Data Aggregation in Cluster-based Wireless Sensor Networks

Kriti Patidar[1], Suyog Malviya[2], Chanchal Soni[3], Deepak Mittal[4]
[1, 2, 3, 4] Department of Electronics and Communication Engineering
[1, 2, 3, 4] SVVV Indore, INDIA

Abstract*-The rapid advancement of hardware technology has enabled the development of small, powerful, and inexpensive sensor nodes, which are capable of sensing, computation and wireless communication. This revolutionizes the deployment of wireless sensor network for monitoring some area and collecting regarding information. However, limited energy constraint presents a major challenge such vision to become reality. We consider energy constrained wireless sensor network deployed over a region. The main task of such a network is to gather information from node and transmit it to base station for further processing. So the aim of any data forwarding protocol is to conserve energy to maximize the network lifetime. Sensor nodes are capable of performing in-network aggregation of data coming from more than one source. In this thesis we have concentrated on energy consumption issue and aim to develop an energy efficient data aggregation protocol. To provide energy efficiency we have considered a cluster-based wireless sensor network. Our protocol executes on each cluster independently and provides an energy efficient data aggregation in a cluster and hence maximize network lifetime for whole network.*

Keywords-wireless sensor networks, energy conservation, energy consumption, data aggregation, network lifetime

I. INTRODUCTION

A wireless sensor network is a wireless network consisting of tiny devices which monitor physical or environmental conditions such as temperature, pressure, motion or pollutants etc. at different regions. The tiny device, known as sensor node, consists of a radio transceiver, microcontroller, power supply, and the actual sensor. Initially sensor network were used for military applications but now they are widely used for civilian application area including environment and habitat monitoring, healthcare application and so on. Normally sensor nodes are spatially distributed throughout the region which has to be monitored; they self-organize in to a network through wireless communication, and collaborate with each other to accomplish the common task. With the going time, sensor nodes are becoming smaller, cheaper, and more powerful which enable us to deploy a large-scale sensor network. Basic features of sensor networks are self-organizing capabilities, dynamic network topology, limited power, node failures & mobility of nodes, short-range broadcast communication and multi-hop routing, and large scale of deployment [12]. The strength of wireless sensor network lies in their flexibility and scalability. The capability of self-organize and wireless communication made them to be deployed in an ad-hoc fashion in remote or hazardous location without the need of any existing infrastructure. Through multi-hop communication a sensor node can communicate a far away node in the network. This allows the addition of sensor nodes in the network to expand the monitored area and hence proves its scalability & flexibility property.

A. Clustering in WSN

It is widely accepted that the energy consumed in one bit of data transfer can be used to perform a large number of arithmetic operations in the sensor processor [13]. Moreover in a densely deployed sensor network the physical environment would produce very similar data in close-by sensor nodes and transmitting such data is more or less redundant. Therefore, all these facts encourage using some kind of grouping of nodes such that data from sensor nodes of a group can be combined or compressed together in an intelligent way and transmit only compact data. This can not only reduce the global data to be transmitted and localized most traffic to within each individual group, but reduces the traffic and hence contention in a wireless sensor network. This process of grouping of sensor nodes in a densely deployed large-scale sensor network is known as clustering. The intelligent way to combined and compress the data belonging to a single cluster is known as data aggregation. There are some issues involved with the process of clustering in a wireless sensor network. First issue is, how many clusters should be formed that could optimize some performance parameter. Second could be how many nodes should be taken in to a single cluster. Third important issue is the selection procedure of cluster- head in a cluster. Another issue that has been focused in many research papers is to introduce heterogeneity in the network. It means that user can put some more powerful nodes, in terms of energy, in the network which can act as a cluster-head and other simple node work as cluster-member only. Considering the above issues, many protocols have been proposed which deals with each individual issue.

II. RESEARCH BASIS AND WORK PERFORMED

Data aggregation protocols aims at eliminating redundant data transmission and thus improve the lifetime of energy constrained wireless sensor network. In wireless sensor network, data transmission took place in multi-hop fashion where each node forwards its data to the neighbour node which is nearer to sink. That neighbour node performs aggregation function and again forwards it on. But performing data forwarding and aggregation in this fashion from various sources to sink causes significant energy waste as each node in the network is involved in operation. Since closely placed nodes may sense same data, above approach cannot be considered as energy efficient. An improvement over the above approach would be clustering where each node sends data to cluster-head (CH) and then cluster-head perform aggregation on the received raw data and then send it to sink. Performing aggregation function over cluster-head still causes significant energy wastage. In case of homogeneous sensor network cluster- head will soon die out and again re-clustering has to be done which again cause energy consumption.

A. Tree Based Approach

The simplest way to aggregate data is to organize the nodes in a hierarchical manner and then select some nodes as the aggregation point or aggregators. The tree-based approach perform aggregation by constructing an aggregation tree, which could be a minimum spanning tree, rooted at sink and source nodes are considered as leaves. Each node has a parent node to forward its data. Flow of data starts from leaves nodes up to the sink and therein the aggregation done by parent nodes. The way this approach operates has some drawbacks. As we know like any wireless network the wireless sensor networks are also not free from failures. In case of packet loss at any level of tree, the data will be lost not only for a single level but for whole related sub-tree as well. In spite of high cost for maintaining tree structure in dynamic networks and scarce robustness of the system, this approach is very much suitable for designing optimal aggregation technique and energy-efficient techniques. S. Madden et al. in [14] proposed a data-centric protocol which is based on aggregation tress, known as Tiny Aggregation (TAG) approach [14]. TAG works in two phases: distribution phase and collection phase. In distribution phase, TAG organizes nodes in to a routing tree rooted at sink. The tree formation starts with broadcasting a message from sink specify level or distance from root. When a node receive this message it sets its own level to be the level of message plus one and elect parent as node from which it receives the message. After that, node rebroadcast this message with its own level. This process continues until all nodes elect their parent. After tree formation, sink send queries along structure to all nodes in the network. TAG uses database query language (SQL) for selection and aggregation functions. In collection phase, data is forwarded and aggregated from leaves nodes to root. A parent node has to wait for data from all its child node before it can send its aggregate up the tree. Apart from the simple aggregation function provided by SQL (eg: COUNT, MIN, MAX, SUM, and AVG), TAG also partitions aggregates according to the duplicate sensitivity, exemplary and summary, and monotonic properties. Though TAG periodically refresh tree structure of network but as most of the tree-based schemes are inefficient for dynamic network, so TAG may be. C. Intanagonwiwat et al. in [3] proposed a reactive data-centric protocol for applications where sink ask some specific information by flooding, known as directed diffusion paradigm. The main idea behind directed diffusion paradigm is to combine data coming from different source and en-route them by eliminating redundancy, minimizing the number of data transmission; thus maximizing network lifetime. Directed diffusion consists of several elements: interests, data messages, gradients, and reinforcements.

Figure 2.1 Simplified schematic for directed diffusion. (a) Interest propagation. (b) Initial gradients setup. (c) Data delivery along reinforced path [3].

The base station (BS) requests data by broadcasting an interest message which contains a description of a sensing task. This interest message propagates through the network hop-by-hop and each node also broadcast interest message to its neighbour. As interest message propagates throughout the network, gradients are setup by every node within the network. The gradient direction is set toward the neighbouring node from which the interest is received. This process continues until gradients are setup from source node to base station. Loops are not checked at this stage but removed at later stage. After this path of information flow are formed and then best path are reinforced to prevent further flooding according to a local rule. Data aggregation took place on the way of different paths from different sources to base station or sink. The base station periodically refresh & resend the interest message as soon as it start to receives data from sources to provide reliability. The problem with directed diffusion is that it may not be applied to applications (e.g. environmental monitoring) that require continuous data delivery to base station. This is because query driven on demand data model may not help in this regard. Also matching data to queries might require some extra overhead at the sensor nodes. Mobility of sink nodes can also degrade the performance as path from sources to sinks cannot be updated until next interest message is flooded throughout the network. To cope up with above issue if introduce frequent flooding then also too much overhead of bandwidth and battery power will be introduced. Furthermore, exploratory data follow all possible paths in the network following gradients which lead to unnecessary communications overhead.

B. Multi Path Approach

One of the main drawbacks of tree-based approach is the scarce robustness of the system. To overcome this drawback, a new approach was proposed by many researchers. Instead of sending partially aggregated data to a single parent node in aggregation tree, a node could send data over multiple paths. The idea behind is that each node can send the data to its possibly multiple neighbours by exploiting the wireless medium characteristic. Hence data will flow from sources to sink along multiple paths and aggregation can be performed by each intermediate node. Clearly schemes using this approach will make the system robust but with some extra overhead. One of theaggregation structures that fit well with this approach is ring topology, where network is divided in to concentric circles with defining levels according to the hop distance from sink. S. Nath et al. in [15] presented a data aggregation technique using multi-path approach, known as synopsis diffusion. Synopsis diffusion works in two phases: distribution of queries and data retrieval phase.

During distribution of queries phase, a node sends a query in the network. The network nodes then forma set of rings around the querying node. The node which is i hop away from querying node is considered is ring Ri. In the second phase, aggregation starts from outermost ring and propagate level by level towards the sink. Here a source node can have multiple paths towards sink.

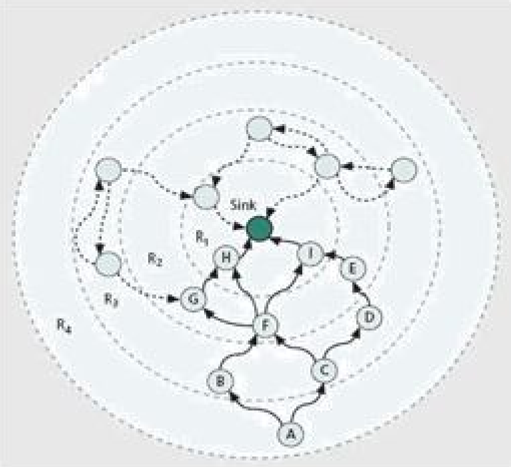

Figure 2.2 Examples of aggregation paths over a ring structure [7].

L. Gatani et al. in [5] describe a new strategy for data gathering in wireless sensor network that consider both issues: energy efficiency and robustness. Authors first say that single path to connect each node to the base station is simple & energy-saving approach but expose a high risk of disconnection due to node/link failures. But multi-path approach would require more nodes to participate with consequent waste of energy. Authors present a clever use of multi-path only when there is loss of packet which is implemented by smart caching of data at sensor nodes. Authors also argue that in many practical situation data may be gathered only from a particular region, so they use a different approach that relies on a spanning tree and provides alternative paths only when a malfunctioning is detected. Algorithm adopts a tree-based approach for forwarding packets through the network. In the ideal situation when no failures occur, this is certainly the best choice, as the minimum number of nodes is engaged in the transmission phase. In the presence of link or node failures, the algorithm will discover alternative paths, so as ensure the delivery of as many packets as possible within the time constraints. The problem with this approach is that it may cause the arising of hot spots and nodes along preferred paths will consume their energy resources quickly, possibly causing disconnection in the network.

C. Cluster-Based Approach

We talked about hierarchical organization of the network in tree-based approach. Another scheme to organize the network in hierarchical manner is cluster-based approach. In cluster-based approach, whole network is divided in to several clusters. Each cluster has a cluster-head which is selected among cluster members. Cluster-heads do the role of aggregator which aggregate data received from cluster members locally and then transmit the result to sink. The advantages and disadvantages of the cluster-based approaches is very much similar to tree-based approaches. K. Dasgupta et al. in [16] proposed a maximum lifetime data aggregation (MLDA) algorithm which finds data gathering schedule provided location of sensors and base-station, data packet size, and energy of each sensor. A data gathering schedule specifies how data packet are collected from sensors and transmitted to base station for each round. A schedule can be thought of as a collection of aggregation trees. In [6], they proposed heuristic-greedy clustering-based MLDA based on MLDA algorithm. In this they partitioned the network in to cluster and referred each cluster as super-sensor. They then compute maximum lifetime schedule for the super-sensors and then use this schedule to construct aggregation trees for the sensors. W. Choi et al. in [1] present a two-phase clustering (TPC) scheme. Phase I of this scheme creates clusters with a cluster-head and each node within that cluster form a direct link with cluster-head. Phase I of this scheme is similar to various scheme used for clustering but differ in one way that the cluster-head rotation is localized and is done based on the remaining energy level of the sensor nodes which minimize time variance of sensors and this lead to energy saving from unnecessary cluster-head rotation. In phase II, each node within the cluster searches for a neighbour closer than cluster-head which is called data relay point and setup up a data relay link. Now the sensor nodes within a cluster either use direct link or data relay link to send their data to cluster head which is an energy efficient scheme. The data relay point aggregates data at forwarding time to another data relay point or cluster-head. In case of high network density, TPC phase II will setup unnecessary data relay link between neighbours as closely deployed sensor will sense same data and this lead to a waste of energy.

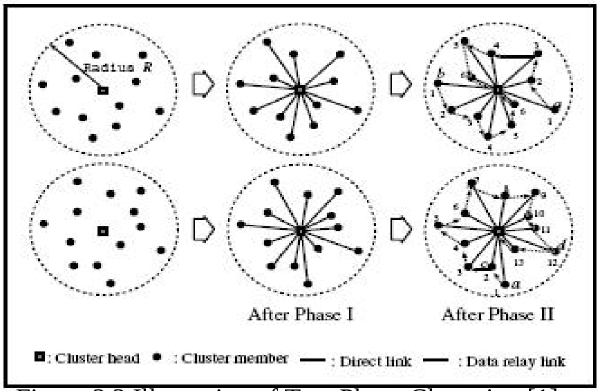

Figure 2.3 Illustration of Two Phase Clustering [1].

H. Cam et al. in [4] present energy efficient and secure pattern based data aggregation protocol which is designed for clustered environment. In conventional method data is aggregated at cluster-head and cluster-head eliminate redundancy by checking the content of data. This protocol says that instead of sending raw data to cluster-head, the cluster members send corresponding pattern codes to cluster-head for data aggregation. If multiple nodes send the same pattern code then only one of them is finally selected for sending actual data to cluster-head. For pattern matching, authors present pattern comparison algorithm.

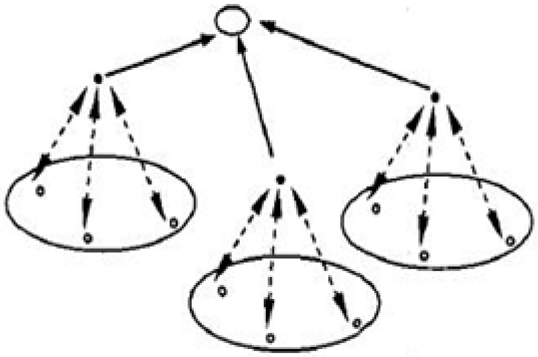

Figure 2.4 Data transmission using ESPDA [4].

III. PROPOSED WORK

A. System & Energy Model

Consider a homogeneous network of n sensor nodes and a base station or sink node distributed over a region. The location of the sensors and the base station are fixed and known priori. Each sensor produces some information as it monitors its vicinity. We assume that the whole network is divided in to several clusters; each cluster has a cluster-head (CH). The clustering and the selection of cluster-head (CH) can be done by using any existing protocol like LEACH, such that cluster-head (CH) is maximum k-hop away from any node in cluster. We also assume that after the formation of cluster the transmission power of all nodes is adjusted in such a way that they can perform single hop broadcast. Single hop broadcast refers to the operation of sending a packet to all single-hop neighbours [8]. Our energy model for the sensors is based on the first order radio model described in [17]. A sensor consumes Eelec = 50nJ/bit to run the transmitter or receiver circuitry and Eamp = 100pJ/bit/m2 for the transmitter amplifier. Thus, the energy consumed by a sensor i in receiving a l-bit data packet is given by,

$E_{Rx_i} = E_{elec} . l$ (1)

while the energy consumed in transmitting a data packet to sensor j is given by,

$E_{Tx_{i,j}} = E_{elec} . l + E_{amp} . d_{i,j}^2 . l$ (2)
where $d_{i,j}$ is the distance between nodes i and j.

B. Protocol Description

In a cluster-based wireless sensor network, our algorithm is designed to provide energy-aware in-network data aggregation in a cluster. Each cluster uses this algorithm independently. In a cluster, the nodes can be categorized as: one cluster-head (CH) and other cluster member node.

1) Function of cluster-head (CH):

Receive a query from base station.

Cluster-head (CH) sends configuration packets to all single-hop neighbours.

Receive data packets from all single hop neighbours. Finally aggregate the data packets received and route it to base station.

2) Function of cluster member:

Receive configuration packets from neighbour nodes.

Update and forward configuration packets to all single-hop neighbours.

Receive data packets from neighbour nodes.

Aggregate all data packets by applying redundancy factor and send it to selected parent node.

The algorithm works in two phases: Configuration packet flow and Data packet flow that are described below. Initial cluster position Flow of configuration packets Flow of data packets through selected parent.

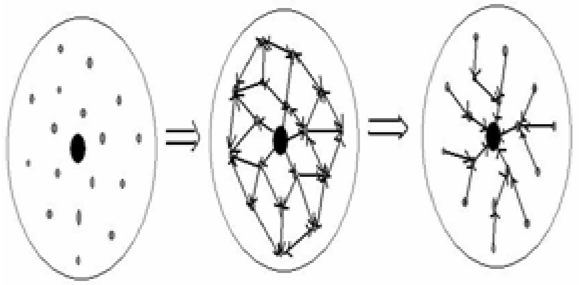

Figure 3.1 A typical scenario of data aggregation in a cluster.

C. Data Packet Flow

When all nodes receives configuration packets, each node now select the parent to which it can forward the data packet. The parent selection procedure is shown in algorithm 3.2. Each node looks in to the list of all its possible parents. The neighbour node which has least hop distance, ie closest to cluster-head, is selected as parent by a node. In case when two neighbour nodes have the least but equal hop distance, the node checks the residual energy of two neighbour nodes. The neighbour node that has greater residual energy is now selected as parent. In both the cases, node also calculate the difference of residual energy of two neighbour nodes, which have least hop distance, and checks whether this difference is less than the threshold or not. If it is then the node selects the parent as usual. But if it is not then the node selects other neighbour node as its parent.

```
Define:
    Er[i]: residual energy of node i
    dh[i]: distance from CH in terms of hop count of node i
    Edij: difference of residual energy of two nodes i & j
    te: threshold of residual energy difference of two nodes i & j for balancing load
    nid: id of a node
    S: set of configuration packets received by node i
    ParentSelection(nid)
        select two nodes j & k such that dh[j] & dh[k] is minimal in S
        if (dh[j] < dh[k] AND |Edjk| < te)
        then return nid of node j
        else if (dh[j]==dh[k] AND |Edjk| < te)
        then return nid of node with max(Er[j] , Er[k])
        endif
        else return nid of node k
        endif
```

Algorithm 3.2 Parent selection procedure.

This allows a node that has more available resources to be selected as a parent node. This also balances the consumption of energy of nodes in the cluster and leads to die out of nodes nearly at same time. After selecting the parent node, each node now forwards its data to its parent. When a parent node receives multiple data packets from its neighbour nodes, it performs aggregation operation by eliminating redundancy in the data. Each parent node checks the equation below:

$$| VNi – VNj | < K \quad (3)$$

where, VNi data value of node i VNj data value of node j

K redundancy factor

If this equation satisfies, the parent node perform aggregation by applying any aggregation functions like MIN, MAX, and AVG on the values of data packet and send only one packet while discarding other packets. But if this equation do not satisfies, the parent performs aggregation by simply concatenating two data packet in to one keeping value of both packets intact. The selection of value for redundancy factor (K) has a trade-off between precision and energy consumption. If the application wants more precision, it should select a low value for redundancy factor otherwise a high value. Selecting high value for K means sending only one value thus less number of bits needs to be transmitted and hence low energy consumption.

IV. CONCLUSION

Wireless sensor networks are energy constrained network. Since most of the energy consumed for transmitting and receiving data, the process of data aggregation becomes an important issue and optimization is needed. Efficient data aggregation protocols not only provide energy conservation but also remove redundancy in the data and hence provide useful data only. There exist several protocols for data aggregation which uses different approaches to provide energy efficiency. In cluster-based approaches, nodes send their data directly to cluster-head and cluster-head then aggregate and forward the data towards sink. We exploited this approach and proposed a new protocol called Energy-aware Balanced In-Network Aggregation (E-BINA). E-BINA uses the advantageous features of cluster-based and tree-based approaches. E-BINA requires a wireless sensor network which is divided in to several clusters, each having a cluster-head. Each cluster then uses E-BINA independently and avoids aggregation only at cluster-head by constructing an aggregation tree rooted at cluster-head. During the construction of aggregation tree, each cluster member node chooses its parent node among its neighbours based on the information of residual energy and hop distance from cluster-head. After the construction of aggregation tree, when a parent node receives data from its different child neighbour nodes, it eliminates the redundancy in the data received from different nodes and then forward. The difference between E-BINA and other cluster-based approach lie in the reduction of transmission power of node as in E-BINA a node send data to its neighbour node instead of sending to cluster-head. This is the main performance improvement factor of our protocol. The simulation result shows that when the data from source node is send to cluster-head through neighbours nodes in a multi-hop fashion by reducing transmission and receiving power, the energy consumption is low as compared to that of sending data directly to cluster-head. The simulation result shows that when the data from source node is send to cluster-head through neighbours nodes in a multi-hop fashion by reducing transmission and receiving power, the energy consumption is low as compared to that of sending data directly to cluster-head.

REFERENCES

[1] W. Choi, P. Shah, and S. K. Das, “A Framework for Energy-Saving Data Gathering Using Two-Phase Clustering in Wireless Sensor Networks”, in Proceedings of the International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous), Boston, 2004, pp. 203-212.

[2] M. Lee, and S. Lee, “Data Dissemination for Wireless Sensor Networks”, in Proceedings of the 10th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC’07).

[3] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, “Directed Diffusion for Wireless Sensor Networking”, IEEE/ACM