowman, and F. White, "Revisions to the JDL data fusion model," in *Proceedings of the SPIE Sensor Fusion: Architecture, Algorithms, and Applications*, **3719**, 1999.
5. D. L. Hall and J. Llinas, *Handbook of Multisensor Data Fusion*, CRC Press, Boca Raton, 2001.
6. R. Niu, P. Varshney, M. Moore, and D. Klamer, "Decision fusion in a wireless sensor network with a large number of sensors," in *Proceedings of the Seventh International Conference on Information Fusion*, **1**, pp. 21–27, (Stockholm, Sweden), June 2004.

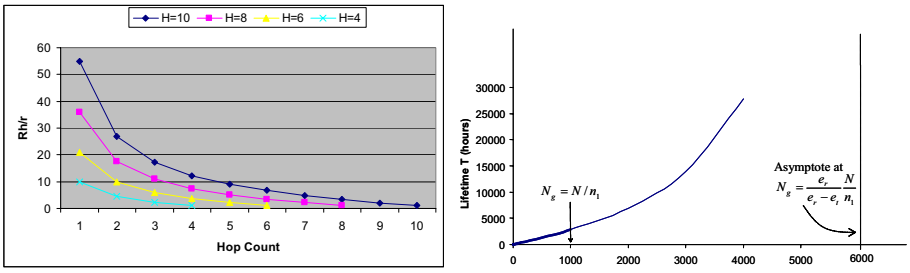
7. J.-F. Chamberland and V. V. Veeraalli, "Asymptotic results for decentralized detection in power constrained wireless sensor networks," *IEEE Journal on Selected Areas in Communications* **22**, pp. 1007–1015, August 2004.
8. J. Chen, L. Yip, J. Elson, H. Wang, D. Maniezzo, R. Hudson, K. Yao, and D. Estrin, "Coherent acoustic array processing and localization on wiireless sensor networks," *Proceedings of the IEEE* **91**, pp. 1154–1161, August 2003.
9. G. Barriac, R. Mudumbai, and U. Madhow, "Distributed beamforming for information transfer in sensor networks," in *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks (IPSN'04)*, (Berkeley, CA), April 2004.
10. J.-F. Chamberland and V. V. Veeraalli, "Decentralized detection in sensor networks," *IEEE Transactions on Signal Processing* **51**, pp. 407–416, February 2003.
11. G. Polychronopoulos and J. N. Tsitsiklis, "Explicit solutions for some simple decentralized detection problems," *IEEE Transactions on Aerospace and Electronic Systems* **26**, pp. 282–292, March 1990.
12. W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of American Statistical Association* **58**, pp. 13–30, 1963.
13. V. Rodoplu and T. Meng, "Minimum energy mobile wireless networks," *IEEE Journal on Selected Areas in Communications* **17**, pp. 1333–1344, August 1999.
14. R. Wattenhofer, L. Li, P. Bahl, and Y. Wang, "Distributed topology control for power efficient operation in multihop wireless ad hoc networks," in *Proceedings of INFOCOM 2001 Volume 3*, pp. 1388–1397, IEEE, IEEE, April 2001.
15. Y. Xu, S. Bien, Y. Mori, J. Heideman, and D. Estrin, "Topology control protocols to conserve energy in wireless ad hoc netowrks," tech. rep., USC/ISI, 2003.
16. J. Al-Karaki and A. Kamal, "Routing techniques in wireless sensor networks: a survey," *IEEE Wireless Communications* **11**, pp. 6–28, December 2004.
17. I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *IEEECommunications Magazine* **40**, pp. 102–114, August 2002.
18. K. Romer and F. Mattern, "The design space of wireless sensor networks," *IEEE Wireless Communications* **11**, December 2004.
19. Perkins, C. and Royer, E., "Ad-hoc On-Demand Distance Vector Routing," *2nd IEEE Workshop on Mobile Computing Systems and Applications*, pp. 90–100, New Orleans, LA, Feb. 1999.
20. Perkins, C. and Belding-Royer, E. and Das, S", "Ad Hoc On Demand Distance Vector (AODV) Routing," IETF Request for Comments, RFC 3561, July 2003.



(a) Sensor field with three gateways

(b) Sensors in one gateway area

Figure 1. A uniform grid sensor field



(a) Forwarding rate as a function of distance from gateway

(b) Network lifetime as a function of number of gateways

Figure 2. Network performance

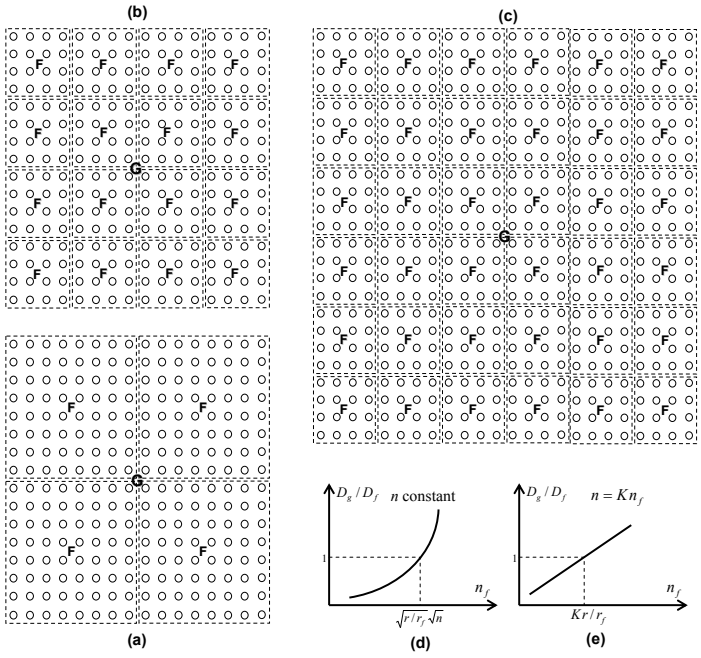


Figure 3. Gateway area with fusion nodes, F = fusion node, G = gateway node.