INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

CORREP: AN EFFICIENT CORRELATION AND REPLICATION BASED DISTRIBUTED DATA STORAGE PROTOCOL WITH MOBILE SINK

Neenu M.Nair[1], J. Sebastian Terence[2]

[1] PG scholar, Dept. Of CSE, Karunya University, neenumnair@gmail.com
[2] Assistant Professor, Dept. Of CSE, Karunya University, jsebinfo@gmail.com

Abstract

Wireless Sensor Networks (WSNs) key role is to perform sensing, collecting and storing measures from the surrounding environment. Distributed data storage proved as the best method for data storage through previous studies since it help to store the data among storage nodes in a distributed manner. This paper presents CORREP, an efficient correlation and replication based data storage protocol for large scale wireless sensor networks with mobile sink. Contrarily to related protocol CORREP effectively manages the data replication among the storage nodes in the network. The proposed protocol considers intelligent correlation mechanisms for summarization of the data packets without affecting the data quality and also it can overcome the node failures that may occur in related protocols. CORREP can guarantee improved network lifetime and energy efficient data storage schemes when compared with existing protocols for distributed data storage. The proposed efficient correlation and replication schemes can maintain the data collection efficiency as same as in the related protocols.

Keywords: Distributed data storage, Data correlation, Data replication, Network lifetime, Energy efficiency.

1. Introduction

A wireless sensor network (WSN) is a distributed network of autonomous devices called sensors which are spatially arranged in order to monitor and collect data or events from a specified geographical area. The main application areas of WSN include industrial monitoring, healthcare monitoring and area monitoring. The gathered events by sensors latter collected by sink node which is responsible for retrieval of data from sensor networks. Since sink node are responsible for storing and processing sensed data it has more power, storage and computational capabilities compared with other sensors. Sink can be either static or mobile (Viana et al., 2010), (Neenu et al., 2013), (Vecchio et al., 2010) and (Piotroski et al., 2009). In most of the monitoring based WSN application areas data storage schemes play a vital role in order to achieve acceptable data gathering

efficiency, Data dissemination efficiency, data availability, network lifetime and system robustness. Thus well organized data storage schemes are inevitable.

There are several schemes were introduced for data storage in wireless sensor networks such as local, centralized and distributed data storage (Wei et al., 2010), (Wen-Hwa et al., 2009) and (Neenu et al., 2013) which is shown in fig.1. In local storage, events are stored at the local memory of each sensor node alternative to local storage centralized storage responsible for store data in a centralized sink ,since sink node are external to the network it can be consider as a external storage. Finally distributed data storage schemes are proposed which can offer efficient data storage than all other existing techniques.

Distributed data storage (Gonizzi et al., 2013) allows store the sensed data either locally or to some other storage node (a node having the ability store data sensed by itself and from other sensors) in the geographical area. Which can be further classified as data centric storage (DCS) or fully distributed data storage (Neenu et al., 2013). In DCS (Wen-Hwa et al., 2009) some designated nodes act as a storage node, data storage normally carried out based on the event type with the help of GHT (geographic hash table). Fully distributed schemes consider all the nodes in the geographical area as storage nodes. Fully distributed schemes can perform well compared with all other existing data storage techniques.

This work introduce an efficient correlation and replication based fully distributed data storage scheme which can overcome the disadvantages of existing protocols such as RaWMS (Bar-Yossef et al., 2008), SUPPLE (Viana et al., 2010), DEEP (Vecchio et al., 2010) and ProFlex (Maia et al., 2013). The proposed technique is based on fully distributed data storage with mobile sink, so that this protocol consider all sensors are responsible for sensing and storing. The key research challenge of this paper is how to achieve good level of energy efficiency and network lifetime.

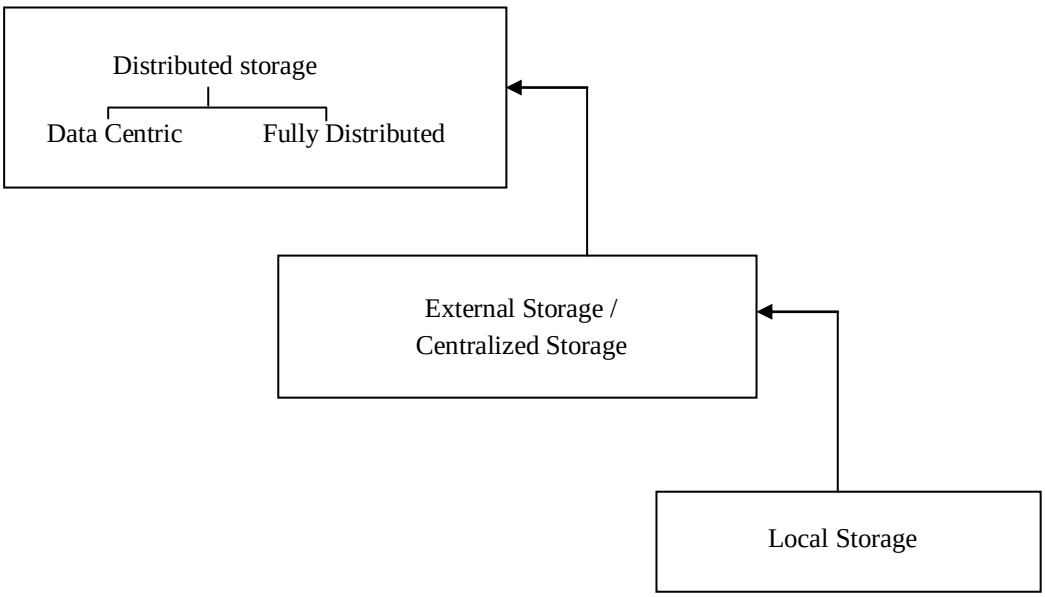

Figure 1: Data Storage techniques in WSN

In the remainder of this paper: Section 2 presents the Problem statement; Section 3 describes the Related works done about the data storage in WSN; Section 4 discusses Distributed data storage protocol ProFlex (Maia et al., 2013) and its disadvantages and Section 5 proposes CORREP approach and its formal analysis that shows how to overcome the disadvantages of ProFlex; Section 6 concludes the paper.

2. Problem Statement

Data storage play a vital role in most of the WSN applications since it is responsible for gain good data gathering efficiency, data availability, network lifetime etc. Since sensor nodes are prone to failure local storage methods may affected by node failure in the case of buffer overflows due to hot spot storage problem. Also it adopts blind flooding since the sink node does not know about where the data of interest is residing, this leads to high communication overhead. The concept of external storage has to face the problems such as sink node bottle neck due to high event frequencies. The failure of sink node affect the connectivity of network thus network life time reduces drastically. Due to the difficulties induced by both local and external storage WSN applications move on to distributed storage since it is robust against node failure. Data centric approach of distributed storage can avoid blind flooding since sink node is aware of where the data is stored. Challenges of Data Centric Storage such as hot spot storage problem, data availability, energy efficiency etc. addressed by fully distributed data storage schemes.Several data storage protocols (Viana et al., 2010) and (Maia et al., 2013) uses fully distributed concept. Such schemes also uses mobile sink so that energy hole problem can be avoided. These protocols combine the concept of distributed storage with replication inorder to improve the data gathering efficiency and data availability (Shen et al., 2013). Hot spot storage problem overcome by the use of uniform selection criterion for the selection of storage nodes. Since the implementation of replication may produce high communication overhead efficient data correlation is necessary so that it is possible to reduce the total number of transmitted messages. The existing protocol doesn't make effective number of replicas and efficient correlation mechanism thus network lifetime and energy efficiency will become an issue.

In our work a fully distributed data storage scheme for data storage is developed which introduce a correlation mechanism so that network lifetime, energy efficiency, message overhead and data gathering efficiency can be improved to a better level.

3. Related Work

This section presents some of the proposals for data storage in WSNs.

Fully distributed storage and Data Centric Storage (Neenu et al., 2013) are the recent advances in data storage in WSN; both are a kind of distributed storage. There are several reviews presented to study about data storage. This is shown in Table 1.

Main Classification	Sub classification	Title	Data Availability	Security	Energy efficient	Network lifetime
Fully Distributed Data Storage [FDDS]	Topology Based	ProFlex	Y	N	Y	Y
		SUPPLE	Y	N	N	Y
	Security Based	C&R- DS	N	Y	N	N
		S&D –DS	N	Y	Y	N
	Load-Balancing Based	C-STORAGE	N	N	Y	Y
		DS for IOT	Y	N	N	Y
	Reliability Based	TinyDSM	Y	N	N	N
		DS for CDA	Y	N	Y	Y
Data centric Storage [DCS]	Topology Based	SDS	N	N	Y	Y
		KDDCS	N	N	Y	N
	Security Based	Pdcs	N	Y	Y	N
		DS-FBA	N	Y	Y	Y
	Load-Balancing Based	DLB	N	N	Y	Y
		ASR	Y	N	Y	Y
	Reliability Based	ADCS	Y	N	N	Y
		D-DCS	Y	N	Y	N

Table1: Comparison Of distributed data storage schemes in WSNs.

Supple (Viana et al., 2010), ProFlex (Maia et al., 2013) are examples of topology based fully distributed data storage. Supple propose a flexible probabilistic data dissemination protocol for wireless sensor networks which can use either static or mobile sinks. Deep (Vecchio et al., 2010) is density-based proactive data dissemination protocol for wireless sensor networks with uncontrolled sink mobility which is proposed before Supple. The supple protocol (Viana et al., 2010) uses a tree topology for data management. Data management includes both data dissemination and storage. Supple protocol completed its task through its three phases called tree construction, weight distribution and data replication. Data replication phase is responsible for the data storage and creation of number of replicas for each data packet. However supple does not consider the problem of energy consumption and traffic overload at the node closer to the root of the tree. Also message overhead may increase in case if event frequency is high since supple forced to create replicas of each created

data packet during data replication phase. Also number of replicas produced is insufficient to achieve better data availability.

ProFlex (Maia et al., 2013) is a distributed data storage protocol for heterogeneous wireless sensor networks proposed to overcome the difficulties faced by supple protocol. ProFlex introduce multiple replication structures by the initialization of multiple trees to increase the data gathering efficiency and to reduce the traffic overload. Data correlation mechanism proposed by proFlex help to reduce the message overhead.

Clearly, each existing protocols shares its strengths and weakness, thus our solution is to propose an enhancement of ProFlex (Maia et al., 2013), and by taking the best features from it, then combine it with the effective correlation mechanism and also find a solution to increase the energy efficiency further more.

4. ProFlex: Distributed Data Storage Protocol

G.Maia et.al (Maia et al., 2013) proposes a distributed data storage protocol for heterogeneous wireless sensor network. It uses mobile sink inorder to avoid the energy hole problem. ProFLex is composed of three phases called tree construction, importance factor distribution and data distribution. ProFlex can guarantee less message overhead, better data gathering efficiency and data availability through replication of data packets and data correlation mechanisms.

4.1 ProFlex System Model

ProFlex (Maia et al., 2013) system model is given as follows.

4.1.1 Nodes and communication between nodes:

During tree construction phase (Maia et al., 2013), ProFlex construct a tree topology which composed of two types of nodes called H-sensor (High-end sensor) and L-sensor (Low-end Sensor) nodes. These different kinds of nodes are responsible for the heterogeneity of WSN. H-sensor node has more sophisticated resources such as improved storage, battery power and communication range compared with L-sensor nodes. Each node is assigned with a communication radius r and it can communicate only with the nodes which are inside its communication radius.

4.1.2 Importance Factor Function:

In this scenario each node is given with a unique id and an importance factor function (I: S $\longrightarrow$ [0, 1]) which defines whether a node can act as a storage node or not (Maia et al., 2013). Here author proposes a uniform selection criterion for the selection of storage node thus all node can play a role of storage node as a result give importance factor as 1 for all the nodes in the network.

4.2 Working

Distributed data storage protocol ProFlex (Maia et al., 2013) composed of mainly 3 phases namely tree construction, importance factor distribution and data distribution. Finally ProFlex introduces a improvement to overcome the message overhead with the help of data correlation mechanism. During tree construction protocol construct multiple trees based on the number of H-sensor nodes and communication range of each H-sensor and L-sensor nodes in the network. Importance factor distribution is responsible for calculation of number of storage nodes in each tree which is done according to the importance factor for each node in the tree. Finally partial view size (buffer size of each node) is calculated from aggregated storage node for each tree, since partial view size is equal to the square root of aggregated storage node. Second phase of ProFlex uses a importance factor distribution algorithm for the calculation of partial view size. The Protocol make an assumption that number of replicas created by an H-sensor node is equal to the partial view size of corresponding tree. The third phase: data distribution phase is responsible for calculating the number of replicas need store in each storage node in each tree this is calculated based on the partial view size and

number of storage nodes of each trees. This phase is also responsible for forwarding the data packets to the storage nodes. Data distribution algorithm is responsible for doing the above tasks.

For instance, Figure: 2 Shows a network having 100 nodes with four H-sensor nodes (P, Q, X, Y) and their respective trees. In that figure the tree Tx rooted at H-sensor node X has 30 storing nodes ($|S_X|=30$) and having 3 H-senor neighbours P, Q and Y. According to Importance Factor Distribution Algorithm (proFlex) the partial view size (buffer size) of all nodes rooted at H-sensor node X calculated as $|S^X_{aggr}| = |S_X|+|S_Y|+|S_P|+|S_Q|$ (30+20+24+26) =100. Thus partial view size of each node rooted at tree Tx is, $|v|=\sqrt{|s^X_{aggr}|}$ =10. Thus all nodes in X's tree will have a partial view size 10. Similarly we can calculate the partial view size for all other trees. Since the number of replicas r (v) is computed based on the partial view size of nodes in the set of storing nodes, the number of replicas of data packets created by H-sensor node X is 10. Data distribution algorithm in proFlex (Maia et al., 2013) compute how many replicas from r (v) goes to its own tree and to each neighbouring tree. In the above example for the tree rooted at H-sensor node X, 3 replicas will store in its own tree and send 2 replicas to nodes in tree rooted at Y, 2 replicas to storage nodes in the tree rooted at Q and finally send 2 replicas to node P (Based on the data distribution algorithm (Maia et al., 2013)).

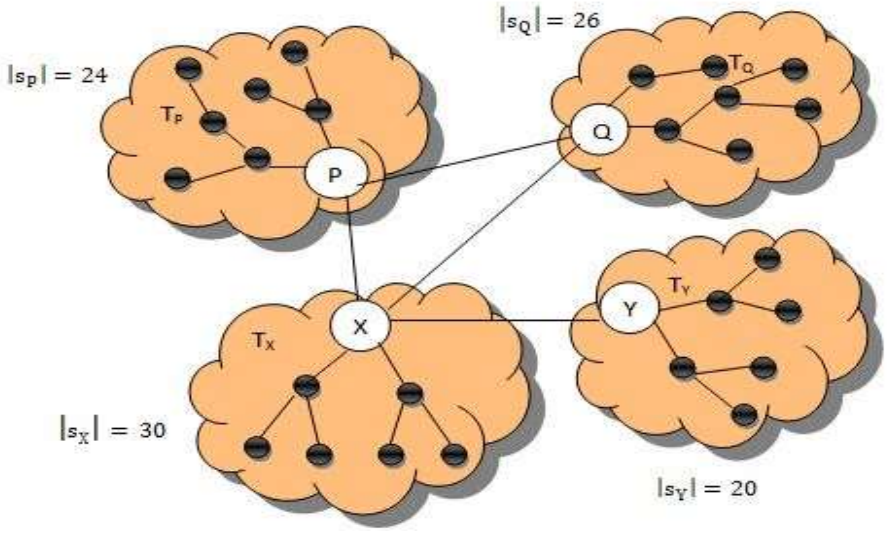

Figure 2: wireless Sensor Network with 100 nodes having 4 as H-sensor node.

4.3 Data Correlation in ProFlex

In order to decrease the message overhead further proFlex introduces a correlation mechanism, which uses an information summarization function to aggregate the data packets reached at each H-sensor node during data distribution algorithm (Maia et al., 2013). Here the data packets are correlated based on a period t and randomly predefined value d (distance). With this approach the H-sensor node continuously receives the data packets from its children's until the time t expires instead of creating the replicas immediately after a single packet received. When period t expires H-sensor node correlate the buffered data packets if and only if the distance between the nodes produces these data packets are less than or equal to the predefined value d. In this way we can reduce the number of data packets and also the replicas as result message overhead can improved to a better level.

4.4 Disadvantages of ProFlex

Data management is the challenging task in wireless sensor networks. ProFlex (Maia et al., 2013) presents advanced distributed data storage in a heterogeneous wireless sensor network. The protocol overcome the

energy hole problem in the previous data storage scheme by introducing mobile sinks. Also it guarantees less message overhead and good data collection efficiency through data correlation mechanisms and replication structure of data. But some difficulties are still their when we consider some special cases. They are listed below:

(a) Excess amount of replicas for data packets.

In proFlex (Maia et al., 2013) multiple replication structure are used inorder to increase the data availability and data gathering efficiency through the use of multiple trees such that each node in each tree loaded with a particular number (Based on the Importance factor distribution algorithm (Maia et al., 2013) of replicas. It will lead to the production excess amount of replicas thus message overhead also increases since protocol need to transmit bulk messages. This also performs unwanted usage of buffer space of each storage node for storing the same copies of data this make the protocol poor in energy efficiency.

(b) Unsuitable Data correlation Mechanism.

The data correlation mechanism in proFlex can reduce message overhead to a better level. But the random selection of distance and time for data aggregation will become a problem if we consider the special case in which the distance may approximately equal to the communication radius and time slot is measured in hours rather than seconds, in such an environment H-sensor want to wait for long period of time and forced to aggregate all the data forwarded by its children's since communication radius of L-sensor nodes are lesser than that of H-sensor nodes. It will lead to H-sensor node failure and may affect the quality of data since unsuitable summarization may take place also network life time become worse because of node failure.

5. CORREP: Efficient Correlation and Replication based Distributed Data Storage Protocol

In this work we propose CORREP, an efficient correlation and replication based distributed data storage protocol. CORREP can overcome the difficulties of ProFlex as mentioned in the previous section 4.4. Finally CORREP can guarantee improved network lifetime and energy efficiency. Also it can keep the data availability and data gathering efficiency same as the original version (Maia et al., 2013). It makes use of mobile sink to collect data from the network so that we can achieve better data collection efficiency. The proposed protocol introduces:

5.1 Energy efficient Replication Method

The replica management systems can provide very high data availability. The use of excess number of replicas may affect the energy efficiency and network lifetime. Thus, CORREP presents a replication scheme for distributed data storage in wireless sensor network which can guarantee both data availability and network lifetime. CORREP achieved the goals by modifying the importance factor distribution algorithm used in ProFlex (Maia et al., 2013). Here we assume the partial view size ($|V|$) of each tree will be equal to $\sqrt{s_{aggr}^{h}}/2$, where $\sqrt{s_{aggr}^{h}}$ is the aggregation of storage node for each tree (Defined in ProFLex). Thus, number replicas will also be reduced since the partial view size is equal to the number of replicas created by the trees ie, r (v) = $\sqrt{s_{aggr}^{h}}/2$ (Maia et al., 2013). Thus the number of total number of replicas in the network is reduced than in ProFlex, as a result storage nodes become energy efficient.

5.2 Data correlation based on communication range

Data correlation in CORREP introduces an advanced method for information summarization scheme. To accomplish this, a small modification to the data correlation carried out in ProFlex is done. When a sensor

node produces a packet, it forwards the data packet to the root of tree as in ProFlex (Maia et al., 2013). H-sensor collects packets from its children and stores the packets in its buffer instead of immediately creating the replicas this process will continue until buffer size of H-sensor become two times of the partial view (buffer size) of Storage nodes of the tree rooted at the corresponding H-sensor node. When H-sensor buffer size reached to two time of the partial view of storage nodes, the H-sensor node summarizes the correlated packets and sending to its own tree and neighbouring tree after creating the replicas as in data distribution algorithm (Maia et al., 2013).We assume the data packets in the buffer can correlate if the distance between the nodes produces the data are less than or equal to the square of the communication range of node that send the data first, ie, $d= r^2$, where r is the communication range of node that send the data to H-sensor node.

5.3 *Performance Analysis of CORREP*

The first step of our performance analysis is based on the impact of reduction made in the number of replicas and there after we analyze the correlation mechanism used in CORREP. Here all the simulations are performed using the NS2 network simulator.

5.3.1 Reduce the Number of Replicas

CORREP used to store less number of replicas in each storage node when compared with ProFlex, so that number of transmitted messages can reduce by maintain the data collection efficiency as similar in ProFlex. This protocol is energy efficient than the ProFLex since buffer usage of each storage nodes executing data storage protocol decreases.

For instance, consider the network in fig.2. We know that $r_P(v) = \sqrt{80}$, $r_Q(v) = \sqrt{50}$ $r_X(v) = \sqrt{100}$ and $r_Y(v) = \sqrt{50}$ Hence each H-sensor node replicates $r_P * |s_P| = 216$; $r_Q * |s_Q| = 234$; $r_x * |s_X| = 300$; $r_Y * |s_Y| = 182$; and thus the total number of replicas in the network using protocol ProFlex is equal to 216+ 243+300+ 182= 932 replicas. The total number of replicas for the same network using the CORREP protocol is 445. (Here, $r_P(v) = \frac{\sqrt{80}}{2}$; $r_Q(v) = \frac{\sqrt{50}}{2}$; $r_X(v) = \frac{\sqrt{100}}{2}$; and $r_Y(v) = \frac{\sqrt{50}}{2}$; Thus number of replicas will be 108 + 117+ 150+ 70 =445). From the above example we can conclude that ProFlex produces 487 more replicas than the CORREP.

5.3.2 Efficient data correlation Mechanisms

CORREP introduces an information summarization with efficient data correlation mechanism. Thus the network lifetime can be improved than the existing protocols. Also it can avoid the complexity in selecting the random values for distance and time slot for data correlation mechanism used in ProFlex (Maia et al., 2013).

For instance, consider the network in fig.2. When sensor nodes in the tree T_X rooted at H-sensor node X send data to node X, it will until the buffer size of X reached two time of the partial view size (buffer size) of storage node in its tree T_X (In this example partial view size of storage node is calculated as 5). Node X starts the summarization of data when its buffer size exceeds two time of partial view size of storage node in tree T_X. Here we assume that the data packets in the local buffer of X node can correlate only if the distance between the storage nodes that produces the data packet less than or equal to the square of the communication radius of storage node that send the data to the H-sensor node X and create replicas of correlated data packets and forwarded to its children's and neighbouring trees as in ProFlex (Maia et al., 2013). Thus network life time can be improved since there is no node failure because of buffer overflow.

6. Conclusion

This paper proposes CORREP, a fully distributed data storage protocol for wireless sensor networks. CORREP is the protocol with high network lifetime and energy efficient when compared with related protocols ProFlex,

supple etc. CORREP transmit less number of messages among the network since it creates only sufficient number of replicas for each data packet. This efficient replication can achieves the similar data collection efficiency as in ProFlex. Moreover, CORREP can guarantee less energy consumption, since the buffer usage is less compared to other protocols. We also propose an improvement to data correlation for information summarization, which is already used in ProFlex. Such an improvement to data correlation was done based on the communication range and partial view size of storage nodes in the wireless sensor networks can improve the network lifetime.

As future work, it would be interesting to add security and data consistency features to CORREP so that situations like attacking of compromising node by the intruders can be avoided and thus confidentiality and quality of sensed data can be improved.

References

[1] C Viana, T.Herault, T.Largillier, S.Peyronnet, F.Zaı¨di (2010), "Supple: a flexible probabilistic data dissemination protocol for wireless sensor networks", 13th ACM International Conference on Modeling.
[2] Guilherme Maia, Daniel L. Guidoni a, Aline C. Viana b, Andre L.L. Aquino c, Raquel A.F (2013), "A distributed data storage protocol for heterogeneous wireless sensor networks with mobile sinks", Ad Hoc Networks 11,pp. 1588–1602.
[3] Krzysztof Piotroski, Peter Langendoerfer and Steffen Peter IHP (2009), "tinyDSM: A Highly Reliable Cooperative Data Storage for Wireless Sensor Networks", 978-1-4244-4586- IEEE.
[4] M. Neenu, T. Sebastian (2013), "Survey On Distributed Data Storage Schemes in Wireless Sensor Networks", IJCSE, Vol. 4 No.6 Dec 2013-Jan 2014, pp.466-473.
[5] M. Vecchio, A.C. Viana, A. Ziviani, R. Friedman (2010), "Deep: density-based proactive data dissemination protocol for wireless sensor networks with uncontrolled sink mobility", Elsevier Computer Communication33 (8) (2010).
[6] Pietro Gonizzi , Gianluigi Ferrari , Vincent Gay b (2013), "Data dissemination scheme for distributed storage for IoT observation systems at large scale" Information Fusion (2013) .
[7] Ren Wei, Ren Yi and Zhang (2010), "Secure, dependable and publicly verifiable distributed data storage in unattended wireless sensor networks", Science China Information Sciences.
[8] Shen Yulong, Xi Ning, Pei Qingqi, Ma Jianfeng (2013), "Distributed storage schemes for controlling data availability in wireless sensor networks", Seventh International Conference on Computational Intelligence and Security.
[9] Wen-Hwa Liao , Kuei-PIng Shih , Wan-Chi Wuaa (2009), "A grid-based dynamic load balancing approach for data-centric storage in wireless sensor networks", Computers and Electrical Engineering 36 (2010) pp.19–30.
[10] Z. Bar-Yossef, R. Friedman, G. Kliot (2008), "RaWMS – random walk based lightweight membership service for wireless ad hoc networks", ACM Transactions on Computer Systems 26 (2008) 5:1–5:66.

A Brief Author Biography

Neenu M. Nair received her B.Tech degree in Computer Science and Engineering from Sarabhai Institute of science and technology (Cochin university), Thiruvananthapuram in 2012 and currently doing her M.Tech-Computer and Communication Engineering degree in Karunya University, Coimbatore. The current project is regarding Data Storage in wireless sensor networks.