A Data-Centric Storage Approach for Efficient Query of Large-Scale Smart Grid

Yan Wang[1, 2], Qingxu Deng[1], Wei Liu[1]
1 School of Information Science and Engineering
Northeastern University
Shenyang, P.R.China
dengqx@mail.neu.edu.cn

Baoyan Song[2]
2 School of Information Sci. and Tech.
Liaoning University
Shenyang, P.R.China
bysong@lnu.edu.cn

***Abstract*—Smart Grid is an important application in Internet Of Things (IOT). Monitoring data in large-scale smart grid are massive, real-time and dynamic which collected by a lot of sensors, Intelligent Electronic Devices (IED) and etc.. All on account of that, traditional centralized storage proposals aren't applicable to data storage in large-scale smart grid. Therefore, we propose a data-centric storage approach in support of monitoring system in large-scale smart grid: Hierarchical Extended Storage Mechanism for Massive Dynamic Data (HES). HES stores monitoring data in different area according to data types. It can add storage nodes dynamically by coding method with extended hash function for avoiding data loss of incidents and frequent events. Monitoring data are stored dispersedly in the nodes of the same player by the multi-threshold levels means in HES, which avoids load skew. The simulation results show that HES satisfies the needs of massive dynamic data storage, and achieves load balance and a longer life cycle of monitoring network.**

Keywords- IOT; large-scale Smart Grid; Hierarchical Extended Storage; extended hash function

I. INTRODUCTION

Based on robust spatial grid structure and communication platform, smart grid should realize the complete integration of electric flow, information flow and business flow, covering every section of power system such as generating, transferring, transforming, distributing and utilizing. Smart grid [1~3] depends on on-line monitoring and real-time information collection of operating parameter in electric power system. For example, monitoring the high pressure electrical installations and power transmission line, smart grid need widely deploy sensors, IED and etc. in order for transmission line to have the capacity to combat natural adversities. Since the monitoring data of large-scale smart grid are massive, real-time and dynamic, data storage and query should partially move to the perception layer of IOT such as WSN for processing data to make decision timely. So, it is critical to research data storage approach for large-scale smart grid for query effectively.

There are several systems used in monitor and control field in pilot projects of Smart Grid now, such as Supervisory Control And Data Acquisition (SCADA)[4], Wide Area Measurement System (WAMS)[5] and etc.. These pilot projects aren't broad in scale, real-time data are transferred to master station and then power network model and historical data are saved in ORACLE relational database in the control centers [6][7]. Nevertheless, the number of electrical installations being monitored would increase rapidly in a large-scale smart grid. It is necessary for transferring and analyzing massive dynamic data timely. Tremendous amounts of data are delivered to relational database server in master center, which would lead to congested collapse. And relational database isn't a better solution for storage and read/write of massive dynamic data in real time.Recently data-centric storage (DCS) [8~12] schemes have been researched by many scholars. More approaches are mentioned as follows: Geographic Hash Table (GHT)[13], Distributed Index for Multidimensional data (DIM)[14], Dynamic load balancing (DLB)[15] and etc. These approaches couldn't fit to data storage of monitoring system in large-scale smart grid.

For the special architecture of monitoring system in large-scale smart grid, the paper proposes a data-centric storage approach in support of a large-scale smart grid: Hierarchical Extended Storage Mechanism for Massive Dynamic Data. HES sorts and stores the massive data according to data types. When data are on the increase gradually, HES can add storage nodes dynamically by coding method with extended hash function for avoiding data loss of incidents and frequent events. Monitor data are stored dispersedly in the nodes of the same player by the multi-threshold levels storage means in HES, which achieves load balance and performance improvement. So, HES can reduce network resources utilization of data transmission and promote efficiency of query and accuracy of intelligent control.

II. STORAGE MODEL OF MONITORING SYSTEM IN LARGE-SCALE SMART GRID

A. Storage Model

The monitoring system of smart grid disposed a large number of sensors and IED nodes on the high-voltage equipment putted in the transformer substation and transmission lines. Sensors are the role of collecting and transmitting data. IED is made of calculation module, memory module, communication module and electronic device with sufficient energy. IED nodes are in charge of accepting the monitoring data from sensors and storing them for a period of time. The intelligent devices have been widely used in the power system.

*Supported by National Natural Science Foundation of China(No. 60973017, 60873068) and Excellent Talent Support Program of Education Department of Liaoning Province(No. LR201017)

978-0-7695-4819-7/12 $26.00 © 2012 IEEE
DOI 10.1109/WISA.2012.27

This paper bases on the practical application of smart grid, and abstracts the monitoring system of large-scale smart grid as the storage model in Figure 1. The model has been divided logically into equal-sized two-dimensional grid called sub-grid based on the unit of transformer substation. Every sub-grid has one ID expressing (X, Y). X and Y are both positive integers. It is supposed that a node in monitor system receives its location through a certain positioning algorithm (such as GPS). Every sub-grid covers many devices，which include transformer, instrument transformer, lighting arrester and switch equipment. They are also placed on the region of substations and their covering transmission lines. A large numbers of sensors, storage IEDs and calculator IEDs are disposed in these devices. Function and description of these nodes are introduced as follows.

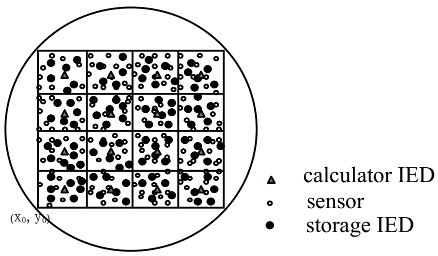

Figure 1. A diagram of storage model of monitoring system in large-scale smart grid

B. Function and Description of Nodes

- Sensor Node

In the grid, a large numbers of data sensors with low-power consumption, low-cost, miniaturization, and high precision have been disposed. Sensors are in charge of collecting and transmitting data. Sensor doesn't have memory module.

- Storage IED Node

The function of storage IED node is mainly the thing that they accept the data detected forwarded from calculator IED nodes and store them for a period of time. Storage IED nodes have a limited compute capacity.

- Calculator IED Node

An intelligent electronic device which be designed according to practice need was put in the center location of every sub-grid. We defined it as the calculator IED. It has three special features: (1) Accept data detected from sensors in the same sub-grid, and calculate the ID of destination sub-grid where the data should be stored, and then transmit the data. (2) Accept the data having the same data type with this sub-grid, and transfer the data to the appropriate storage nodes. (3) Manage the storage IEDs in its own sub-grid, and achieve the function of dynamic expansion of storage IEDs.

To achieve the above functions, the calculator IED has integrated the strong calculation module and amount of memory module. Meanwhile, in these nodes we defined the Node_Grid table. Its structure is shown as follows: (1) G_{ID}: sub-grid ID which storage IED belongs to; (2) B_{ID}: binary code of a storage IED; (3) L_R: real geographic location of a storage IED; (4) L_V: virtual coordinate of a storage IED; (5) T_L: current threshold level of a storage IED; (6) N_L: current storage level of a storage IED; (7) D_T: data type.

III. HIERARCHICAL EXTENDED STORAGE MECHANISM

A. Destination Node Discovery

While a sensor node, which exists in a certain sub-grid, detects data and then generates a Data Message, the sensor firstly transmits it to the calculator IEDs, which belongs to the same sub-grid with the sensor. The structure of Data Message is shown as follows: (1) D_T; (2) G_{ID}: the ID of sub-grid which includes the sensor that detects the data; (3) Data: data detected.

After a calculator IED has accepted the data, it calculated the destination sub-grid ID where the data should be matched in data type, utilizing appropriate hash function according to data type. And then Put Message generated by calculator IED is transferred to the destination sub-grid by GPRS/CDMA/3G. The structure of the Put Message is shown as follows: (1) G_{ID}: destination sub-grid ID; (2) D_T; (3) Data. At last, the calculator IED in destination sub-grid computes the storage IED that is the most appropriate node that stores the data in the own sub-grid. Above steps are shown in Figure 2.

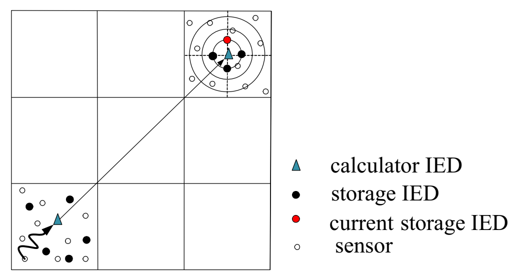

Figure 2. The diagram of data transfering and storing

B. Hierarchical Extended Storage Mechanism

Data detected is forwarded to the destination storage sub-grid, after the Calculator IED computes the destination sub-grid ID based on the data type. As irregular distribution of data detected, the monitoring system difficultly predicts the number of storage IEDs when there are incidents or frequent events. Therefore, to solve the above problems, we propose Hierarchical Extended Storage Mechanism for Massive Dynamic Data in this paper, which can store mass monitoring data according to data types. This mechanism relies on two schemes: one is coding method with extended hash function which could increase the storage nodes dynamically on account of the amount of data actually received to avoid missing data; the other is the multi-threshold levels storage means stores data to multiple storage IEDs of the same layer to avoid suffering from hotspot storage problem, and then achieves load balancing.

1) The initialization of calculator IED node: A large number of storage IEDs are used of storing data detected in monitoring system of smart grid. When system initializing, each storage node sends an Init Message to calculate IED within the same sub-grid. An Init Message contains two parts: (1) L_R: geographic location of a storage node; (2) G_{ID}: sub-grid

ID of storage node belonged to. When calculate IED receives an Init Message, it will add information to Node_Grid table according to G_{ID} of Init Message. And then the calculate IED sorts all storage IED in its sub-grid by distances from the calculate IED ordering by ascending. And then every calculate IED maintains these information in the whole storage procedure.

2) Coding method with extended hash function: Calculate IED chooses the nearest 2^N （N=1,2…）nodes to itself in Node_Grid table as the first layer of storage nodes according to predicting actual data detected, and codes these 2^N nodes by N-bit binary number sequentially. For example, if N=2, the code of storage nodes in the first layer is 00, 01, 10, 11(see Figure 3.). And then calculate IED modifies the N_L value of these nodes to 1 in Node_Grid table, that is N_L=1.With the increasing amount of data detected, the first layer of storage nodes which have no more storage space, now it is necessary to extend storage layer dynamiclly. Then calculate IED chooses the nearest 2^{N+N_L} nodes to itself in other nodes as the next layer of storage nodes, and codes these nodes by N+N_L bit binary number sequentially. And then calculate IED modifies the N_L value, that is N_L=N_L+1. Coding method with extended hash function can increase the storage nodes dynamically on the basis of the actual amount of monitoring data to avoid data losting.

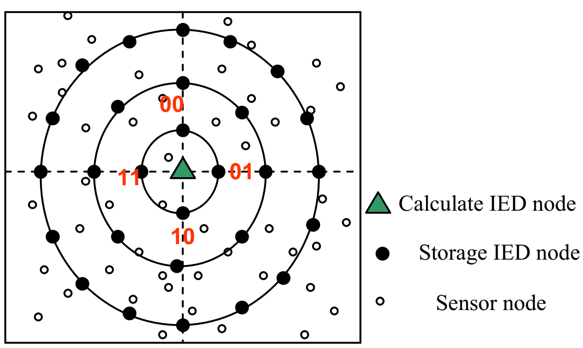

Figure 3. The initial code of Storage nodes with 2^N=4

3) Multi-threshold levels storage means: To avoid load skew, we propose a multi-threshold levels storage means , when plentiful amount of data detected are stored in the storage nodes of the same layer. The storage state of storage node is divided into multi-threshold levles according to storage capacity of hardware and amount of data in smart grid. For example, the total storage capacity of a storage node is S, which is divided into n levels, namely T1, T2,…Tn, Ti < S, and Ti < Ti+1 (i = 1 … n). Storage nodes of every layer store data collected sequently according to their binary code from small to large untill storage levels of all the storage nodes in the same layer come to i leve, and then be transitted to i+1 level （i = 1…n). This method evenly stores data to all storage IEDs of the same layer to avoid a node being exhausted quickly when data storing and querying are focus on it, and then achieves load balance. When storage level of all storage nodes in the i layer comes Ti threshold level, the last one should send Storage_Full Message to the calculate IED in its sub-grid. Thus, the calculate IED changes L_V and T_L in its Node_Grid table on the basis of Storage_Full Message. Storage_Full Message contains three parts: (1) G_{ID}: sub-grid ID of a storage node; (2) L_V: virtual coordinate of a storage IED; (3) T_L: current threshold level of a storage node.

4) HES algorithm: Monitoring data are massive in smart grid and it is difficult to predict the number of storage nodes required for incidents or frequent events. So much monitoring data will be lost if there are no enough storage nodes. HES strategy is used when all storage nodes had coded of a layer is full, calculate IED will extend dynamically storage nodes by coding method with extended hash function. These extended storage nodes are used for storing massive monitoring data (See algorithm 1).

Algorithm 1: Coding method with extended hash function.
Input: N_L and N bit for coding in the first layer.
Output: Binary codes of all nodes in N_L layer.
Step 1: If (N $\leqslant$ 0) then go to step 6.
Step 2: Choose the nearest 2^{N+N_L} nodes as storage nodes of N_L+1 layer.
Step 3: Coding these nodes by N+N_L bit binary number sequentially.
Step 4: Alter Node_Grid table by setting $N_L = N_L+1$ on 2^{N+N_L} nodes chosen in step 3.
Step 5: Output binary code of every node in N_L layer.
Step 6: End.

For all the storage nodes of every layer, monitoring data are stored in these nodes evenly by multi-threshold levels storage means, so that all the operations about storage and query are not concentrated on one storage node. Therefor, this approach can make all storage nodes in the same layer get to load balance(See algorithm 2). In addition, monitoring data stored in storage node have a certain storage period. Along with storage nodes extending hierarchily outward through hierarchical extended storage strategy, the storage node of inner player can execute data backup when the nodes of outer player store data, which can improves system availability. Moreover, monitoring data in sub-grid are stored extended from inner to outer layer according to the sequence of occurrence time and data type, which method is propitious to query of application in smart grid.

Algorithm 2: Multi-threshold levels storage means.
Input: S, N_L and n.
Step 1: T_i =i×S/n, i = 1..n.
Step 2: If $S_j \geqslant T_i$ (S_j is data volume of a node in N_L layer, j = 1..m,) then send Storage_Full Message to Calculate IED.
Step 3: Alter Node_Grid table by setting L_V = (∞, ∞) and T_L = T_L +1 in Calculate IED. Go to step 2 until $S_j \geqslant T_i$ of all nodes in N_L layer
Step 4: Alter Node_Grid table by setting L_V = (0, 0). Set i = i + 1. Go to step 1 until i = n + 1.

C. an Example of HES Mechanism

For instance, data type in a sub-grid is freeze event on transmission lines. Initially, calculate IED adds a nodes' information in Node_Grid table according to G_{ID} value when it receives an Init Message. And calculate IED computes the distance between storage IED and calculate IED, and then stores the node in an appropriate position of Node_Grid table in order (see Table 1). When initializing Node_Grid table, set initial L_V, N_L and T_L of all storage nodes of every sub-grid, which is $B_{ID} = 0$, $L_V = (0, 0)$, $N_L = 0$ and $T_L = 0$.

Calculate IED chooses the nearest 2^N (N=1,2…) nodes to itself in Node_Grid table as the first storage layer according to predicting actual data detected, and codes these 2^N nodes by N-bit binary number sequentially. Here, if N=2, the code of storage nodes in the first layer is 00, 01, 10, 11. And then set $N_L = 1$ of all storage IEDs coded in this layer (see Table 1).

After receiving a Put Message, Calculate IED initially checks ID of a sub-grid (G_{ID}) and data type (D_T) of the Put Message whether equals to corresponding attributes in Node_Grid table. If it is the case, the calculate IED finds the first storage node whose L_V is not equal to (∞, ∞) in the table. It means that the node which $L_V \neq (\infty, \infty)$ is the nearest usable one from calculate IED. And then the calculate IED forward data to it.

Assume the total storage capacity of a storage IED is 100, which is divided into three levels, $T_1 = 30$, $T_2 = 60$, $T_3 = 90$, remaining storage space will be used to other applications. In table 1, Calculate IED has received Put Message whose G_{ID} and D_T are the same with its G_{ID} and D_T in Node_Grid table, then forwards data to the storage node which L_R value is equal to (75,85). When the storage threshold level of this node has arrived T_1 level, it will send Storage_Full Message to the calculate IED in its sub-grid. Thus, the calculate IED changes L_V and T_L in its Node_Grid table on the basis of Storage_Full Message. Data reached subsequently will be stored in other nodes of this layer until threshold levels of all the storage nodes in this layer come to T_1, then the calculate IED will modify their virtual coordinate L_V to (0,0), that is $L_V = (0, 0)$. And so on, it is not halt until the storage threshold levels of all the storage nodes in this layer get to the highest threshold level. If all storage nodes in every layer get to the highest threshold level, L_V need not be reset.

Table 1 Node_Grid table when node 00 got to T1 threshold level

G_{ID}	B_{ID}	L_R	L_V	T_L	N_L	D_T
(2,2)	00	(75,85)	(∞, ∞)	3	1	freeze
(2,2)	01	(78,74)	(∞, ∞)	3	1	freeze
(2,2)	10	(75,73)	(∞, ∞)	3	1	freeze
(2,2)	11	(68,70)	(∞, ∞)	3	1	freeze
(2,2)	0	(60,68)	(0,0)	0	0	freeze
(2,2)	0	(66,55)	(0,0)	0	0	freeze
(2,2)	…	…	…	…	…	…

When storage capacity of all storage nodes in the first layer is full, the calculate IED would choose again the nearest $2^{2+1} = 8$ nodes from other storage nodes which $L_V = (0, 0)$ in Node_Grid table. These eight nodes are as storage nodes of the second layer, and coded by three-bit binary numbers sequentially. They are namely 000, 001, 010, 011, 100, 101, 110, and 111. At the same time, the calculate IED modifies the value of these eight nodes' N_L in the Node_Grid table, that is $N_L = 2$. Data reached subsequently are also stored evenly in the second layer using the multi-threshold levels storage means. Same as above, storage nodes will be extended outward hierarchically. When data of inner layers have expired, the calculate IED backups them to the database servers of center substations. In the meantime, the calculate IED need make the modification on corresponding storage nodes' attributes which are $B_{ID} = 0$, $L_V = (0, 0)$, $T_L = 0$, $N_L = 0$.

IV. EXPERIMENTAL EVALUATION

We conducted extensive experiments on storage performance of HES by using VC++ language. In order to illustrate the effectiveness of HES, we have compared HES with SCADA and WAMS in amount of data stored and load of storage nodes on monitor network.

A. Experimental Parameters

The experimental environment was Pentium 2.0GHzdde CUP, Memory 2G, and the Operating System is Windows XP. Experimental parameters are shown in table 2.

Table 2 Experimental Parameters

Parameters	Values
The number of sensors	2000
The number of RTUs	2000
The number of PMUs	2000
The number of data concentrators	50
The number of IEDs	100
Area of grid X*Y	400m*400m
The number of data types	10
Amount of data	3000
The size of a message	8 bytes
Energy consumption sending a message	50 nJ/bit
Energy consumption receiving a message	50 nJ/bit
Energy consumption forwarding a message	0.1 nJ/bit/m
Transfers radius of sensor	30m
Initial energy of a sensor	1J

B. Amount of Data Stored

SCADA system monitors electrical installations by Remote Terminal Units (RTU) and controls them in power plant, substation and other places. WAMS system has capable of collecting dynamic operation data at high rate with Phasor Measurement Units (PMUs). RTU in SCADA system integrates with sensors. Data collected by PMU are transferred to center stations by data concentrators in WAMS system. Figure 4 illustrates the change of amount of data stored in SCADA, WAMS and HES, when data increase gradually. Above three mechanisms don't loss data, when data increase from 500 to 1500. SCADA and WAMS systems begin to loss data when the number of data exceeds 1500. Storage capacity of RTU is limited in SCADA, which leads to serious loss and poor accuracy. Since data concentrators in WAMS system have a larger storage capacity, performance in WAMS system is better than that in SCADA. HES can add storage nodes dynamically with data increasing, so a small amount of data is lost when data are close to 3000. It ensures data integrity and better accuracy in smart grid.

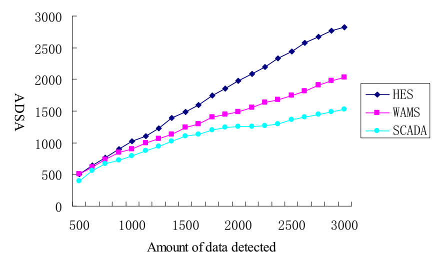

Figure 4. Amount of data stored actually (ADSA) in monitoring systems

C. Load of Storage Nodes

Central storage of data tends to cause load skew of nodes in SCADA and WAMS, nevertheless monitor data are stored dispersedly in the many nodes of the same player by the multi-threshold levels means in HES. The experiment describes the load of above three mechanisms through standard deviation of storage space in Figure 5.

Definition (*Standard Deviation of Storage Space (SDSS)*) It is also called mean square error of storage space which is an average of the differentials subtracting mean of amount of data stored from every storage node. Standard Deviation of Storage Space can reflect the dispersion degree of a data set. The value of SDSS is higher, which indicates more difference between storage size of mostly storage nodes and average storage. The value of SDSS is lower, which indicates storage size of a majority of storage nodes near to average storage.

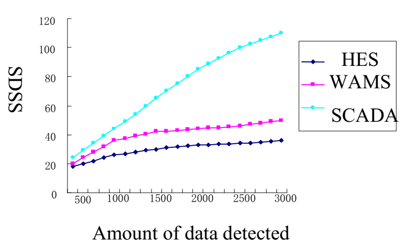

Figure 5. Evaluation of Standard Deviation of Storage Space (SDSS)

The more standard deviation is lower, the better load. When data raising continuously, this result is illustrated that standard deviation increases rapidly on account of limited storage of sensors in SCADA. And standard deviation in WAMS increases more slowly than SCADA because of data concentrators. There are a lot of storage nodes to store data uniformly, hence standard deviation increasingly neatly in HES. This contributes to load balance of storage nodes and efficiency of query.

V. CONCLUSION

In this paper, we have presented a data-centric storage approach in support of a large-scale smart grid: Hierarchical Extended Storage Mechanism for Massive Dynamic Data. HES can add storage nodes dynamically by coding method with extended hash function for avoiding data loss of incidents and frequent events. Monitor data is stored dispersedly in the nodes of the same player by the multi-threshold levels means in HES, which achieves load balance. The simulation results show that HES satisfies the needs of massive dynamic data storage, and achieves load balance. Furthermore, HES has improved usability of monitoring system and query efficiency. The data-centric storage approach proposed in this paper is significant for real-time data processing of a large-scale smart grid.

REFERENCES

[1] Shu-yong CHEN, Shu-fang SONG, Lan-xin LI, Jie SHEN, “Survey on Smart Grid Technology”, Power System Technology, Beijing, Apr. 2009, Vol. 33 No. 8.

[2] Zhi-zhe ZHANG, Xing-yuan LI, Shi-jie CHENG, “Functions and Implementation of United Information System for Smart Grid”, Proceedings of the CSEE, Beijing, Dec.5, 2010, Vol.30 No.34.

[3] Shi-jie CHENG, Xing-yuan LI, Zhizhe ZHANG, “Entire-grid-area Information-sharing and Integrated Applications in United Information System for Smart Grid”, Proceedings of the CSEE, Beijing, Jan.5, 2011, Vol.31(1)

[4] Zheng-feng WANG, Tai-gui HUANG, Di WU, et al., “Application of wide-area measurement system to power systems”, East China Electric Power, Shanghai, 2007, 359（5） : 44-48.

[5] Zheng-feng WANG, Xiao-fei HU, Tai-gui HUANG, “Key technical matters existed in application of wide area measurement system to power system”, Northeast Electric Power Technology, Shenyang, 2008,28（8）:1-3.

[6] Jun-ming Wang, “Some Highlights in Relation to Smart Grid”, Power System Technology, Beijing, Oct. 2009, Vol. 33 No. 18.

[7] Gui-he DU， Zheng-feng WANG, “Design and research on power network dispatching integration of smart grid”, Power System Protection and Control, Henan, Aug. 1, 2010, Vol.38 No.15.

[8] R. Sarkar, X. Zhu, and J. Gao, “Double Rulings for Information Brokerage in Sensor Networks”, Proceedings of ACM MobiCom, 2006.

[9] S. Shenker, S. Ratnasamy, B. Karp, R. Govindan, and D. Estrin, “Data-Centric Storage in Sensornets”, in Proceedings of ACM HotNets, 2002.

[10] S. Ratnasamy, D. Estrin, R. Govindan, B. Karp, S. Shenker, Y. Li, et al., “Data-centric storage in sensornets,” Proceedings of the first workshop on sensor networks and applications(WSNA), 2002.

[11] D. Ganesan, D. Estrin, J. Heidemann, “Dimensions:why do wo need a new data handling architecture for sensor networks”, ACM SIGCOMM Computer Communication Review, January 2003, Volume 33 Issue 1.

[12] M. Aly, N. Morsillo, P.K. Chrysanthis, K. Pruhs, “Zone sharing: a hotspots decomposition scheme for data-centric storage in sensor networks”, International Conference on Very Large Data Bases (VLDB) Workshop on Data Management for Sensor Networks (DMSN 2005), August, 2005.

[13] S. Ratnasamy, B. Karp, L. Yin, F. Yu, R. Govindan, S. Shenker,et al., “GHT:a geographic hash table for data-centric strorage”, Proceedings of the first ACM international workshop on wireless sensor networks and applications, 2002.

[14] Xin Li, Young Jin Kim, Ramesh Govindan,et a1., “Multi-dimensional range queries in sensor networks”, Proceedings of the first ACM international conference on embedded networked sensor systems(SenSys), November 2003, 63-75.

[15] Wen-Hwa Liao, Kuei-Ping Shih, Wan-Chi Wu, “A grid-based dynamic load balancing approach for data-centric storage in wireless sensor networks”, Computers and Electrical Engineering, 36(2010) 19-30.