This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2010 proceedings

An Efficient Data Aggregation Scheme Using Degree of Dependence on Clusters in WSNs

Tetsushi Fukabori*§, Hidehisa Nakayama†, Hiroki Nishiyama*, Nirwan Ansari‡, and Nei Kato*
*Graduate School of Information Sciences, Tohoku University, Sendai, Japan
†Tohoku Institute of Technology, Sendai, Japan
‡Advanced Networking Lab., ECE Department, New Jersey Institute of Technology, Newark, NJ, USA
Email: §teslur@it.ecei.tohoku.ac.jp

***Abstract*—Recently, much effort aiming at achieving ubiquitous networks has been made. A ubiquitous network refers to a network environment, which enables anytime and anywhere access, by possibly any given device or by any user. In a ubiquitous network, applications require many types of information such as temperature and so forth. A great deal of attention has been paid to aggregate this information in Wireless Sensor Networks (WSNs). A WSN consists of tiny nodes comprising sensing and communication devices. The information sensed by each node is relayed via other nodes in the WSN to the destination node called the "sink". One of the most significant challenges pertaining to any WSN is to reduce the energy consumption of its nodes, which run on scarce battery resources. An effective scheme to reduce this energy consumption is to exploit the sink node's mobility, which however presents new challenges to the sink node's routing and information aggregation. In this paper, we propose a new routing and data aggregation scheme based on clustering. Simulation results demonstrate that our scheme can provide better energy efficient data aggregation as compared to the KAT (K-means And Traveling salesman path) mobility.**

I. Introduction

Technological advances in broadband wireless communications, coupled with cost reduction of wireless devices, are empowering a real possibility to design and deploy ubiquitous networks. A ubiquitous network refers to a network environment, which enables access anytime, anywhere, by any device, and by anyone. Recently, many applications on ubiquitous networks have been developed such as traffic control by using car navigator networks, monitoring wild animals, and data aggregation under natural disasters [1]–[3]. In the context of these applications, the aggregation of information such as temperature, humidity, and atmospheric pressure is, indeed, an important issue. Many studies have focused on using Wireless Sensor Networks (WSNs) to aggregate this information in the past few years. A WSN is a wireless network consisting of a large pool of sensor nodes. Each sensor node in a WSN is equipped with sensing and wireless communication devices for sensing and transmitting various information. The overall information gathered by all the nodes is aggregated to a particular destination node known as the sink. The WSN nodes are driven by battery power, which is typically scarce. Once the battery life of a node expires, that node can no longer sense or transmit. Therefore, the reduction of energy consumption is one of the most critical challenges in a WSN. Over the years, much research has been aiming at reducing energy consumptions of the WSN nodes [4], and the deployment of mobile sinks is one of the leading techniques.

In deploying a mobile sink for data gathering, the sink node moves within the target sensing area and aggregates information collected from the nodes deployed in the considered WSN. In WSNs that do not employ the mobile sink, the position of the sink node is fixed and the information sensed by a node far away from the sink is relayed via many intermediate nodes. This results in excessive communications and eventually leads to higher energy consumption. In addition, nodes that lie near the sink node may often relay information from other nodes, and they may consequently exhaust the battery power quickly. These problems may shorten the lifetime of the network. On the other hand, these problems can be mitigated by employing a mobile sink [5]–[8]. Therefore, the mobile sink scheme (i.e., the scheme that employs mobile sinks) can reduce energy consumption and improve the network lifetime. However, this scheme is not immune from other shortcomings. For instance, the mobile sink scheme may contribute to increased delay because the physical mobility of the sink is much slower as compared to electronic/wireless communications. It is worth noting that the battery power of a WSN is consumed over time even if it is not communicating with other nodes.

In this paper, we focus on the efficiency of data aggregation, and propose a new and effective data aggregation scheme based on clustering. Our scheme consists of three parts, namely, the clustering of nodes by using the Expectation-Maximization (EM) algorithm, generating trajectory of the sink node by using the solution to Traveling Salesman Problem (TSP), and aggregating the data collected from the WSN nodes by using a cluster adapted Directed Diffusion mechanism.

The remainder of this paper is organized as follows. Section II describes some related works. Section III elucidates the proposed algorithm. Simulation results are presented in Section IV. Finally, concluding remarks are provided in Section V.

II. Related works

In recent literature, many studies have already focused on data aggregation with mobile sink(s) in WSNs. Shah *et al.* [5] proposed the data aggregation scheme with random walk of the mobile sink node. This scheme called "Data MULEs" aims at saving power consumption. Therefore, data transmission commences only when the mobile sink reaches the proximity

978-1-4244-6404-3/10/$26.00 ©2010 IEEE

of the sensor nodes. The objectives of the mobile sink schemes proposed in [6], [7] are also similar. In [6], the mobile sink scheme achieves four times the network lifetime in contrast to the WSN configuration where the sink remains static. In [7], the mobility of the sink results in smooth energy consumption of the WSN nodes, and this is also shown to prolong the network lifetime.

Finding suitable routes of a mobile sink is critical to reduce energy consumption and prolong network lifetime. Consequently, many studies in WSNs have been carried out pertaining to the network layer such as LEACH (Low-Energy Adaptive Clustering Hierarchy) [9] and Directed Diffusion [10]. The former is a self-organizing, adaptive clustering protocol. To reduce energy consumption, nodes in LEACH are grouped into a number of clusters based on their battery usage. Each of these clusters consists of a cluster head, with which every node belonging to that cluster communicates. The sink aggregates data, relayed via cluster heads, from other nodes. Since it consumes higher energy due to its frequent transmission, the cluster head is sometimes re-selected based on the remaining energy level. As a consequence, LEACH can prolong the network lifetime because nodes having higher remaining energy are preferred to serve as cluster heads. On the other hand, due to its simplicity, Directed Diffusion is one of the most commonly used data aggregation protocols in WSNs. While many studies have proposed modified models of the Directed Diffusion [11], [12] approach, the One Phase Pull model [13] remains the simplest one. In this model, Directed Diffusion uses two types of messages, namely, the "Interest" message and the actual data messages. To aggregate data by using Directed Diffusion, the sink node broadcasts an "Interest" message that contains a time-to-live value, and also the addresses of the source and destination nodes. Upon receiving this request, the destination node transmits to the source, the appropriate data-messages, which contain the sensed data. When more destination nodes in the "Interest" message cannot be reached by the current source but downstream from the current destination node, the current destination node then changes the source node address to its own, reduces the time-to-live value, and rebroadcasts the "Interest" message.

KAT mobility [8] is a cluster based data aggregation scheme, which uses the K-means algorithm for clustering, TSP (Traveling Salesman Problem) for minimizing traveling time, and Directed Diffusion for data aggregation. KAT mobility aims at not only reducing energy consumption but also at increasing the efficiency, i.e., the ratio of the aggregated data volume to the consumed energy. KAT mobility creates clusters based on the positions of the nodes in the considered WSN. Therefore, when some nodes become inactive due to degraded battery, KAT mobility can re-evaluate the trajectory of the mobile sink node and select a more suitable path for performing data aggregation. Thus, KAT mobility achieves fault-tolerance and attains higher efficiency than other conventional mobile sink schemes that employ random walk or fixed trajectory.

III. Proposed scheme

First of all, we focus on KAT mobility because of its high efficiency. However, KAT mobility is not immune from its shortcomings. First, this scheme uses the K-means algorithm, which attempts to minimize the sum of squares of the distances between the nodes and the cluster centroids, based on appropriate assumptions. Because of these assumptions, the K-means algorithm may sometimes fail to minimize these distances. Second, KAT mobility does not take into account the form of the cluster while aggregating data. This may cause too many data relays and eventually lead to high energy consumption. These issues may decrease the efficiency of the KAT mobility scheme. In this paper, we propose a new data aggregation approach based on clustering to ameliorate these shortcomings of KAT mobility. We assume that the mobile sink knows every node's position and can use the algorithms that require geographic information. KAT mobility also functions under the same assumption. In our scheme, the nodes in the sensing area are grouped into K clusters by using the EM algorithm, and the mobile sink traces the trajectory of TSP through these cluster centroids. To aggregate data efficiently, the mobile sink and nodes use cluster adapted Directed Diffusion. We will explain our scheme, in detail, in the remainder of this section.

A. Clustering algorithm

In order to reduce energy consumption in WSNs, transmission distance is one of the most important issues because the required power of wireless transmission is proportional to the square of the transmission distance. The EM algorithm includes minimizing the sum of squares of the distances between nodes and cluster centroids. Moreover, the EM algorithm is constructed under less-strict assumptions than those of K-means algorithm. Therefore, we use the EM algorithm to group the WSN nodes into K clusters to reduce energy consumption. The two-dimensional EM algorithm is only based on an assumption that nodes are distributed according to a two-dimensional Gaussian distribution, known as the Gaussian Mixture Model (GMM), i.e.,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{1}$$

where,

$$\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right\} \tag{2}$$

K: The number of clusters
π_k: The mixing coefficients of the k^{th} cluster
$\boldsymbol{\mu}_k$: The 2-dimensionaol vector indicates the mean of the k^{th} cluster
$\boldsymbol{\Sigma}_k$: The 2×2 covariance matrix of the k^{th} cluster

This distribution of nodes often appears in the actual world. We consider that nodes are scattered from high altitude,

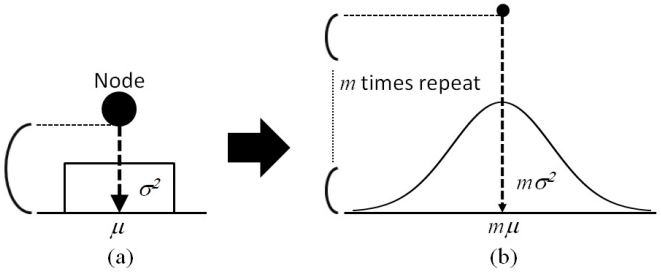

Fig. 1. (a): Probability of locomotion in one step free falling. (b): Probability of locomotion in overall (m steps) free falling.

and free falling nodes can be modeled as m times repeat of locomotion following independent and identical distributions where the mean and variance are denoted by μ and σ^2, respectively (Fig. 1. (a)). In this situation, it is already known from the central limit theorem that the nodes are distributed according to the Gaussian distribution with the mean $\mu_{\text{overall}} = m\mu$ and the variance $\sigma^2_{\text{overall}} = m\sigma^2$ (Fig. 1. (b)). Based on the assumption of this distribution of nodes, we use the EM algorithm to find the maximum likelihood estimates of $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_k\}$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k\}$, and $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_k\}$. In our proposed scheme, all these parameters are used for determining the trajectory and enabling cluster adapted Directed Diffusion.

B. The trajectory of the mobile sink

After clustering the WSN nodes, we will determine the actual trajectory of the mobile sink node. The mobile sink traverses through clusters and aggregates data from various nodes. Since it is possible to increase efficiency by shortening the traveling time, it is preferable that the mobile sink traces the shortest path among the cluster centroids. Therefore, we use the solution of TSP as the trajectory. Since TSP is an NP-hard problem, we resort to an approximate solution of TSP as the trajectory. In essence, the evaluation of the mobile node's trajectory is similar to that used by the KAT mobility scheme.

C. Cluster adapted Directed Diffusion

After partitioning the WSN nodes into K clusters according to GMM and determining the trajectory of the mobile sink, we now address the issue of aggregating the sensed data collected from the WSN nodes. For this purpose, we consider using the Directed Diffusion approach under the already stated assumption pertaining to the nodes' distribution. As the mobile sink transmits the "Interest" message, it may only reach up to the centroid of the next cluster rather than the edge of that cluster. A message broadcasted in this fashion using the time-to-live value spreads in a circle. However, the two-dimensional Gaussian distribution may be distributed in a circular shape as well as an elliptical shape. As a consequence, since the nodes are distributed following the GMM, the density of the nodes increases with the proximity to the cluster centroid as implied by Eq. (1). Therefore, if the "Interest" message from the k^{th} cluster centroid reaches that of another cluster, an excessive number of nodes may send back data messages far from the k^{th} cluster centroid. This causes an overwhelming volume of communications among the WSN nodes, thus resulting in high energy consumptions. To resolve this issue, we propose the cluster adapted Directed Diffusion. In GMM, each data point (i.e., a node) may be considered as part of the different mixes of distributions, and thus its contribution to a respective distribution is weighed with a certain ratio. This ratio is referred to as the "degree of dependence" $\gamma(z_{nk})$, and is expressed as follows.

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \tag{3}$$

If the node n lies near the centroid of the k^{th} cluster, the value of degree of dependence $\gamma(z_{nk})$ is greater than that in case of other nodes lying farther away from the centroid. We propose probabilistic response to improve efficiency by using this degree of dependence metric. Thus, in our proposed scheme, each node evaluates its own degree of dependence, $\gamma(z_{nk})$. When a node receives an "Interest" message from the centroid of the k^{th} cluster, it checks the following condition:

$$\gamma(z_{nk}) \geq r, \tag{4}$$

where r is chosen randomly from the (0,1]. If this condition is satisfied, the node transmits the data message to the sink node. Otherwise, the node will not transmit any sensed data. In this scheme, when a node lying near the k^{th} cluster centroid receives an "Interest" message from the next cluster, the node usually does not send back data messages because the node's value of degree of dependence on the k^{th} cluster almost equals to one, and almost zero with respect to other clusters. Thus, even if a node receives an "Interest" message from far away nodes, the previously described problem can be dealt with effectively by using our scheme.

D. Summary of the proposed scheme

Our proposed data aggregation scheme based on clustering can be summarized as follows:

Step 1. The mobile sink node groups all nodes into K clusters by using the EM algorithm in the following manner.

(1) Initialize $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$ and the convergence criterion θ_{EM}, and evaluate the initial value of the log likelihood $\mathcal{P}$:

$$\mathcal{P} = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}, \tag{5}$$

where N is the number of nodes.

(2) Evaluate the degree of dependences expressed in Eq. (3) by using the current parameter values.

(3) Re-estimate the parameters by using the current degree of dependences:

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \tag{6}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathcal{D}_{nk} \tag{7}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{8}$$

where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}) \tag{9}$$

$$\mathcal{D}_{nk} = (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^{\text{T}} \tag{10}$$

(4) Evaluate the log likelihood $\mathcal{P}^{\text{new}}$ by using re-estimated parameters, and if

$$|\mathcal{P} - \mathcal{P}^{\text{new}}| < \theta_{\text{EM}} \tag{11}$$

is not satisfied, return to (2).

Step 2. Determine the trajectory by using the approximate solution of TSP.

Step 3. Aggregate data from nodes by using cluster adapted Directed Diffusion as follows.

(1) All nodes evaluate their own degree of dependences on all clusters, and initialize θ_{Interest}.

(2) When the mobile sink reaches the cluster centroid, it broadcasts an "Interest" message.

(3) When the node receives the "Interest" message, if Eq. (4) is satisfied, the node sends back data messages to the mobile sink node. At the same time, if the following condition is satisfied, the node re-broadcasts the "Interest" message.

$$\gamma(z_{nk}) \geq \theta_{\text{Interest}} \tag{12}$$

Otherwise, the node drops the "Interest" message and does not re-broadcast.

IV. Performance evaluation

We have used Qualnet Simulator (Ver. 3.9.5) to evaluate the effectiveness of our proposed scheme in WSNs.

A. Network and battery models

We have used TCP/IP based networks. The universal standard (IEEE 802.11b) is adopted for designing the physical and link layers for constructing the wireless networks. The cluster adapted Directed Diffusion mechanism, as described in Section III, is used to formulate the network layer. The transmission rate is set to two Mbps. The battery model is adopted from the KAT mobility [8].

B. Parameter settings

In the first experiment, nodes are uniformly distributed, and those of the second experiment are distributed according to the Gaussian mixture distribution. The number of Gaussian random variables used to model the GMM in the second experiment is set to four. Common parameters are set as follows. An area of $5 \times 5\text{km}^2$ is considered. The mobile sink node's velocity is varied from 20 m/s to 30 m/s arbitrarily. In addition, the mobile sink pauses at the cluster centroids for 20 s. The mobile sink broadcasts the "Interest" message every 3s to cluster centroids. Each sensor is equipped with a buffer of 10 MB, and is able to generate a constant rate of data at 512 B/s. The number of clusters K in both schemes is fixed at 10. The simulations are conducted by varying the population of WSN nodes from 20 to 200, and the simulation time is set to 80 min. The simulations are repeated 40 times by using different seeds of random digits, and each simulation run uses the same parameters. Averages of all the simulation runs are used as results. The values of θ_{EM} and θ_{Interest} are set to 1.0×10^{-15} and 0.3, respectively.

C. Performance metrics

To evaluate efficiency of the proposed scheme, we use the efficiency metric, denoted by E, originally proposed in KAT mobility [8]. E is expressed as follows.

$$E = \frac{\text{(Received Bytes by all mobile sinks)}}{\text{(Consumed Energy by all sensors)} \times N} \tag{13}$$

This metric refers to the volume of aggregated data received by mobile sinks per unit consumed energy and per node. Therefore, the more data aggregated and the less energy consumed, the higher efficiency is achieved. To understand the efficiency improvement in the proposed approach easily, the relative improvement is quantified by the following ratio:

$$E_{\text{ratio}} = \frac{\text{(The value of } E \text{ of the target scheme)}}{\text{(The value of } E \text{ of KAT mobility)}} \tag{14}$$

Therefore, the value of E_{ratio} of KAT mobility is always one and that of our proposed scheme appears as the efficiency ratio of the proposed mechanism and KAT mobility.

D. Experimental results

Fig. 2 demonstrates the efficiency when nodes are distributed according to the uniform distribution. The efficiency of the proposed scheme starts to improve when N exceeds 80 and eventually reaches 1.5 times that of KAT mobility. When the number of nodes is less than 60, the efficiency of the proposed scheme is less than that of KAT mobility because the node density is too sparse and some nodes are beyond the wireless communication range. When the number of nodes is larger than 60, the proposed scheme can improve the efficiency. Even if the nodes are distributed according to the uniform distribution, there is still some distinction of the nodes' density, e.g., the nodes are distributed sparsely in some parts of the field, but densely in some other parts, somehow

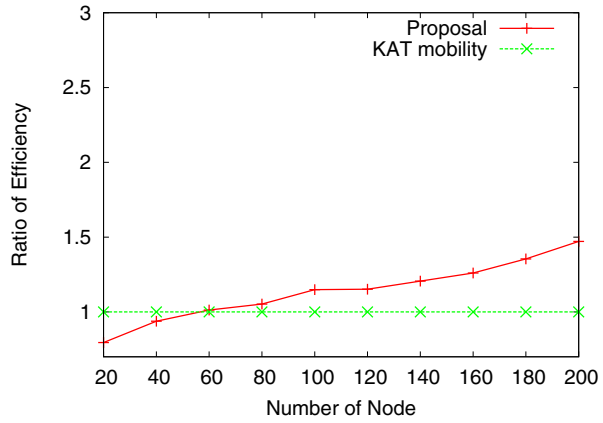

Fig. 2. Ratio of efficiency when nodes are distributed according to the uniform distribution

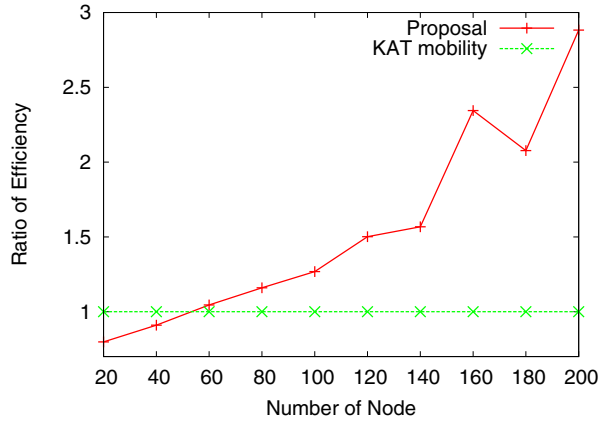

Fig. 3. Ratio of efficiency when nodes are distributed according to the Gaussian mixture distribution

following the Gaussian mixture distribution. Therefore, the proposed scheme works more efficiently than KAT mobility.

Fig. 3 depicts the efficiency when nodes are distributed according to the Gaussian mixture distribution. The efficiency of the proposed scheme surpasses KAT mobility at around $N = 60$, and finally reaches 2.9 times that of KAT mobility. When the number of nodes is less than 60, the efficiency is not improved because of the reason described above. When the number of nodes is larger than 60, the proposed scheme improves the efficiency due to the distribution of the nodes. As described in Section III, Fig. 3 demonstrates the efficiency when the nodes are dropped from the planes or helicopters and are scattered. In this experiment, the locations of the nodes are shown in Fig. 4. This mixture distribution consists of four Gaussian distributions indicated as green circles in Fig. 4, and the nodes follow these particular distributions in each experiment.

V. Conclusion

In this paper, we have proposed a new data aggregation scheme based on clustering in WSNs. The simulation results verify that our proposed scheme can increase efficiency for

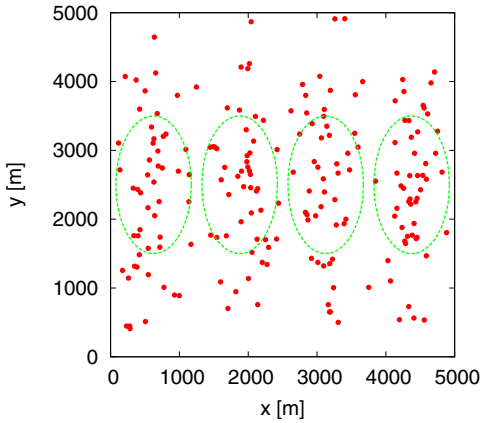

Fig. 4. Location of nodes when nodes are distributed according to the Gaussian mixture distribution

different numbers and distributions of nodes. Moreover, if the nodes are distributed according to the Gaussian mixture distribution, our proposed scheme can reach up to 2.9 times that of the previously proposed KAT mobility. Since nodes are scattered from high altitude according to the Gaussian mixture distribution, the proposed scheme, which is developed based on this practical assumption, is readily applicable.

References

[1] M. Sichitiu and M. Kihl, "Inter-vehicle communication systems: a survey," *IEEE Commun. Surveys Tuts.*, vol. 10, no. 2, pp. 88–105, 2008.

[2] P. Juang, H. Oki, Y. Wang, M. Martonosi, L.-S. Peh, and D. Rubenstein, "Energy-efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet," *Proc. ASPLOS-X*, pp. 96–107, Oct. 2002.

[3] F. Chiti, R. Fantacci, L. Maccari, D. Marabissi, and D. Tarchi, "A broadband wireless communications system for emergency management," *IEEE Wireless Communications*, vol. 15, no. 3, pp. 8–14, Jun. 2008.

[4] N. Pantazis and D. Vergados, "A survey on power control issues in wireless sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 9, no. 4, pp. 86–107, 2007.

[5] R. Shah, S. Roy, S. Jain, and W. Brunette, "Data mules: modeling and analysis of a three-tier architecture for sparse sensor networks," *Proc. IEEE SNPA*, pp. 30–41, May 2003.

[6] W. Wang, V. Srinivasan, and K.-C. Chua, "Extending the lifetime of wireless sensor networks through mobile relays," *IEEE/ACM Trans. Netw.*, vol. 16, no. 5, Oct. 2008.

[7] M. Marta and M. Cardei, "Using sink mobility to increase wireless sensor networks lifetime," *WoWMoM 2008*, pp. 1–10, Jun. 2008.

[8] H. Nakayama, N. Ansari, A. Jamalipour, and N. Kato, "Fault-resilient sensing in wireless sensor networks," *Computer Communications archive*, vol. 30, no. 11-12, pp. 2375–2384, 2007.

[9] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 660–670, Oct. 2002.

[10] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed diffusion for wireless sensor networking," *IEEE Netw.*, vol. 11, no. 1, pp. 2–16, Feb. 2003.

[11] A. Marcucci, M. Nati, C. Petrioli, and A. Vitaletti, "Directed diffusion light: low overhead data dissemination in wireless sensor networks," *IEEE VTC 2005-Spring*, vol. 4, pp. 2538–2545, 2005.

[12] C. Jisul and K. Keecheon, "Eadd: Energy aware directed diffusion for wireless sensor networks," *ISPA 2008*, pp. 779–783, Dec. 2008.

[13] B. Krishnamachari and J. Heidemann, "Application-specific modeling of information routing in sensor networks," *Proc. IPCCC 2004*, pp. 717–722, 2004.