Data Duplication Detection and Avoidance in Data Collection Applications

Sanooja M[1*], Chinnu Ravi[2]
[1*]PG Scholar, Dept. of computer science and engineering
[2]Asst.Professor, Dept. of computer science and engineering TKM Institute of Technology, Kollam, India

***Abstract*— RFID (Radio frequency identification) and Wireless sensor networks are two traditional wireless technologies. Here integrating these two technologies for efficient data collection in large mobile monitoring applications, which is called hybrid RFID and WSNs. A special node used in this network, it is called smart node. It combines the functions of RFID reader, reduced function of sensor and RFID tag. For efficient data collection cluster based data collection is performed. The main problem in data collection application is data duplication. The data duplication may waste the energy of senor nodes and create congestion. In cluster based data collection intra and inter cluster duplication should occur. Intra cluster duplication efficiently handled by using different methods. Here propose a new algorithm to avoid inter cluster duplication. It will improve the performance of data collection applications.**

INDEX TERMS*— *Radio frequency identification, wireless sensor network, Hybrid RFID and WSNs, Distributed hash table.

I. INTRODUCTION

Radio Frequency Identification and (RFID) systems and Wireless Sensor Networks (WSNs) represent two key technologies. The integration of RFID and sensor networks can increase their utilities to other scientific and engineering fields by exploiting the advantages of both technologies. RFID systems are mainly used to identify objects or to track their location without providing any indication about the physical condition of the object. RFID technology provides many benefits, such as RFID technology provided a low cost form of data collection and asset management, it enable data collection in environments that are unsuitable for workers as RFID tag can provide data in harsh environments and RFID is able to provide many reads and write functions per seconds, so it is sufficient for most data monitoring applications. RFID systems are mainly used to identify objects or to track their location without providing any indication about the physical condition. of the object . WSNs on the other hand, are networks of small, low cost devices that can cooperate to gather and provide information by sensing environmental conditions such as temperature, light, humidity, pressure, vibration and sound.

The evolution of RFID and WSNs has followed separate research and development paths and has led to distinct technologies. It have many applications where the identity or the location of an object is not sufficient and extra information that can be retrieved through sensing environmental conditions is important. However sensor networks may be used in these environments, the location and identity of an object remain critical information that can be retrieved through RFID systems. The best solution in these cases is the integration of both technologies because they complement each other.

II. RELATED WORK

Transmission of redundant data causes network delay and uses network resources unnecessarily within the network. To reduce the redundant transmission, in-network data filtering has been discussed. To eliminate the in-network redundancy transmissions, [5] Proposed a filtering mechanisms have two phase. In first phase consider only serial number of EPC data. If serial numbers are equal then go to second phase(*Backward-First Filtering*).At the second phase, filter considers the other values of EPC data. This approach uses two height-balanced trees deployed on each node; a tree indexed by ID is used for searching the buffer to decide whether or not input data are redundant. The purpose of other trees indexed by arrival time is to update the contents of the trees to maintain the size of both trees depending on the data arrival time. Paper [6, 7] propose a hash table based in-network filtering that works on RFID readers. To check the data redundancy, a hash table is faster than a tree or list.

Exact and Heuristic Algorithms for Data-Gathering Cluster-Based Wireless Sensor Network Design Problem [10] implies an integrated topology control and routing problem in cluster-based WSNs. To prolong network Lifetime via efficient use of the limited energy at the sensors, adopt a hierarchical network structure with multiple sinks at which the data collected by the sensors are gathered through the cluster heads (CHs).

III. SYSTEM MODEL

A. Data duplication

In order to apply the RFID technology to large scale warehouses, seaports or airports, many RFID readers should be required to cover the large area. To extend the radio

frequency coverage, the authors of [1] Proposed RFID system integrated with wireless sensor system. In their work they use wireless sensor node with RFID reader and one host node connected directly to the host computer. However there are several physical characteristics of wireless sensor nodes to support RFID system, such as power consumption, communication range, and size of RFID readers.

If we use many number of readers to extend the radio frequency coverage and cover the area completely, multiple readers should share the coverage area. So, they generate duplicate readings. In that case, duplicate data will be sent to the host computer thorough wireless sensor network. If we use many number of readers to extend the radio frequency coverage and cover the area completely, multiple readers should share the coverage area. So, they generate duplicate readings.

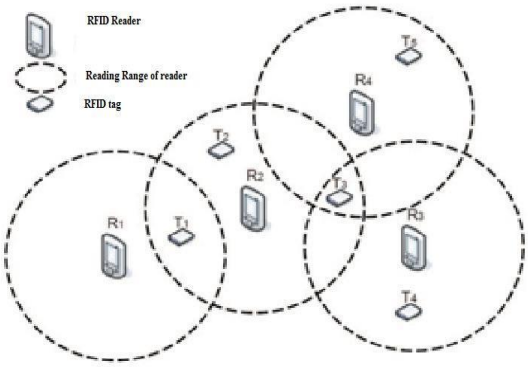

Fig 1. Overlapped reading range of RFID reader

The author of [2] classified these duplications as follows:
Duplication at the data level is generated when individual tags are read multiple times or multiple copies of tags are read once or more.
Duplication at the reader level is incurred when tags in the overlapped reading areas are read by multiple readers. As shown in Fig. 1, tag T1 in the overlapped area can be read by two neighbor readers, R1 and R2. Similarly, tag T3 can be read by three readers at similar time.

B. *Data model*

In this system, readers send the reading data to the host computer, and the data are stored in each reader as historical data for redundancy analysis. The data consist of five fields: <EPC>, <Reader ID>, <Neighbor IDs>, <Reading Timestamp> and <Number of remaining Filtering>.

Field	Tag Id	Reader id	Time stamp	Number of remaining filtering
Bytes	8	4	4	4

Table 1. Structure of RFID data

(Electronic Product Code)
EPC was designed to uniquely identify each object instead of identifying groups or classes of objects as occurred in existing identification techniques such as bar code.
Reader ID
A Reader ID is an identification value used to identify each RFID reader in the system. Reader IDs in the data field show which reader generated the reading data after reading the tag.
Neighbor IDs
Neighbor IDs keep track of all of the Reader IDs within one-hop communication range, since the communication range is larger than read range in most cases. We assume that neighboring readers may have overlapped reading areas; the system uses the Neighbor IDs to decide whether or not two readers have overlapped reading area.
Reading Timestamp
The Reading Timestamp field tracks the time when the readers read the EPC value of tagged objects.
Number of remaining filtering
For filtering, we assign two kind of initial values to *number of* remaining *filtering*, 1 and *fe*. In case of intra-cluster nodes, value of *f* will be 1. Whereas, in inter-cluster node value will be *fe* as shown below:
Number of remaining filtering (f) = 1: need to be filtered at local CH OR fe: need to be filtered at intermediate CH.

C. *Clustering*

Replicated data between any two encountered smart nodes generates a high cost. Concurrent data transmission from number of nodes to an RFID reader causes channel access congestion. For efficient data collection and transmission clustering is implemented. Here, describe two enhanced algorithms called cluster-member based and cluster-head algorithms, in which smart nodes are clustered to number of virtual clusters and each cluster has a cluster head. In the cluster head based algorithm, cluster members replicate their tag data to its cluster head. When a cluster head of a cluster reaches an RFID reader, the RFID reader receives all information of nodes. This enhanced method greatly reduces channel access congestion and reduces the information exchanges between nodes and makes it easy to erase duplicate information in a cluster.

Algorithm of cluster head determination and data transmission conducted by smart node i.

---

1 :Receive cluster head candidates from an RFID reader
2: for each cluster head candidate j do
3: Calculate (fnij - frj)
4: end for
5: Choose the cluster head with max(fnij - frj)
6: if it is a cluster head and meet its cluster member then
7: Read data from the cluster member
8: end if
9: if it is a cluster head and meet an RFID reader then
10: Send its data to the RFID reader
11: end if

To create the clusters in the cluster-member based Algorithm, nodes report their encountering frequency to the server through the RFID readers. The server creates nodes with high encountering frequency into a cluster using the method in [9] and notifies the cluster nodes through the RFID readers. The cluster head for a cluster can be determined in a number of ways depending on the application requirement. RFID readers record the meeting frequency with each node and report the data to the back-end server. The server calculates the sum of the frequencies from different readers for each node j, represented by frj, and selects N nodes with the highest frj as the cluster heads. The information about the selected cluster heads along with their fr is transmitted back to the RFID readers, it will forward the information to the nodes. Here use fnij to denote the meeting frequency between node i and a cluster head j. A node measures its fnij*frj for each cluster head candidate, and selects the one with the highest value as its cluster head. The metric of fnij * frj indicates how fast cluster head j can forward node i's data to an RFID reader. Using RFID readers, each node reports its selected cluster head to the server and the server then notifies all heads about their cluster members.

The head determination performed at the server to reduce the communication. As a result, each cluster head is associated with a group of nodes, and it can quickly forward the data to RFID readers for its cluster members. In this HRW system, since the data is stored in tags, active nodes can extract the information at any time from a sleeping node. In traditional WSNs, nodes in sleeping mode cannot conduct data transmission. Therefore, the HRW system can greatly enhance packet transmission efficiency with the RFID technology. In cluster based data transmission, two type duplication occurred. Intra cluster duplication and Inter cluster duplication. Different methods are implemented to overcome intra cluster duplication. Here proposed new algorithms to overcome this inter cluster data duplication.

D. Inter cluster duplication

Inter-cluster duplication can't be detected by a single CH without transmitting information with neighboring CHs. It will results in a huge communication overhead. Here provided a mechanisms to detect inter-cluster duplications. But, first we need to differentiate among readers that overlap within clusters or across the boundary of a cluster, for which, here introduced the Neighbor Discovery Message [7] . In this method after cluster formation, each node exchanges an *ND* message with neighboring nodes. The ND message includes node *ID* and cluster *ID*. A node which receives *ND* messages from its neighbors keeps the cluster *ID* in an *ND* array. From the *ND* array of a node, we can identify whether it has the *ID* of any neighboring clusters or not. If *IDs* of two or more clusters exist in a node *ND* array that node will be considered as inter-cluster node. However, at the same time one node can form duplication with nodes of the same cluster and with nodes of different clusters.

Algorithm for inter cluster duplication detection

```
Function inter_cluster_duplicated_data_filtering
  Seek the data.tag_id from the tag list.
   If found then
    Decrease the value of f by 1
      If the data is duplicated then
        If the data comes from my cluster then
           Update the value of β field as 'D'
        Else
             If fe – α ≥ f then
              Send Feedback Message
             End if
        End if
      Else
            send the data to the next hop
      End if
  Else
       Insert the data into tag_list
      If the data come from my cluster then
         /*only their own cluster head keeps the
           observation and the relay ratio record*/
        If tag list has the value of N and β is 'D'
        then
       send the data to the N
       Else Update the tag_list with the value of β field as 'R'
       End if
       Else
       send the data to the next hop
       End if
End if
```

When a cluster head receives data packets from its cluster members, it checks from the tag list whether the incoming RFID data packet is already received or not. The structure of the tag list is shown in Table 2.

Tag Id	Reader id	Time stamp	Observation	Redundant Reader
8	4	4	1	4

Table 2. Structure of tag list

The observation *(β)* field has two flags such as *R* and *D*. *R* means that the RFID packet is successfully relayed to the sink node and *D* represents that the RFID packet is dropped for duplication at an intermediate node. Redundant reader *ID (N)* indicates the reader that reads the tag and generates the intra-cluster duplications. If tag *ID*, reader *ID*, and time stamp all match; and value of *β* is as '*D*' and *N* exists, it means this data is already dropped at previous readings

After inter-cluster duplicate detection, intermediate CHs will inform with a feedback message to CHs whose nodes are generating duplicate data packets. Later, those CHs can change routing paths of duplicate data to eliminate it close to source, at neighboring CHs, to avoid redundant transmissions from data generation point to detection point

IV. PERFORMANCE EVALUATION

Transmission of redundant data causes network delay and consumes network resources unnecessarily in an RFID system integrated with WSN. Previous literature proposed methods that reduce redundant transmissions. However, they still induce network delay because of inter cluster duplication existing there.

This paper proposed an algorithm to detect and eliminate inter cluster duplication. The sensed data from a member node get transmitted to the sink node through different cluster heads. Cluster head determine intra and inter cluster duplication. Inter cluster duplication determine at nearby nodes and inform the source cluster head through feedback message. Then the duplicated cluster head change the route of the duplicate data to the nearby cluster head. Nearby cluster head filter the data and send the original data to the sink. By applying this, duplication transmission can be reduced. As a result transmission delay reduced, efficiency increased, congestion reduced.

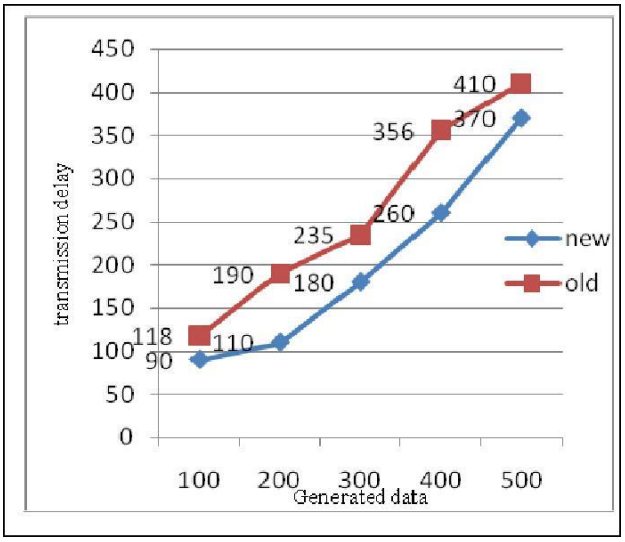

Fig 3.Comparisson of transmission delay

V. CONCLUSION

This system propose Hybrid RFID and WSN System (HRW) which integrates the multi-hop transmission mode of WSNs and direct transmission mode of RFID systems to improve the efficiency of data collection, to meet the requirements of low economic cost, high performance and real-time monitoring in mobile monitoring applications. HRW is composed of RFID readers and hybrid smart nodes. Proposed system performs efficient cluster based data collection and also avoiding intra and inter cluster duplication accurately. This work saves communication and computational cost and increases the network lifetime compare to other literature solutions.

REFERENCES

[1] Mark L. McKelvin, Jr. Mitchel L. Williams and Nina M. Berry, "Integrated Radio Frequency Identification and Wireless Sensor Network Architecture for automated Inventory Management and Tracking Applications," In the *Proceedings of the 2005 conference on Diversity in computing*, October 2005]

[2] Elimination for RFID Data Streams", International Journal of Principles and Applications of Information Science and Technology, Vol.1, No.1
December, 2007.

[3] D. R. Howe, *Data Analysis for Database Design*, 3rd ed., Butterworth- Heinemann, 2001, pp. 39-46.

[4] R. Derakhshan, M.E. Orlowska, X. Li, "RFID Data Management:Challenges and Opportunities", IEEE International Conference on RFID,2007.

[5] Choi, W.; Park, M.S. In-network Phased Filtering Mechanism for a Large-Scale RFID Inventory Application. In *Proceedings of the 4th International Conference on IT & Applications (ICITA)*,Harbin, China, January 2007; pp. 401-405.

[6] Kim, D.S.; Kashif, A.; Ming, X.; Kim, J.H.; Park, M.S. Energy Efficient In-Network Phase Rfid Data Filtering Scheme. In *Proceedings of the 5th*

International Conference on Ubiquitous Intelligence and Computing, *UIC 2008*, Oslo, Norway, 23–25 June 2008; pp. 311-322.

[7] *Ali Kashif Bashir, Se-Jung Lim, Chauhdary Sajjad Hussain and Myong-Soon Park*,"Energy Efficient In-network RFID Data Filtering Scheme in Wireless Sensor Networks", *www.mdpi.com/journal/sensors*, *6 July 2011*

[8] Dong-Hyun Lee, Eun-Mook Lee, Ali Kashif Bashir and Myong-Soon Park," Efficient In-Network Redundancy Filtering in RFID System Integrated with Wireless Sensor Networks", Department of Computer and Radio Communications Engineering, Korea University Seoul, Korea,

[9] F. Li and J. Wu, ''MOPS: Providing Content-Based Service in Disruption-Tolerant Networks,'' in Proc. ICDCS, 2009,pp. 526-533.

[10] Hui Lin and Halit Üster," Exact and Heuristic Algorithms for Data-Gathering Cluster-Based Wireless Sensor Network Design Problem", ieee transactions networking,vol. 22, june 2014.