Cooperative Data Reduction in Wireless Sensor Network

Qingquan Zhang*, Ting Zhu†, Ping Yi‡, Yu Gu§
*Department of Electrical and Computer Engineering, University of Minnesota, Twin City, MN 55414, USA
†Department of Computer Science, Binghamton University, Binghamton, NY, 13905 , USA
‡School of Information Security Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
§ Singapore University of Technology and Design, 138682, Singapore

***Abstract*—Due to the limited power constraint in sensors, dynamic scheduling with data quality management is strongly preferred in long lifetime monitoring applications. But typical techniques treat data management as an isolated process on only selected individual nodes, e.g. the centroid node. In this paper, we propose and evaluate an aggressive data reduction algorithm based on error inference within sensor segments. The architecture integrates three parallel dynamic error control mechanisms to optimize the trade-off between energy saving and data validity. We demonstrate that not only substantial energy savings can be achieved but also that an error bound specified by the application can be guaranteed. Moreover, we have investigate the system performance by using the realistic historical soil temperature data as an experimental context. The experimental results demonstrate that the system error meets the specified error tolerance and produces up to a 50 percent of the energy savings compared to several sensing schemes.**

I. Introduction

Wireless Sensor Networks (WSNs), consisting of thousands of low-cost sensor nodes having different type of sensors (e.g. Light sensor, Temperature Sensor and etc) installed to detect interested physical properties, have been used in many application domains. Due to the limited power supply and difficulties in harvesting ambient energy, low power energy management obviously is a critical research problem in WSNs [1], [2], [3], [4], [5]. In order to achieve missions, such as persistent surveillance and tracking, energy efficient sensing schemes have to be developed to extend the lifetime of the network. Energy consumption determines the lifetime of a sensor network, and communicating wirelessly consumes more power than any other activities. Hence, it is important to design protocols so as to minimize the amount of communication required by the sensor nodes. In the past few years, many solutions [6], [7] have been proposed for energy conservation by applying the data aggregation protocol in which one centroid node (i.e., the sensor node in a group that all other sensors report their data to at a given schedule) collects and packs several communication packets into one packet. Most data aggregation methods [6], [7] achieved energy saving by taking advantage of the fact that most of the data packets sent to the centroid nodes will share similar packet headers, which include training sequences for transceiver clock synchronization, framing information, destination address, and error control codes, typically makes up a large component of the sensor data packets, especially at low data rates such as those in environmental sensing applications. More saving can be achieved by combining those packets with similar header into one or few packets; Even though those methods show some interesting results, data inherent correlation to simplicity and efficient manner is underestimated which can be exploited in our data reduction algorithm. Generally energy saving in data aggregation has to be achieved by using a robust sensor node equipped with strong power supply and relative super computation capacity, which imply the increase in design complexity. Moreover those aggregation approaches are not easy to achieve in practice due to connectivity and reliability issues, as all the sensors have to coordinate scheduling and routing their packets so that the data can be aggregated along the path.

To release the strict requirement of hardware and design complexity, we propose a dynamic and systematic data reduction approach, called DR3, which is created based upon three levels of communication reduction in wireless sensor network architecture. Data reduction, differently to data aggregation, cuts down significant amount of the data communicated among sensor nodes between/within sensor groups. It doesn't requires all the sensor nodes to remain active most of the time, yet still achieve high accuracy of data even at situations when quality connectivity cannot be maintained.

Most data aggregation approaches aggregate data within a group. Data are sampled at each sensor nodes most of the time. Suppose that the data intended is the mean value, the sample data get a random distribution within certain data range, the possibility will increase when there exists at least one data among the sample group whose reading is close enough to intended results. Let's take a temperature sampling as an example. Suppose temperature ranging from $0^o C$ and $35^o C$ is uniformly distributed along an area. If sensor nodes are placed uniformly in this area in a grid topology, it is expected that the higher the intensity of the nodes, the higher possibility that one grid having a node will cover where sample mean will fall into. This characteristic motivates us to explore potentials of data reduction extensively by creating the DR3 system framework.

The highlights of such design can be categorized as a novel three-level data reduction: dynamic centroid selection, duty-cycle adjustment with machine learning prediction and correlative inter-group data reduction. According to our knowledge, this is the first design that supports self-adaptive centroid node selection and uses centroid's reading to represent the group's reading instead of waiting for all data ready to further reduce the communication packets in wireless sensor network.

II. Overview and Objectives

Typical sampling systems, such as habitat monitoring [8], make use of the data aggregation and compression techniques. In wireless sensor network applications, the energy cost associated switching operating status for transceivers are significantly large enough compared to energy consumption in sensing. It is wisely to compress several packets into one or aggregate the data collected from different packet into one data sector before transferring to transceiver for communication as done in TAG [9]. Since data compression and aggregation such as TAG need to buffer a relatively large amount of sensing data from different location, they are not quite suitable for time-critical monitoring application and impose some requirement on hardware.

A. Data Management in DR3

The strategies investigated in our DR3 scheme are mainly motivated by a long lifetime environment monitoring application in which sustainability and data accuracy are of the best interest. We intend to construct a system which can cover large area for data acquisition while requiring less maintenance and less complexity in data processing. These design requirements are quite critical for practical application especially when energy renovation technique is not accessible due to cost issues. Several features are considered in term of power conservation in our design:

Self-adaptive Sensing: due to the dynamic nature of environment application, sensor system is required to provide continuous sensing service though the full space coverage may not be necessary. However its sensing frequency together with its length of prediction cycle will be self adjusted according to system prediction accuracy. This feature becomes one of the most important issues in our data management scheme to study the energy conservation performance of our strategies.

Data Accuracy: instead of achieving long lifetime operation in wireless sensor network, DR3 embraces consideration in data accuracy in sensing. Aggressive data management scheme tends to undermine the data integrity by throwing some data which could be quite critical. Certain technique is exploited in DR3 to accommodate such scarification.

In order to successfully implement those features into our system architecture, three levels of data management are carefully introduced and investigated to meet those requirements. We developed a prototype system by adding a combination of internal group and inter-group data reduction while trying to maintain data integrity at an acceptable level. Our previous work [10] has shown that a statistical approach can help to predict the future sensing data under predefined tolerance. However using machine learning approach in data reduction is new as the data processing center can predict the data both inside group and inter-group with sending request which normally consume large amount of energy in packets dissemination. The objectives we intend to achieve are: 1) optimization in dynamic centroid selection. Dynamic centroid selection is a method to find the node whose reading is the best approximation to that of group interest.2) duty-cycle optimization. It is an approach to find the error rate of group centroid's prediction at a given schedule. We intend to figure out the trade-off between energy saving and error rate in prediction 3) opportunity coupling, an idea to identify correlated sensor group in a network thus to reduce the sensing data from a whole network perspective. A prototype of sensor network implemented with such energy efficiency scheme will also be presented.

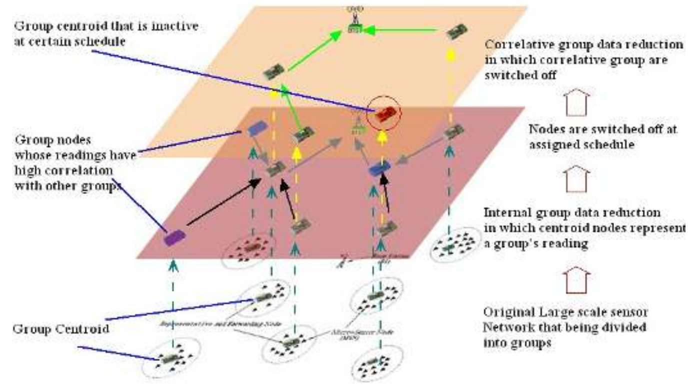

Fig. 1. The Architecture schematic of DR3

III. Design of DR3

We propose divide the sensing action into three phases. To enable only one sensor's reading while calling other sensor nodes to sleep. The statistical approach is also introduced into our system to reduce the error of sampling during resampling cycles. We also rely on the inter dependence of those group representative readings with intention to switch off groups whose readings are highly correlated to other groups. This correlation implies that one group's data can be extracted from others data with acceptable error rate. This achieves great energy saving for large sensor network, especially for monitoring application while maintaining data integrity through dynamic resample strategy described later in detail.

The performance of design will be investigated under the monitoring application so as to explore potential benefits and issues in implementation. Our design architecture can be categorized into three levels of data reduction:

A. Internal Group Data Reduction (IGDR)

In IGDR, instead of all the sensor nodes continuously transmit their sensing data to group centroid used in conventional data aggregation, only one node, also a centroid, will be active in sensing during certain sensing schedule while the remaining sensors are called to sleep. Since majority of communication packets associated with originally active nodes have been compromised due to status switching off, it can be anticipated that large amount of energy could be saved for those sleeping nodes.

Our strategy is independent to the category of data to be aggregated or compressed, and then it is not important for us to choose what kind of data we choose to apply our data reduction strategy. In ideal case, a centroid node will be fault-free one. However, in a real system, sensing at each sensor node could be irregular and affected by many factors including environment nose, obstacles, et al. Therefore, the irregularity of node performance should be carefully measured and considered. Because typically centroid node switching involves communication topology modification, which costs extra energy consumption overhead, environment irregularity should be factored in decision making process. Before going further, there are two questions need to answer firstly: (i) how to evaluate the measurement risk of the selected node; (ii) is there a monotonic relationship between the environment irregularity and centroid nomination? To solve the first question, a node's reading error to the reference will be exploited. The average of all the sensors' data in a group, the most typical reference metric in data aggregation, will be compared against each node's sensing data within study group and the one whose reading has the least square error rate in sampling period has high possibility to be chosen. This factor allows us to minimize the error of using one sensor node rather than operation of whole local group. To tackle the second question, we argue that a smoothing approach should be applied to eliminate the spontaneous environmental irregularity. Therefore, sensor node's selected frequency, denoting how many times it has been as centroid, weights into final selection. As inferred, the higher frequency, the better chance that the node is picked. To bridge gap between one centroid's reading and reference (group reading average), polynomial regression approach used in machine learning are introduced so that an approximation mapping function from one centroid node's reading onto reference can be constructed. This strategy, by its nature, guarantees the data stability and reliability in sensor network.

1) Centroid Selection : An algorithm for centroid node selection is proposed according to relationship between interested data of a group and data from individual sensor nodes within a group. Our algorithm, based on previous analysis, applies these two most critical factors, the instant measured risk and historical node nominated frequency, to determine optimal centroid node. To be more clear, measured risk is defined as the minimizing empirical risk (by least square error) R toward the reference measurement, and the node nominated frequency is the frequency of a node being a centroid node in the group. Given data from all sensor nodes within groups, those two conditions are applied repeatedly into our selection module to produce the winning centroid node for each sensing schedule according to the algorithm described in Table 1. Given the specific sensing schedule L (L= 1,2...,n) the number of sensor nodes in a group, denoted as S_i, i= 1,2..m, the number of cycle j the past frequency of winning selection for each node $f(i)$, and the sensing data of each sensor node at cycle j as D_{ij}, the centroid selection algorithm iteratively performs the following steps:

Noted that the if only least square error based selection algorithm only gives an optimal selection for that specific schedule, the past frequency $f(j)$ of nodes winning could help to select the node who has got the most approximate results to the average data in a long time frame. Now the centroid determination problem turns into an optimal utility function $F(e_i, f)$ issue:

Algorithm 1 The Selection Algorithm

1: Calculate the average Q_j for each group at cycle j by taking average of all the data sampled from each sensor node;
2: The data is store in a matrix for later comparison;
3: Calculate the least square error for each sensor node i by taking the least square error function for the entire sensing schedule;
4: Identify the node as centroid node at schedule L;
5: $C(L)$ = {j |argmin{utility function(e_j, $f(j)$) }}
6: Update the frequency counter of selected centroid node
7: $f(j)$ = $f(j)$ +1;

$$\text{Maximize} \quad \sum_{i=1}^{N} F(e_i, f)$$
$$\text{subject to} \quad \frac{\sum\limits_{i=1}^{N} e_i}{N} \leq e_t \tag{1}$$

In practice, the utility function can vary upon the application purposes. If a sensor node is capable of being the centroid nodes with multiple nominations, it will be reliable to lean toward this node in decision. Given the current measurement risk e_i at node i, and nomination frequency f_i, and Q modalities for the time L, one utility function, $F(e_i, f_i)$ can be obtained by smoothing the multiplication of those two factors over the modality k according to its sensitivity:

$$F(e_i, f_i) = \sum_{k=1}^{!} e_i \cdot f_i \tag{2}$$

For practical design purpose, the modalities can be determined based on requirement of system sensitivity to the environment irregularity.

2) Translation Model Building: This section describes the statistical method used to compare the centroid node's sensing data to the group average. The model function,whose input are centroid's sensing data $C(L)$ is built to bridge the difference between individual readings to the averages. By introducing the translation function, we expected to reduce error rate caused by the natural difference of using one node's reading to approximate the group average.

The critical problem in such approximation with finite samples is the model selection. It involves choosing the model complexity optimally for a given training sample. The training samples here refer to the centroid node's reading in the past n cycles. As a result, the number of training samples also determines the empirical risk for each model complexity. Practical model building involves two tasks which are estimation of model parameters and estimation of the prediction risk. The first task can be achieved via minimization of the empirical risk, i.e. least square fitting approach used in our work. The second can be done through data resampling following the finish of first task.

Recall that our work focuses on environment monitoring application, it makes sense not to use advanced models and high degree complexity m into our model selection. Therefore typical polynomial regression statistical approach with a model complexity from 1 to 2 will be used in our model construction. In intuition, the centroid node's reading is proportional to the group average, which makes us believe complexity of 1 and 2 are sufficient enough to keep approximation risk at a low level. It can be noticed that the empirical risk depends on the selection of model and the number of training samples. Then the performance of such estimator requires carefully selection of model parameters, model complexity degree and the number of inputs. However, the distribution of input data also affects system performance so that those raw data are needed to be preprocessed before entering into estimator. The purpose of preprocessing is to offer fast convergence of parameters in approximation process, which may increase model accuracy due to better fitting of input training samples.

The first level data reduction achieves the both the centroid selection and translation model building mission, and the latter one establish a solid basis for second level data reduction discuss in the following section.

B. Adaptive Lower Duty Cycle Data Reduction (ALDCDR)

In LDCDR, the centroid node can lower its operating duty cycle in sensing schedule, meaning centroid can even switch into sleeping status to save its energy. This idea comes from the fact that sensor's reading could form a recognizable pattern during certain period especially in environment monitoring application. Those patterns can be well approximated and used for prediction future outcome from sensor node if specific application is well studied. As a result, following the model building schedule, we introduce the data resampling phase and prediction phase. In data resample phase, we try to achieve the latest data to update the model built in training cycle. In prediction phase, the centroid node will switch off to conserve energy and the assumed sensing results are generated through predictor that has been updated. Although similar polynomial regression approach used in IGDR provides us necessity of model selection and parameter optimization, it is wise to use empirical model in specific application i.e. environment monitoring in our case, to save more energy related to computation and data storage. Based on the prediction accuracy provided by our empirical model, the sensor system will adjust its resample rate accordingly yet to ensure that a balance for accuracy and energy saving could be well maintained. The details of model have been provided in .

Fig. 2. the new developed Test-bed of Wireless Sensor Network

C. Correlated Group Date Reduction (CGDR)

In CGDR, we intend to aggressively reduce sensing data that need to be communicated through setting sensor group whose sensing reading have high correlation with other groups into sleep status. The sleep sensor group's sensing reading can be estimated using its correlated group's reading. This process obviously brings additional estimation error for the true reading but could be well worthy when large scale sensor network embeds significant amount of correlated sensor groups. The standard statistical sample correlation coefficient is used to measure the group correlation. The threshold of bias pair data correlation is the key in balancing the error and energy saving among groups. Its decision mainly depends on the duty cycle for each group and the environment stability in which wireless sensor network is deployed. We propose to use error/energy saving ratio as a metric to find out the optimal threshold of correlation under different duty-cycle setup and stability of sensing environment. The results can eventually suggest a guideline in communication protocol design for other application.

D. SYSTEM IMPLEMENTATION

The architecture will be implemented on our newly constructed test-bed with more than 100 sensor nodes which provides a realistic controllable environment for design verification and performance improvement shown in Fig. 2. The design is implanted on Berkeley TinyOS/Micaz systems, scanning light patterns will be projected onto test-bed for sensor to detect Sensors are divided into several groups according to space proximity. Those data sensed are sent to a powerful workstation where complicated and energy consumption calculation will be performed. The evaluation results (e.g. Error rate vs Energy Saving) allow further analysis to optimize the overall system.

IV. EVALUATION

A. System Evaluation

To evaluation the performance of three levels data reduction strategies, a simulation program over maximum 1000 nodes is carried out to emulate the deployment of nodes over a large area which is divided into 100 regions. The data feed is collected from WI-MN Cooperative Extension Agricultural Weather Page where they monitor the soil temperature hourly every day. The whole area under monitor is divided into over 25 groups as show in Figure 4. The soil temperature will be sampled twice per hour and 24 hours per day. Their full record of the soil temperature data in the past 10 years allows us to test extensively efficiency of our strategy. Over 7 groups of data are exploited to verify the performance of our strategies under different system setup. To overcome the continuity issues for those areas that are not covered, a weighted average data distribution method is introduced to generate the reading to cover the whole area where there are also sensors deployed in our experiment. Their sampling data, though sampled in hour basis, is treated continuously that provides flexibility of our experimental control.

B. Performance Analysis

1) Estimation Error Analysis: In this sector, we evaluate the architecture error rates on different key design parameters, which include the density of sensor nodes in each group, the effect of risk tolerance, the frequency of state interchange between re-sampling phase and prediction phase. Although those three parameters are not independent, study of their effects separately could provide us better insight of improving system performance. The methods used to calculate the error rate are described in the methodology section.

2) Error Rate vs Risk Tolerance: During such evaluation, the level of risk tolerance varies from 1 percent to 90 percent at a step of 3 percent while other two key parameters, the density of sensor nodes in each group and number of groups are set to be constants respectively. And to avoid the possible problems from the length of training phase, it is also set to be a fixed value 49 unit cycles at the beginning of each test. Since we attempt to achieve aggressive energy saving, the minimum length of prediction phase is set to be 1. As described in methodology section, we divide the region regularly into 100 groups across the whole area 1500m $\times$ 1500m. The sensor nodes are deployed into those groups according to the density setup initially. According to our experiments, we do not observe a notable difference for the positions of sensor nodes in each group.

All sources vary their intensity according to the data provided trying to mimic the environment hour by hour. The total cycles of evaluation is about 9000, approximately more than 1 year reading, which we believe significantly large enough to reduce the unsystematic risk caused by limited sample size. At this moment, we haven't considered the effect of switching off correlated sensor groups, as IGDR and ALD-CDR level consideration have focus on accuracy of prediction and translation models created. To prevent unexpected error at the initial state, we start with the training schedule right from the beginning, during which both the translation model and predictions continue to feed the inputs, sensor nodes' sensing at each cycle, to construct and train their model individually.

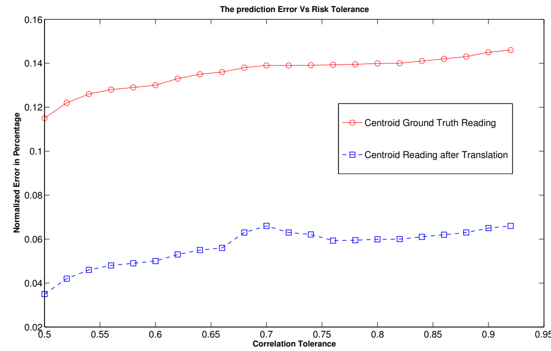

Fig. 3. The influence of risk tolerance

Figure 3 shows the result of estimated error rate with the change of risk tolerance for estimator w/o using the translation model. As suggested from the simulation results, the prediction error increases proportional to the increase of risk tolerance. And it can be told from the projection that the risk won't be equal to zero if risk tolerance is extended to zero. This is the systematic error introduced by our methodology nature in which a minimum of prediction step is set. The average prediction risk for tolerance at 90 percent is around 14.2 percent, a highly affordable error for a sensor system targeting for long term usage. At such a low level of prediction error, it will be more flexible for system to choose other appropriate parameters.

We also measure average system prediction error under the assistance of our translation model. The prediction error assisted with the translation model we build is about two times less than the error without using the translation model, a number more benign than what we expect. And the gap of error in both methods shows low correlation to risk tolerance we choose. This can be explained that translation model takes effect to reduce error from location mismatch from the weighted average center of all reading in a group since relationship between centroid's reading and the real group average reading is linear within a certain space.

3) Energy Saving Vs Risk Tolerance: The risk tolerance determines the switching rate of the prediction cycle which is the key effect to energy saving by our system. It is crucial to study risk tolerance's impacts onto the energy saving. During our simulation experiments, we steadily increased our system risk tolerance, then measured the average energy saving by sensor system. The results are shown in Figure 6. From intuitive perspectives, the energy saving increases as the risk tolerance becomes higher due to the fact that the prediction cycles have been extended, thus allow the system to get into energy saving mode more frequently than low risk tolerance. It is not surprised that even at 10 percent risk tolerance level, energy saving can reach to 68 percent because the main energy saving has been achieved through the IGDR approach. We note here that the energy saving's performance can only improve the system energy saving level to certain extend since the

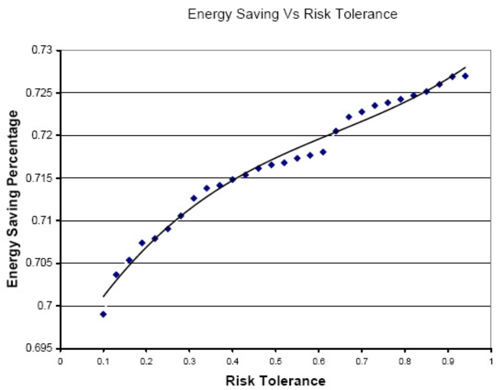

Fig. 4. The influence of risk tolerance on the average energy saving

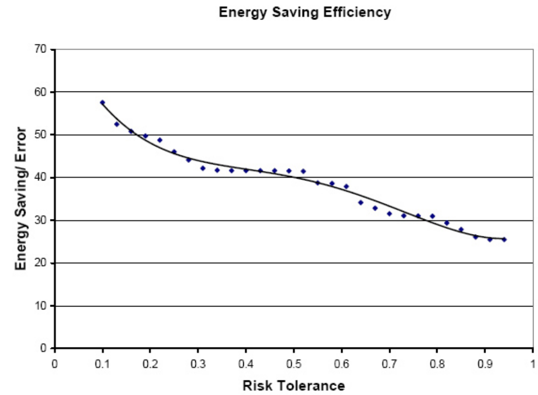

Fig. 5. Energy saving efficiency related to the risk tolerance

existence of training cycles and re-sampling cycles used to build the model. It hints that more aggressive energy saving is possible by reducing those cycle length but apparently we will have that at the loss of accuracy and roughness of the prediction model.

4) Energy Saving Efficiency Vs Risk Tolerance: It will be difficult to judge whether the energy saving at a risk tolerance level is acceptable or not without any comparison standard. Therefore an indicator called energy saving efficiency is defined as

$$ESE = \frac{Energy\_Saving}{Risk\_Tolerance} = \frac{E_s}{\tau_{risk}} \quad (3)$$

E_s is the energy saving percentage by referring to TAG approach and τ_{risk} to be the risk tolerance of system. It is shown in Fig. 5.

From the result we can see that even energy saving at low risk tolerance, its efficiency is higher than that for higher risk tolerance. At the middle range, the energy saving efficiency remains at a stable rate, which is quite interesting for practical application. It indicates system can be tuned at a range of risk tolerance for different energy saving while keeping the stable energy saving efficiency. This feature provides flexibility for users to adjust the system under different application requirement domains.

V. CONCLUSION

In the paper, a system framework offering 3 degree of data reduction is proposed in an attempt to minimize energy consumption in wireless sensor network especially for long life time application. The lower level of data reduction scheme serves as the basis for higher level one with consideration of their integration issues. Our statistical local error minimization approach ensures centroid nodes selected in each group can best represent oriented data point (i.e. data average) without giving up data accuracy which is quite important in data analysis. Furthermore, an empirical data model is created to lower computation complexity and groups whose data have high correlated coefficient will be reorganized into divisions so that more aggressive data reduction can be achieved. Our system implementation suggests a possible practice to extend to a much larger scale of wireless sensor network. The potential benefits for a system to exchange over 70 percent energy saving with prediction error rate about 2 percent to 12 percent are demonstrated through our extensive simulation by taking advantage of those readily available soil temperature data. As future work, we will need to consider the routing and connectivity issues for three level data reduction scheme integration to improve the network's robustness.

VI. ACKNOWLEDGEMENT

This work is partially supported by Singapore-MIT IDC IDD61000102a Grant, SUTD SRG ISTD 2010 002 Grant and Binghamton University academic program and faculty development fund.

REFERENCES

[1] S. Guo, S. M. Kim, T. Zhu, Y. Gu, and T. He, "Correlated Flooding in Low-Duty-Cycle Wireless Sensor Networks," in *ICNP*, 2011.

[2] T. Zhu, Y. Gu, T. He, and Z. Zhang, "Achieving Long-Term Operation with a Capacitor-Driven Energy Storage and Sharing Network," in *ACM Transactions on Sensor Networks (TOSN)*, Vol. 8, No. 3, Aug. 2012.

[3] T. Zhu, A. Mohaisen, Y. Ping, and D. Towsley, "DEOS: Dynamic Energy-Oriented Scheduling for Sustainable Wireless Sensor Networks," in *INFOCOM*, 2012.

[4] T. Zhu, Z. Zhong, Y. Gu, T. He, and Z. Zhang, "Feedback Control-based Energy Management for Ubiquitous Sensor Networks," in *IEICE Transactions on Communications*, Vol. E93-B, No.11, pages 2846-2854, Nov. 2010.

[5] T. Zhu, Y. Gu, T. He and Z.-L. Zhang, "eshare: A capacitor-driven energy storage and sharing network for long-term operation," in *SenSys*, 2010.

[6] I. Solis and K. Obraczka, "in-network aggregation trade-offs for data collection in wireless sensor networks," *International Journal of Sensor Networks*, 2006.

[7] M. Lee and V. Wong, "An energy-aware spanning tree algorithm for data aggregation in wireless sensor networks," Victoria, BC, Canada, 2005//, pp. 300 – 3, energy-aware spanning tree algorithm;data aggregation;wireless sensor networks;distributed protocol;directed diffusion;network routing;.

[8] A. Cerpa, J. Elson, D. Estrin, L. Girod, M. Hamilton, and J. Zhao, "Habitat Monitoring: Application Driver for Wireless Communications Technology," in *Proc. of the 2001 ACM SIGCOMM Workshop on Data Communications in Latin America and the Caribbean*, April 2001.

[9] S. Madden, M. Franklin, J. Hellerstein, and W. Hong, "TAG: A Tiny Aggregation Service for Ad-Hoc Sensor Networks," in *Operating Systems Design and Implementation*, December 2002.

[10] Q. Zhang, Y. Gu, L. Gu, Q. Cao, and T. He, "Collaborative scheduling in highly dynamic environments using error inference," in *7th International Conference on Mobile Ad-hoc and Sensor Networks*, 2011.