# An Efficient Approach for Data Aggregation Routing using Survival Analysis in Wireless Sensor Networks

Dr.B.Vinayaga Sundaram, Rajesh G, Khaja Muhaiyadeen A, Hari Narayanan R, Shelton Paul Infant C,Sahiti G, Malathi R, Mary Priyanga S

Department of Information Technology, Anna University, Chennai
bvsundaram@annauniv.edu, raajiimegce@gmail.com, khaja.it@gmail.com, hari.zlatan@gmail.com, sheltonpaul89@gmail.com, sahiti.mit@gmail.com, angelmalathi@gmail.com, marypriyanga@yahoo.com

***ABSTRACT***

*Wireless Sensor Network (WSN) is a collection of small sensor nodes with a communications infrastructure to achieve mutual communication and to monitor and record conditions at diverse locations. The major constraints of WSN are limited availability of power and it is prone to frequent node failures. In order to prolong the lifetime of the sensor nodes, the sensor data should efficiently reach the base station and there should be a reduction in message transmission, which consumes the majority of the battery power. Aggregation of data at intermediate sensor nodes helps in saving the energy that would be spent if the nodes send directly to the base station. In addition to aggregation, the mechanism to overcome node failures is also essential to ensure the successful delivery of the data packets to the base station. This paper proposes an efficient way based on inverse-square law along with survival analysis for aggregating data without the formation of an explicit structure and to overcome node failures. Our evaluation of performance shows a considerable decrease in the number of transmissions required to carry the sensed data to the base station and also a considerable increase in packet delivery ratio. By using this approach, it is possible to minimize power usage of sensor nodes effectively.*

***KEYWORDS***

*Inverse Square law, Survival Analysis, Kaplan-Meier estimator, Enhanced Random Delay, Wireless Sensor Networks, Node failure*

## 1. INTRODUCTION

Wireless Sensor Network (WSN) consists of spatially distributed autonomous sensors to cooperatively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants. A number of sensor nodes are densely deployed in a field of interest and they observe the phenomena at different points in the field, which are sent to a data sink or base station, located either at the centre or out of the field, for processing [8]. Wireless sensor networks are now used in many civilian application areas, including environment and habitat monitoring, health applications, home automation, and traffic control. Size and cost constraints on sensor nodes result in corresponding constraints on resources such as energy, memory, computational speed and bandwidth. Specific applications for WSNs include habitat monitoring, object tracking, nuclear reactor control, fire detection, and traffic monitoring. In a typical application, a WSN is scattered in a region where it is meant to collect data through its sensor nodes. Instead of directly sending the data to the sink, it is highly desirable to aggregate the data through effective data-aggregation techniques to minimize the power consumption of the sensor nodes during data transmission and thereby prolonging the overall network lifetime. Since the sensor nodes cannot be recharged, the sensed data should reach the base station through multihop routing. Many approaches were common. Mostly, the

DOI : 10.5121/ijwmn.2011.3219

sensor network is grouped into clusters or a tree like structure and data moves up hierarchically to the base station. Here those nodes in top of the hierarchy, gets affected more. In order to handle this, we propose the structure-less approach where data gets aggregated quickly. A two step framework is used. In first, node that is more probable to contain data is found and then the time for which a node must wait for data to arrive from other nodes is found. We show the gain in performance by means of the reduction in number of transmissions in the network. Finally, our simulation results will also substantiate our claim of the gain in performance. In addition, sensor nodes in WSNs are prone to failure due to energy depletion, hardware failure, software bugs, communication link errors, environmental interference, malicious attack, and so on. Fault tolerance is the ability of a system to deliver a desired level of functionality in the presence of faults. Since the sensor nodes are prone to failure, fault tolerance should be seriously considered in the sensor network applications. Hence a fault tolerant mechanism is proposed to facilitate the transmission of the data packets to the base station so that the data packets reach the destination without any loss even in the presence of intermediate node failures along the multihop route to the base station.

# 2. RELATED WORK

All of the previous work done can be broadly classified into 2 categories viz structured approach and structure less approach. This section delves deeper into the various approaches proposed in these categories.

## 2.1 Structured Approach:

*Rumor routing* [2] routes the queries to nodes that observed a particular event rather than flooding, but maintaining agents and event-tables are sometimes infeasible. In [3], a protocol called *Low Energy Adaptive Clustering Hierarchy (LEACH)* forms cluster and randomly selects cluster heads and rotates this role to evenly distribute the load among nodes. In [4], *Power-Efficient Gathering in Sensor Information Systems* enhancement of *LEACH,* where nodes communicate only with their closest neighbors and once the turns of all the nodes are over, a new round will start. All the structured approaches results in fixed delay, which would be intolerable in large network deployments.

## 2.2 Structure-less Approach:

Studying the effect of the changing network topology is essential for analyzing the performance of structure less algorithms. [5] proposes a *Distributed Random Grouping(DRG)* algorithm that uses a probabilistic grouping to answer aggregate queries like computation of sum, average, maximum, minimum, etc. Through randomization, all values will progressively converge to the correct aggregate value (the average, maximum, minimum, etc.) in this method. The disadvantage with this approach is that it involves periodic and frequent transfer of message exchange between the nodes of a group. [6] proposed a novel data aggregation protocol for event based applications via 2 mechanisms namely *Data Aware Anycast (DAA)* at the MAC layer and *Randomized Waiting (RW)* at the application layer. DAA mechanism used RTS and CTS packet transmissions in order to determine whether the neighbor node has data. Since sensor nodes need to wait for data from other nodes, RW was proposed where each node chooses a random delay value within a maximum delay $\tau$. We call this as *Structure Free Data Aggregation (SFDA)* and compare our approach with SFDA to show the improvement in performance.

# 3. PROBLEM FORMULATION

Since we are primarily concerned with a structure less network, our aim here is to find a neighbor node within its communication range which is most probable to contain data and given that a node has sensed data, it needs to know how long it should wait for data from other nodes.

In addition to finding the next node to forward the data, another major concern of wireless sensor networks is to detect node failures so that the packets reach the base station without any loss. Our aim is to detect node failure with less number of control message transmissions and increase packet delivery ratio.

### 3.1 Neighbor Node Detection

In order to detect which node is most probable to contain data, we propose a prediction based approach as described in [1]. [6] proposed DAA method for the same problem of finding the neighbor node with data. The main problem with this approach is that it uses RTS and CTS packet transmissions for every data transmission. This can induce a serious load and can reduce the lifetime of sensor nodes considerably. Hence, we propose an approach which does not involve any kind of communication between the nodes. Each node analyses the collected upstream nodes data and calculates a probability value and using this probability value, it decides the node for which it has to forward data.

**Inverse-Square Law:** In physics, an Inverse-Square Law is any physical law stating that some physical quantity or strength is inversely proportional to the square of the distance from the source of that physical quantity. This law is very suitable for our environment since all electromagnetic waves has to obey the inverse square law. Since we do not know the position of the source, we take the position of node with highest sensed value as the position of source, as the node is closest to the source. The value sensed by the sensor node which is at a particular distance from the event is analogous to the intensity value at that distance from the source. Next, we need to determine the rate at which the sensed value changes with respect to the distance. If we obtain this, then we can predict the most probable sensed value for the downstream nodes. The Inverse-Square Law is,

$$I = \frac{k}{r^2} \tag{1}$$

Where I is Intensity, r is distance from source and k is a constant. Differentiating (1) gives the rate at which the sensed value changes with respect to distance.

$$\frac{dI}{dr} = -2K_m r^{-3} \tag{2}$$

The negative sign indicates that the intensity decreases with respect to distance. Here $K_m$ represents the mean value of K for all the upstream nodes.

$$K_m = \frac{\sum_{j=1}^{N} I_j r_j^2}{N} \tag{3}$$

From this, we can determine the rate at which the intensity has varied with respect to distance for all the upstream nodes. It is most likely to vary at the same rate for downstream nodes also.

**Survival Analysis (SA)** It is a branch of statistics which deals with death in biological organisms and failure in mechanical systems with respect to time t. We correlate this to our environment where we define the survival function S(r) as the probability that data is sensed by a node which is beyond distance r. It is defined as,

$$S(r) = P(R > r) \tag{4}$$

Similar to the lifetime distribution function of SA, we define the *Intensity distribution function*, which is a compliment of the survival function S(r), as

$$F(r) = P(R \le r) = 1 - S(r) \tag{5}$$

The above equation (5) indicates the probability that data is sensed by a sensor node that is located at distance r or below. *Event density function* which denotes the rate of data sensed with respect to distance is given by

$$f(r) = F'(r) = \frac{d}{dr}F(r) = 2K_m r^{-3} \tag{6}$$

**Kaplan Meier Estimator (KPE)** is a probabilistic measure of SA which denotes the probability that a living being survives up to some point of time. In other words, it denotes the probability that P(S≤ t). For our environment,
According to KPE,

$$S\wedge(r_i) = \prod_{x\le ri}(1-m\wedge(x)) \tag{7}$$

Here, Hazard function m^(x) is defined as the event rate at time *t* conditional on survival until time *t* or later (i.e., $T \ge t$). For our environment is given by,

$$m\wedge(x) = \frac{d(x)}{n(x)} \tag{8}$$

Where d(x) is number of nodes without event detection and n(x) is number of events under study. We now calculate the probability that data exists in the next downstream node. The probability that data don't exists in next downstream node is,

$$P(R \le r_0 + {}^{r}\!/_{R} > r_0) = \frac{\int_{r_0}^{r_0+r} f(r)dr}{S(r_0)} \tag{9}$$

The node with the least probability is chosen as the node that is most probable to have data. We then send data to that node and data aggregation is most probably achieved.

## 3.2 Delay Calculation

Once a node has collected data, it needs to know how long it should wait for data to come from its downstream nodes. In SFDA, they used a randomized waiting scheme wherein each node takes a random delay value within a certain maximum delay value. Deterministically assigning the waiting time to nodes such that nodes closer to the sink wait longer can avoid the problem but results in a fixed delay for all packets, which would be intolerable in large network deployments. Therefore, randomized waiting scheme is the optimal approach for assigning delay values to the sensor nodes. However, we propose Enhanced Random Delay (ERD), a subtle difference to that approach wherein, instead of making the maximum delay value fixed for the entire network, we make the maximum delay value dependent on the distance of the node from the sink. This provides an improvement in performance because of lesser probability for a node to choose a delay value that will make it wait longer than is necessary.

## 3.3 Node failure mechanism

Nodes in WSNs are prone to failure due to energy depletion, hardware failure, software bugs, communication link errors, environmental interference, malicious attack, and so on. Fault tolerance is the ability of a system to deliver a desired level of functionality in the presence of faults. Since the sensor nodes are prone to failure, fault tolerance should be seriously considered in the sensor network applications. Hence a fault tolerant mechanism (FTM) to facilitate the transmission of data packets to the base station is proposed so that the data packets reach the destination without any loss even in the presence of intermediate node failures along the multihop route to the base station.

Since sensor nodes are frequently subjected to unexpected failures, the wireless sensor network should be able to function and transmit messages within the network in spite of node failures. Under this scheme it is assumed that the failed nodes fall under either of the following two categories:

- The failed node is an intermediate node alone that just forwards the messages from the sender node to the destination which in most of the cases will be the base station
- The failed node is the one that senses the data along with performing aggregation and forwarding.

Under the second case, the failed sensor node cannot be used as an aggregation point and these nodes can be detected as failed only if this node falls under the route of any other sensor node that transmits. In case of node failures, we provide a solution to detect the failed node with comparatively less overhead of message transmissions and transmit the data packets successfully to the base station without affecting its transmission because of the node failure.

The proposed fault tolerant mechanism (FTM) is that whenever a node calculates a probability value to identify the most probable downstream node to which it can transfer the data, the sender node will check the presence of that node by sending a beacon signal and starting a timer. If the node doesn’t respond to the beacon signal within the stipulated time then, the node is considered to be failed and thus, the sender will select another node in an alternate route to transfer the data packets. While selecting the new destination node to which it intends to send the data, the sender will select the node based on the same criteria of calculating the probability based on equation (9) that the selected node will contain data. Additionally, when a node detects that a particular node has failed it sends message to its nearby nodes about the id of the failed node so that this failed node id can be removed from their routing table list. This transmission of messages to indicate the node failure is in a localized area and thus will not increase the overall overhead of the entire network to a great extent. By this localized transmission, it further avoids the transmission of beacon signals by other nodes to detect that failed node again and again every time they try to transmit to this failed node. Since a nearly equal alternate route is found, the packet delivery ratio increases in spite of node failures.

# 4. PERFORMANCE ANALYSIS

In order to evaluate the performance of our approach, we cannot use the number of aggregation points as the metric because the packets may be aggregated after travelling many hops. Expected number of transmissions and packet delivery ratio is the correct metric for evaluating the performance of our algorithm. To evaluate the performance of detecting the node failures, packet delivery ratio is to be used to compare if most of the packets that are sent by the event sensing nodes, reach the destination.

## 4.1 Expected Number of Transmissions

In this section, we will first calculate the probability for the packet to get aggregated at a node. After this, we will calculate the expected number of transmissions. In SFDA [5], they assume that the delay chosen by each node is distinct from each other. Using this assumption, they calculate the expected number of transmissions in the network. In a practical situation, each node is independent of each other i.e., each node chooses a random number that is independent of the random value chosen by another node. So, in the worst case, we compare the expected number of transmissions in the network of our approach with that of theirs.

Consider a chain topology of nodes from $v_0$ to $v_n$ where $v_0$ is the sink and all nodes have data to send. Let the number of nodes in the network be 7. Let *Y* be the discrete random variable representing the number of hops a packet has been forwarded before it is aggregated. As an example, for 7 nodes shown, the node $v_n$ can choose its delay so that its sending order (*l*) ranges from 1 to 6, and the node $v_{n-h}$ can take its order (*k*) from 2 to 7. The remaining (*h*-1) nodes in between take delay values in $l^{h-1}$ ways. The number of ways $N_0$ in which the nodes from $v_n$ to $v_{n-h}$ take their delay values is therefore,

$$N_0 = \sum_{k=2}^{n} \sum_{l=1}^{k-1} l^{h-1} \text{ , } if\ 0<h<n \tag{10}$$

From [1] it can be shown that the using equation (10) the expected number of transmissions is derived as,

$$E_0 = \sum_{h=1}^{n-1} h \text{ x} \left[ \frac{N_0}{n^h} \right] + \frac{\sum_{h=1}^{n} h^{n-1}}{n^{n-2}} \tag{11}$$

Based on equation (11) as derived in [1], the maximum allowed delay value for all nodes is chosen as τ. In our approach, we fix the maximum delay for each node based on the node's distance from the sink. The node will choose a random delay within that fixed maximum delay. Obviously, since we reduce the limit of delay value for each node, nodes farther from the sink will choose a lower delay than the nodes that are closer to the sink thereby attaining early aggregation. Therefore as described in [1], the expected number of transmissions for our approach is given by,

$$Ep = \sum_{i=2}^{R} \left[ \sum_{h=1}^{R-1} \frac{h(i-1)!}{(i+h)!} \text{x } N_p \right] + \sum_{h=1}^{R} \frac{h^2}{(h+1)!} \tag{12}$$

Where *Np* denotes the number of ways in which the nodes from $v_n$ to $v_{n-h}$ takes its order whose value is shown in [1]. Thus for increasing network size the expected number of transmissions in the ERD decreases drastically compared to the SFDA approach.

### 4.2 Packet delivery ratio (PDR)

Packet delivery ratio can be defined as the ratio of the number of sensed data packets received by the base station to the number of data packets sent by the event sensing nodes.

$$PDR = \frac{P_{recv}}{P_{sent}} \tag{13}$$

PDR is in the range of 0 to 1. The higher the value of PDR indicates the packets are delivered successfully to the base station through different routes that doesn't include the failed nodes. Therefore, this parameter is compared with the number of nodes that are failed in the network.

## 5. PERFORMANCE EVALUATION

The OMNeT++ 4.0 simulator is used along with MiXiM simulation framework to provide mobility among the events. The nodes are arranged in a grid topology with inter-node separation of 30 m between the sensor nodes. Intel-Lab data [7] is also used to calculate the number of transmissions for this network using our approach. The simulation scenario used is same as described in [1].

### 5.1 Simulation Scenario

**Table 1.** Default Parameters Used in the Simulation

| Parameters | Values |
|---|---|
| Network Topology | 300 m x 300 m |
| Data Rate | 38.4 Kbps |
| Communication Range | 55 m |
| Mobility Model | ConstSpeedMobility |
| Packet Size | 50 bytes |
| Sensing Interval | 10 s |
| Event Size | 50 m to 200 m |
| Internode Separation | 30 m |
| Event Moving Speed | 10 m/s |
| Maximum Delay | 0.8 s to 4 s |

The Packet Aggregation Ratio (PAR) is used as metric to compare different protocols. PAR determines how effective a protocol is in aggregating packets and is (Number of nodes in which a packet gets aggregated/Number of nodes through which packet is transmitted to sink). PAR will be in the range 0 to 1.Maximum the value of ratio determines the packet is effectively aggregated in the route. Table.1 shows the various in the default parameters used in the simulation.

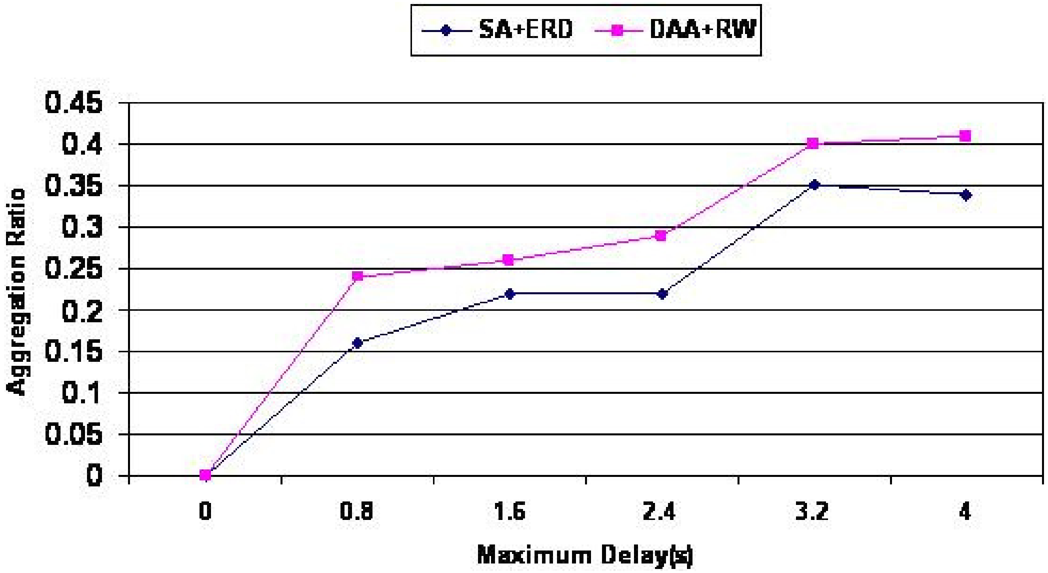

Figure1. Aggregation ratio vs Different Maximum Waiting time

The Packet Delivery Ratio (PDR) is used as a metric to compare the proposed node failure avoidance mechanism with the network without using this node failure mechanism. The PDR value is calculated for different number of nodes that have failed in the network. The simulation results show that the packet delivery ratio is high when the node failure avoidance mechanism is used.

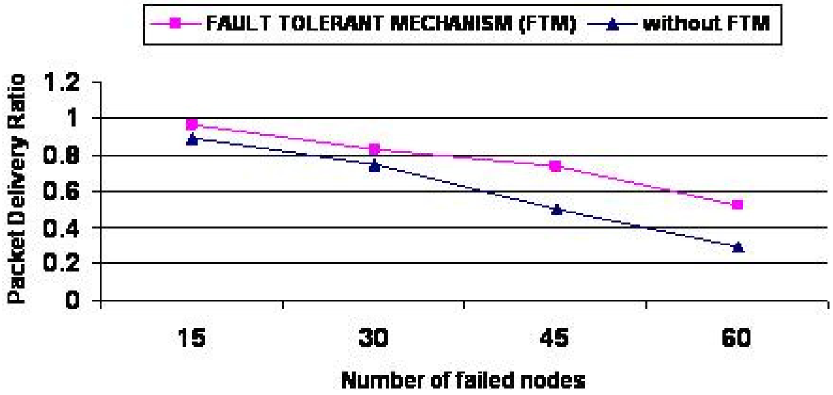

Figure 2. Packet delivery ratio vs number of failed nodes

Maximum Delay (MD) is used as a varying parameter and the corresponding number of transmissions for the Intel-Lab network is plotted as shown in Figure 3. For increase in MD, the total number of transmissions in the network should decrease. From the graph Figure 3, it is shown that, the proposed approach performs better than SFDA approach. On average, the number of transmissions decreases by 9.1 percent.

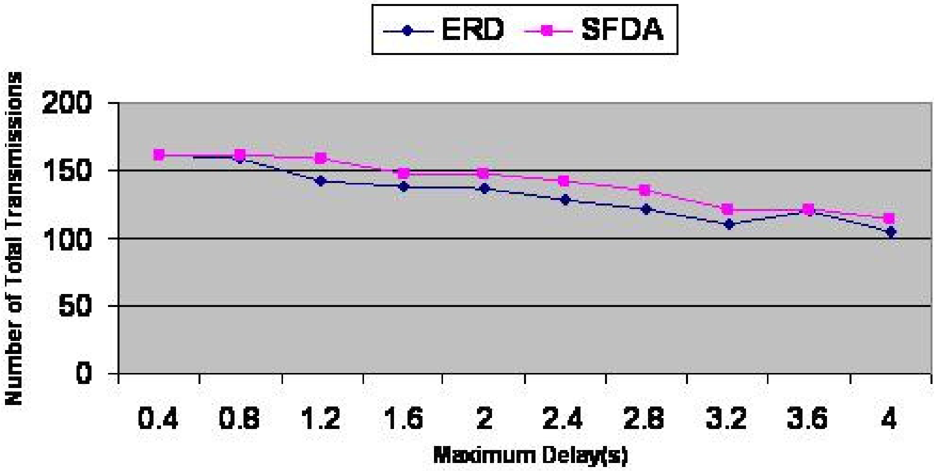

Figure 3. Number of Transmissions for Varying Maximum Delays

## 6. CONCLUSION

In this paper, we proposed a cost forming any explicit structure in the network. Instead of spending more energy in communication, SA, neighbor node that is more probable to contain data. Using the ERD method, the unwanted message transmissions are further reduced by finding the random delay within a fixed delay. From extensive simulation exhibits almost equal performance to SFDA by minimizing energy consumption and unnecessary message transmissions. By the node failure avoidance mechanism we have proposed, it results in high packet delivery ratio in spite of node failures.

## REFERENCES

[1] Khaja Muhaiyadeen A, Hari Narayanan R, Shelton Paul Infant C, Rajesh G, Inverse Square Law Based Solution for Data Aggregation Routing using Survival Analysis in Wireless Sensor

Networks, International Conference on Networks and Communication (Netcom), Bangalore, India,2010

[2] Braginsky, D., Estrin, D.: Rumor Routing Algorithm for Sensor Networks. In: Proceedings of the First Workshop on Sensor Networks and Applications (WSNA), Atlanta, GA, October 2002.

[3] Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy Efficient Communication Protocol for Wireless Microsensor Networks. In: Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS'00), January 2000.

[4] Lindsey, S., Raghavendra, C.: PEGASIS: Power-Efficient Gathering in Sensor Information Systems, IEEE Aerospace Conference Proceedings, 2002, Vol.3, 9-16 pp.1125-1130.

[5] Jen-Yeu Chen., Jianghai Hu.: Analysis of Distributed Random Grouping for Aggregate Computation on Wireless Sensor Networks with Randomly Changing Graphs. In: IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 19, NO. 8, AUGUST 2008.

[6] Kai-Wei Fan., Sha Liu., Prasun Sinha.: Structure-Free Data Aggregation in Sensor Networks, In: IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 6, NO. 8, AUGUST 2007.

[7] Intel Berkeley Research Lab Sensor data, http://db.csail.mit.edu/labdata/labdata.html.

[8] F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, ''Wireless sensor networks: A survey,'' Computer Networks Journal, volume 4, no. 12, Mar. 2002, pp. 393—422.

[9] Giuseppe Anastasi a, Marco Conti, Mario Di Francesco, Andrea Passarella, Energy conservation in wireless sensor networks: A survey, Adhoc networks journal Elsevier, volume 7, 2009 , pp. 537–568.

[10] Kemal Akkaya *, Mohamed Younis, A survey on routing protocols for wireless sensor networks, Ad Hoc Networks journal, Elsevier, volume 3, 2005, pp. 325–349.

[11] C.R. Lin and M. Gerla, “Adaptive Clustering for Mobile Wireless Networks”, IEEE Journal on Selected Areas In Communications, volume 7, September 1997, pp. 1265-1275.

[12] S.Basagni, “Distributed Clustering for Ad Hoc Networks”, Proceedings of International Symposium on Parallel Architectures, Algorithms and Networks, pp. 310-315, June 1999.

[13] D. Chen, and P. Varshney, “QoS Support in Wireless Sensor Networks: A Survey,” International Conference on Wireless Networks, Las Vegas, Nevada, USA, June 21-24, 2004.

[14] G. Bravos, A. Kanatas, “Energy Consumption and Trade-offs on Wireless Sensor Networks,” IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, Vol. 2, pp. 1279- 1283, September 2005.

[15] Ameer Ahmed Abbasi a,*, Mohamed Younis, A survey on clustering algorithms for wireless sensor networks, Computer Communications , Science Direct, Volume 30 2007 , pp.2826–2841.