Data centric modeling of environmental sensor networks

Ram Dantu*, Kaja Abbas†, Marty O'Neill II† and Armin Mikler†
*Network Security Lab
†Network Research Lab
Department of Computer Science and Engineering
University of North Texas
Denton, TX 76203, USA
Email: rdantu@cs.unt.edu, kaja@cs.unt.edu, me@martyo.com, mikler@cs.unt.edu

Abstract

Meteorological and hydrological sensors deployed over several hundred kilometers of geographical area comprise an environmental sensor network. Large amounts of data need to be processed in minimal time and transmitted over the available low speed and low bandwidth links. This paper describes algorithms for optimal data collection and data fusion. An inductive model using exponential back-off policy is used to collect optimal amount of data. The data measurements for temperature, pH and specific conductance collected for a year from the sensors deployed at Lake Lewisville are used to test the inductive model. Energy savings of 90% are achieved even with 1% of degree of tolerance. The problem of data fusion is addressed by the introduction of a novel concept of a super-sensor, based on self-organization and collaboration among sensors. A histogram application is described that uses recursive doubling for global collaboration between sensors. The performance of the networked super-sensor in comparison to a centralized polling approach is analyzed for optimality on two different geographical areas.

I. Introduction

The efforts to predict the dynamics of large watersheds and landscape ecosystems have led to the deployment of meteorological and hydrological sensors and their integration to remote sensing data in order to span vast areas. To gain insight into the intricate functional dependencies that occur in complex ecosystems, environmental and biological sensors have been deployed to acquire the real-time vital ecosystem information [1][2][6]. The complexity of the sensor network is manifested in the number and the variety of sensors, and the extent of the geographical area.

The objective of this paper is the design, implementation, and analysis of data centric algorithms for the collaborative retrieval and processing of information in large environmental sensor networks.

A. Toxicity in the watershed

The quality of the environment is measured using chemical analyses to find the pertinent analytes. Exposure may be defined as the magnitude, duration, and frequency with which organisms interact with bioavailable toxicants. Living organisms are needed to indicate the quality of the environment. The bivalves used in the biomonitoring scheme developed at UNT and deployed as part of an EPA sponsored EMPACT grant [6][22] are individual sensors surveying the different variables. The geographic location of the sensors in a watershed permits more detailed analysis (GIS, NPDES permits, etc.) of the watershed and can lead to implementation of best management practices reducing or eliminating the sources of the toxicants.

B. Why sensor networks in a watershed

Predictions for ecosystems dynamics are multivariate in nature and require event correlation of a large number of different types of environmental and biological sensors. The identification of environmental events and their ecological effects requires a statistical analysis of the acquired

datasets. Data mining techniques [9][10][21] on these datasets are useful to correlate ecological events as captured by the sensors. The deployment of a sensor infrastructure in a watershed may contain hundreds or thousands of sensor nodes over an area of several hundred square kilometers. Networked sensors embedded in the watershed will reveal previously unobservable phenomena.

C. Problem description

Real time acquisition of sensor data from a watershed and the assembly of widely distributed and disparate sensor information into a composite image will prove useful to the environmental scientist. Different algorithms complete the same tasks with different efficiencies, prompting the use of varied algorithms for fuller optimization of the system [11]. The energy of the sensors is expended in data collection and dissipation of the processed data. The time intervals for data retrieval should be alterable in real time by each individual sensor, depending on the dynamics of the incoming data.

Table I illustrates the data handling requirements for the sensors [19]. Large amounts of data need to be processed in minimal time and transmitted over the available low speed links. The amount of storage needed for a sensor network that is continually sampling grows exponentially with the required sensing resolution. Real time parallel data processing by the sensors is vital for environmental protection applications. A combination of power optimization algorithms with respect to data collection, and data dissipation and fusion are the desired requisites of a sensor network.

Table I. Sensor data requirements.

Sampling frequency	1 Hz
Bit rate	7 bits/sample
Amount of sensing data from one hour per sensor	4 Kb
Amount of sensing data for one hour for a geographic area of few tens of kilometers	400 Mb – 4 Gb

D. Outline

Section II gives an overview of the recent work pertaining to data dynamics in sensor networks. Section III illustrates the data collection algorithm and the performance analysis using the water quality data collected from Lake Lewisville. Section IV describes the concept of super-sensors and data fusion algorithm for the generation of a histogram. The paper is concluded in Section V.

II. Background and prior work

Wireless sensor networks are constrained by the low energy constraints of the individual sensors. The efficient use of energy is a prime criterion for the longevity of the sensors. A comprehensive review of recent research in sensor networks is provided by [18]. The spatial configuration of the sensors is not predetermined, so the sensors must dynamically fit into the network, sense the requisite data, process this data, and communicate any results with other sensors.

Individual sensors are powered by small batteries and/or small solar panels that provide sufficient energy for computation on a local microprocessor, capabilities of receiving signals from the global positioning system (GPS) for geographical location, storage of sensor data in local memory, and communication with other sensor-nodes in the proximity. The use of the radio should be optimized by turning off the radio transmission when there is no data of interest [15]. The sensors should operate at a minimal power rating until the event of interest occurs to wake up the sensors [7]. The data should be processed and compressed before transmission for the efficient use of the limited resources [20].

Collaborative sensor processing is needed to conserve power, memory, and bandwidth. A technique called directed diffusion was introduced in [12] for collaborative processing. Directed diffusion uses data-centric routing instead of address-based routing. In this approach, a sink sensor requests data by sending interests for named data [8][12][4]. Data matching an interest are then drawn from source sensors towards the sink sensor. Intermediate sensors can cache or transform data and may direct interests based on previously cached data.

A data dissemination method based on data-centric storage proposed a geographic routing algorithm [16]. An architecture for monitoring sensor networks uses an aggregated view of the system state [24]. Another data-centric architecture considers a sub-optimal data aggregation tree and illustrates that gains are largest when the sources are relatively close to each other and far from the sink [13]. The energy gain in the use of short-range messages over long range messages is elucidated in [3].

III. Data collection

The power consumption by the sensors for data retrieval should be optimal to extend the life of the sensors. Instead of centralized algorithms, localized algorithms are the key for the optimal use of the energy expended by sensors [14]. An inductive data collection model is proposed in this paper, that represents the senor data at an acceptable level of approximation. The sensors are let to sleep at periodic intervals and conserve energy expended in data collection compared to active continual monitoring.

A. Inductive model

The power of an individual sensor is limited by the battery energy and/or energy generated from the embedded solar panels. Sensor power is the key element for the longevity of the sensors. The sensors are defined by two states: *sleep* and *wakeup*. During the wakeup state, the sensor is active and retrieving the corresponding data from its reachable region. During the sleep state, the sensor is let to be passive with null energy spent in retrieval of data. During this sleep state, the data carried by the sensor refers to the last recorded data. The difference between the sensor data and the current real data leads to an error. If this error is uncritical and is within the acceptable levels of approximation, then sensor energy is saved during the sleep state.

The criticality of the data carried by the sensors is dependent on a given environmental application. Some environmental applications may need exact current data in the sensor observation regions which pushes the sensors to continually monitor the network at all times. There are other applications that shall need data within acceptable error rates. This approximation feature is exploited to let the sensors sleep during lesser fluctuations in recorded data.

Degree of tolerance (Δ) is the parameter used by the sensors in the decision-making process of changing states from sleep to wakeup and vice versa. The exponential back-off scheme is used to change the sleep time of the sensor nodes with an additional feature of letting the sleep time alter in either way, that is, increase or decrease. If the percentage difference in two consecutive recorded values of the sensor is not more than Δ, then the sensor is let to sleep twice its last sleep time. On the other hand, the sensor is let to sleep half its last sleep time. Once the sleep time is over, the sensor is changed to wakeup state to record the next observation value. The state changes from sleep to wakeup and vice versa is a continual process and terminates only when the sensor runs out of power. The sensor state diagram [Figure 1] illustrates the conditional checks executed by the sensors during the change of states.

The environmental data is monitored by the sensors for a time range T. During the sleep time, the value stored in the sensors deviates from the real value resulting in an error. δ is the error percentage at any given instant between the current recorded value in the sensor and the real value in the environment. Error rate is the average deviation percentage between the recorded and real values over a range of time.

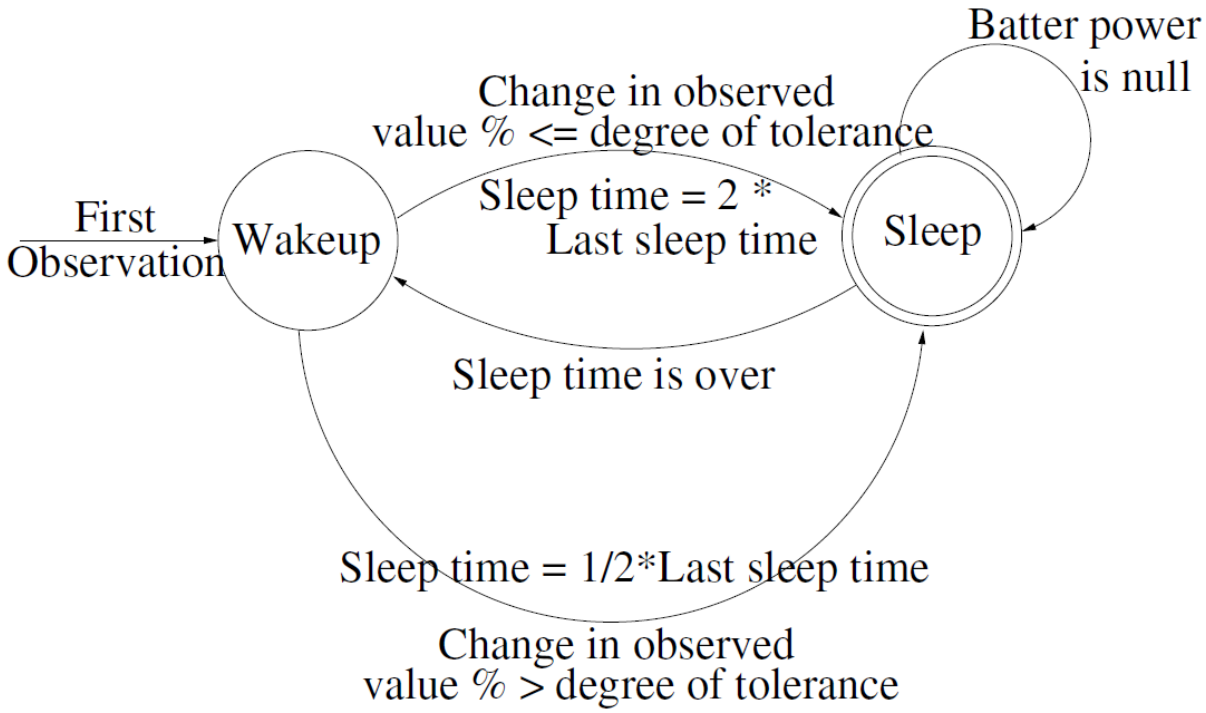

Figure 1. Sensor state diagram.

$$Error\ rate(\%) = \frac{\sum_{time=0}^{time=T} |\delta|}{T} \qquad (1)$$

B. Algorithm

The following algorithm is used by the sensors in the decision-making process of changing states from *wakeup* to *sleep* and vice versa.

1) state = Wakeup
2) Record the current value
3) diff = | current value – last recorded value |
4) diff % = (diff / last recorded value) * 100
5) state = Sleep
6) if diff % <= degree of tolerance
   sleep for twice the last sleep time
   else
   sleep for half of the last sleep time
7) go to step 1

C. Performance analysis

The environmental data has been collected from the Ecoplex project database [6]. The data pertains to the water quality characteristics at Lake Lewisville (Table II) for one year. The sensors at the lake monitor the water quality parameters at an interval of every five minutes. Temperature, pH, and specific conductance are the three parameters used for analysis of the inductive model for data collection.

Table II. Water quality data from Lake Lewisville.

Start date	9/1/2002
End date	8/31/2003
Sampling interval time	5 minutes
Observed parameters	Temperature pH Specific conductance

Figure 2 illustrates the efficient energy savings for varying degrees of tolerance Δ. An exponential growth in energy savings of 90% is observed up to Δ of 1% for all the three data parameters. For higher Δ of more than 1%, the energy savings monotonically increase towards a saturation point of close to 100% at Δ of 5%. Figure 3 describes the error rate induced in the data with respect to Δ. The error rate induced in temperature data is the minimal with a linear growth rate. Specific conductance shows similar characteristics of temperature. pH has a high initial growth rate with error rates compared to Δ. The pH data values of the lake water exhibit higher fluctuations while temperature data values show a steady stream with lower fluctuations. All three data parameters tend to a stable state at Δ of 4%.

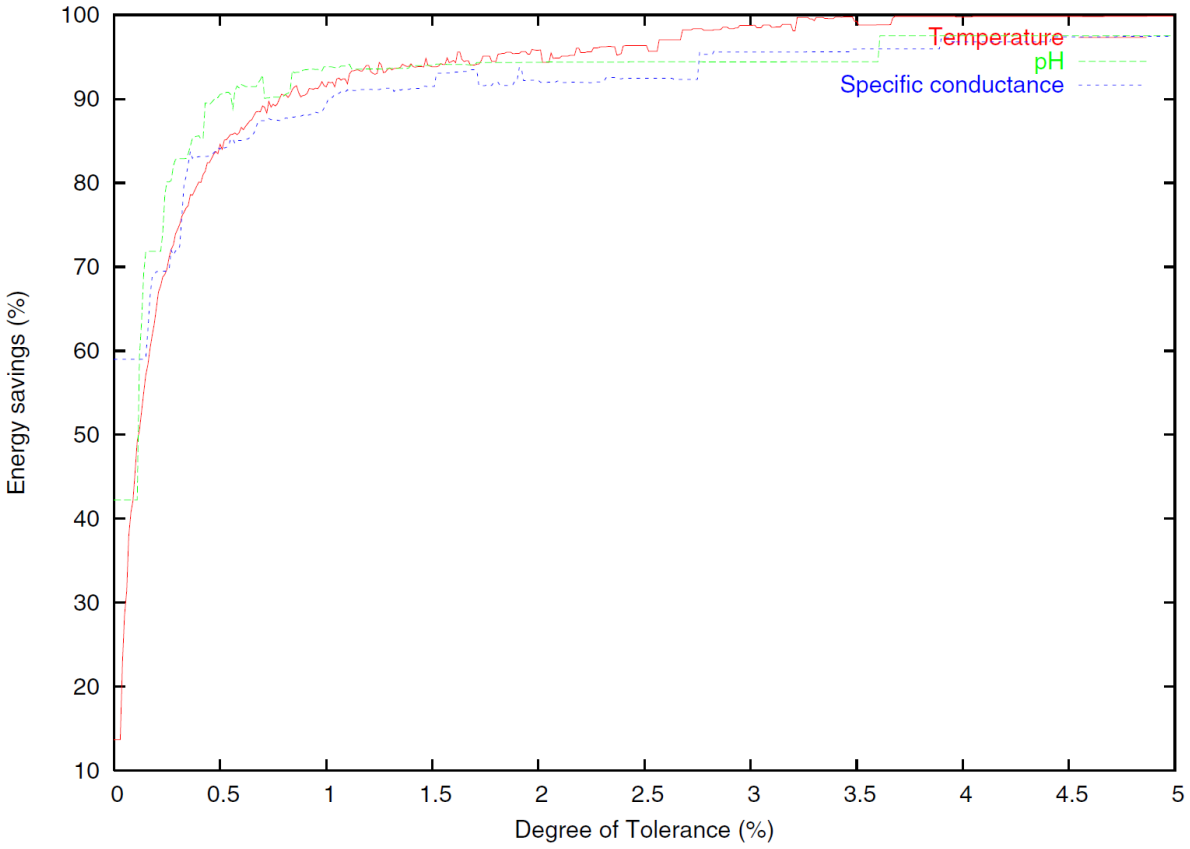

Figure 2. Energy savings performance.

The acceptable limits of error rate are key to the optimal choice of Δ. In case of temperature and specific conductance, due to the linear characteristics of error rate to Δ and an exponential growth rate of up to 1% in energy savings, Δ of 1% is an optimal choice, provided the observed error rate is acceptable. In case of pH, although the energy savings are similar to temperature and specific conductance, the observed error rate is higher. For a degree of tolerance of 0.25%, it results in an error rate of 5%. If this is within the acceptable limits of error, 0.25% for Δ shall be an optimal choice for pH for corresponding energy savings of 80%.

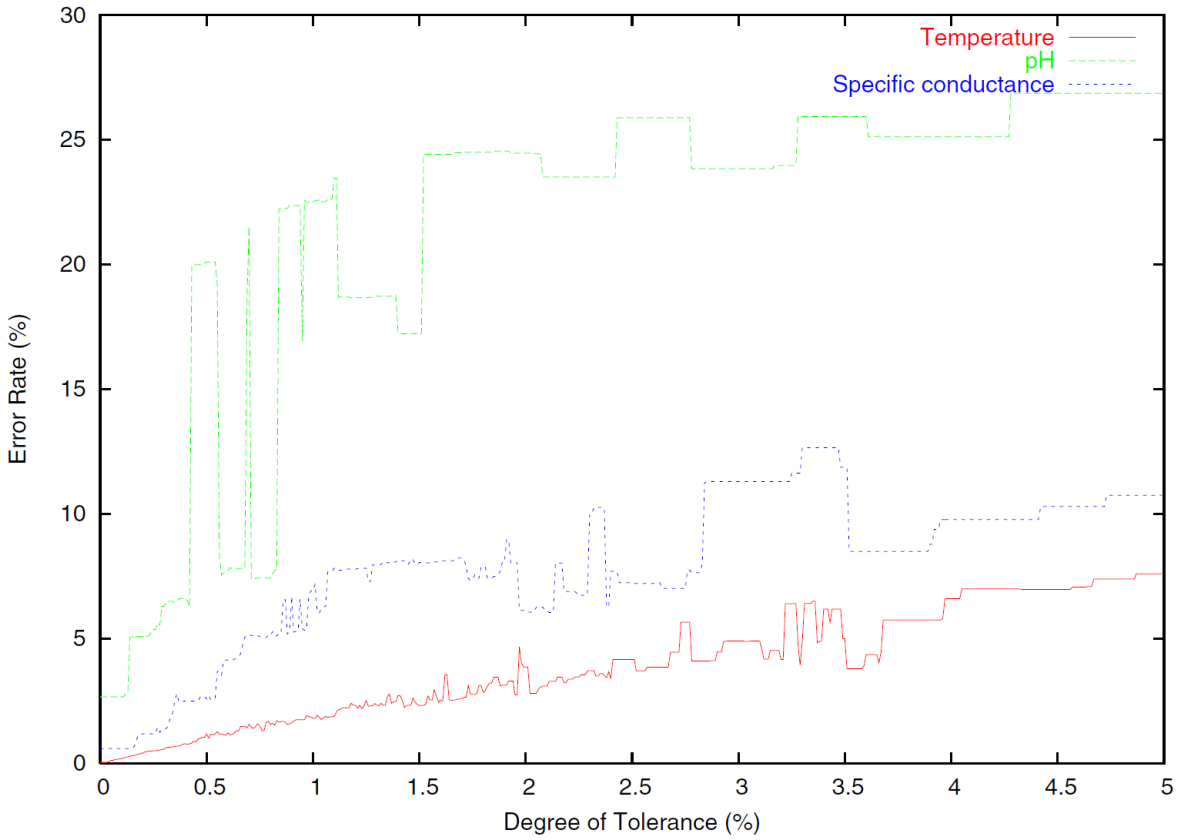

Figure 3. Error rate characteristics.

IV. Data fusion

This section addresses the design of real-time monitoring systems, coupled with feedback systems, that provide water quantity and quality information for improved environmental decision-making concerning watersheds, streams, and lakes. The problem is that of real-time acquisition of sensor data and the assembly of widely distributed and/or disparate sensor information into a composite image that can be interpreted by environmental scientists. The sensor nodes can be equipped to internally process the information and broadcast only the requisite information thereby saving on network bandwidth [17][15][23]. A collaborative scheme based on direct diffusion and recursive doubling for creating a composite view of sensed data is presented in this paper.

A. Image composition

In the Ecoplex project [6], the widespread use of phosphorus-containing fertilizers on P-rich soils in urban and suburban watersheds and the discharge of this excess phosphorus into storm water were studied. Runoff from this storm water has impaired the water quality in the North Texas urban lakes. In-stream sensors were used to collect real-time data including clam gape, pH, temperature, dissolved oxygen, conductivity, wind speed, air temperature, and rainfall. The sensors in this network form a super-sensor by cooperating with each other to obtain a single image. Figure 4 illustrates the logical concept of a super-sensor comprised of 14 cooperating sensors over a geographical area.

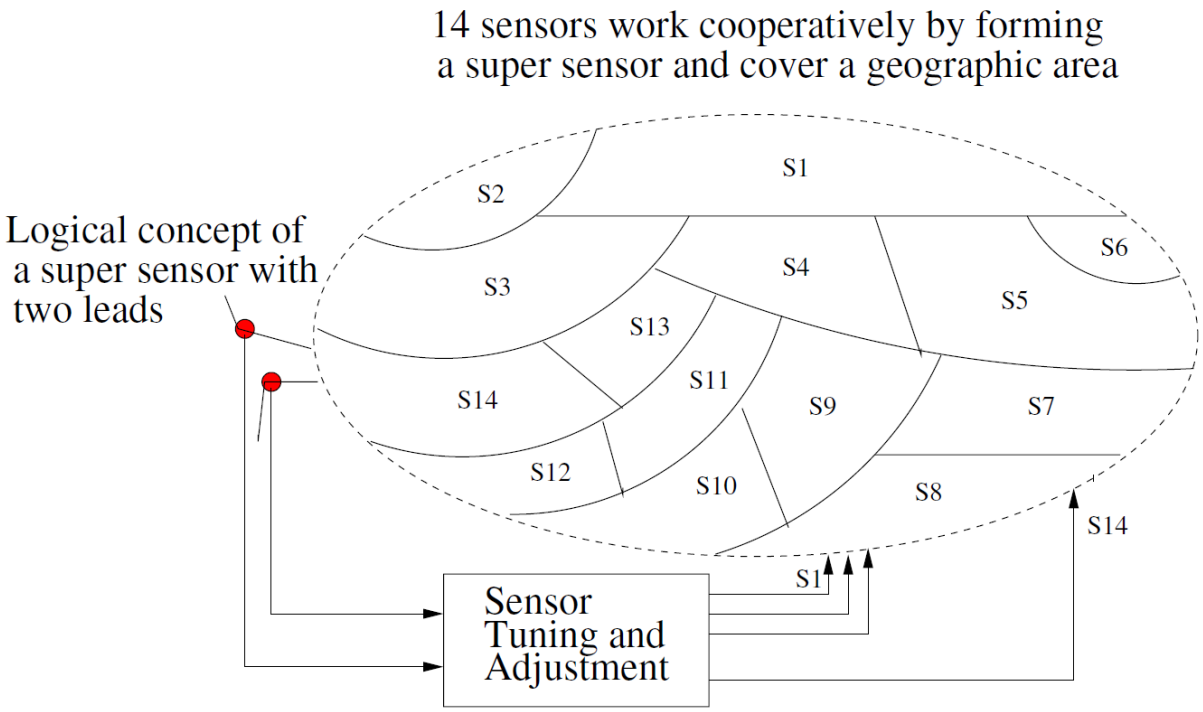

Figure 4. Feedback control for sensing a geographic area.

B. Composite image generation

A composite image consists of an array of sensed data, based on spatial and/or temporal measurements. The image may contain smoothed, filtered, and summarized values of temperature, pH, and nitrate over a geographic area. The image is a composite view of multiple sensors captured by the super-sensor.

A pivot sensor is initially selected. This selection is communicated to all of the sensor members of the super-sensor. The following steps describe the procedure for capturing the image from the super-sensor:

1) Sensors collect the local fragment of the data image and compute the local results.
2) Sensors transfer intermediate results to their designated neighbors.
3) Intermediate results are processed.
4) Continue steps 2 and 3 until the result reaches the pivot sensor.
5) The pivot sensor processes these intermediate results and computes a final result.
6) Sensing parameters are adjusted and fed back to all members of the super-sensor.
7) The results are used to select an optimal pivot and communicate it to all of the sensors.
8) Return to step 1.

At any given moment, the pivot will have the latest final result. Two different types of measurements, such as temperature and pH, may use different pivots in the network. In the next section an algorithm for histogramming temperature from different sensors is described.

C. Analytic model for histogramming

In a super-sensor system, each sensor captures an image under its purview, and exchanges the contents of the image with its neighbors. There are N sensors in the neighborhood network. If a set of $M * M$ measurements forming an image of a geographical area are divided into N strips, these strips will have M^2/N measurements each. Each strip is allocated to a sensor that calculates a partial histogram of the strip assigned to it. These partial histograms are merged using recursive doubling [5].

Suppose that there are B values or bins for any given measurement in the image. Initially, sensors $P_{2l+1} : l = 0, 1, 2, \ldots (N/2 - 1)$ merge the $B/2$ least significant bins of the partial histograms contained in their own memory, as well as those of their right neighbors. Similarly, sensors P_{2l+2} merge the $B/2$ most significant bins located in their own memory as well those of their left neighbors. At the end of these operations, sensors P_{2l+1} hold the least significant halves of the merged partial histograms, while their right neighbors' sensors P_{2l+2} hold the most significant halves [Figure 5].

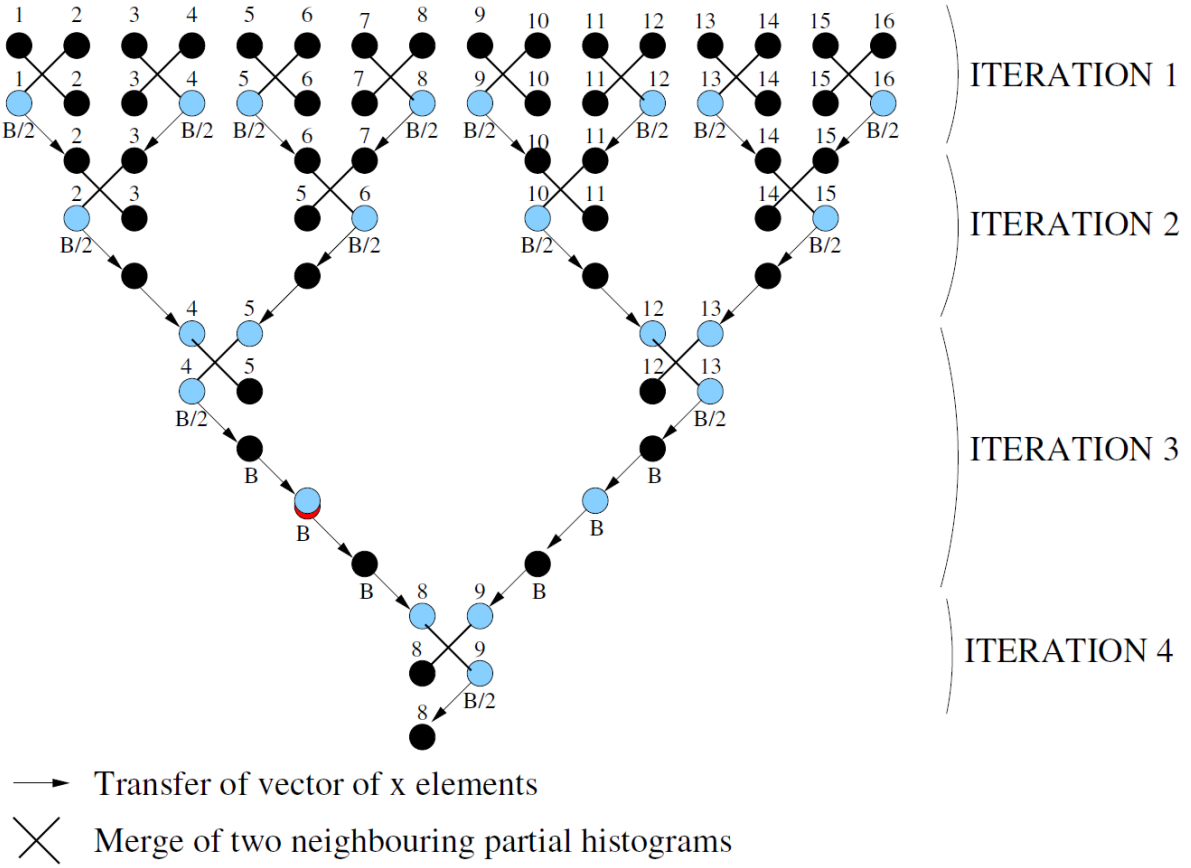

Figure 5. An example of the distributed merge algorithm for histogram calculation using sixteen sensors (an example for composite image generation).

Proceeding to the next level, sensors $P_{4l+1} : l = 0, 1, 2, \ldots (N/4 - 1)$ transfer the $B/2$ least significant bins of their merged histogram to sensors P_{4l+2}, and similarly sensors P_{4l+4} transfer the $B/2$ most significant bins to sensors P_{4l+3}. At this point, nodes P_{4l+2} and P_{4l+3} contain partial B bin histograms and the process is repeated. The final completed histogram is to be found in sensor $P_{N/2}$. The algorithm merges the partial histograms in a tree structure of sensors embedded on the sensor network.

As the process continues, partial histograms are located in sensors that are progressively further away from their immediate neighbors. Their merging requires the transfer of data between distant sensors. This communication is accomplished directly through the intermediary links with the neighbors.

The described model uses the form of a *bucket brigade* to transfer long vectors between distant sensors. In order to transfer a B-bin vector from sensor P_i to sensor P_j, the intervening sensors form a pipeline through which the B-vector is transferred in $O(j - i + B)$ steps. Each sensor in the pipeline moves data from its left neighbor to the right neighbor, and the transfer of data between two sensors i and j requires $B + (j - i)/2 - 1$ steps.

The histogram merging algorithm is carried out in *log (N)* iterations. Each of the iterations consists of a partial merging step requiring *B/2* additions and *B/2* transfers, together with *B/2* transfers needed to locate the merged histogram in the appropriate sensor. Iterations *2* to *log (N – 1)* require the pipelining of intermediate sensors, and the m^{th} iteration requires $B + 2^{m-2} - 1$ transfers. τ_α is the time required to perform a single addition, and τ_{ttr} is the time required for a single transfer from the left neighbor to the right neighbor. T_{MRG} is the time required for the histogram merging algorithm. The time spent to merge each of the $B/2$ bins of two partial neighboring histograms is $\tau_{ttr} + \tau_\alpha$ which corresponds to the transfer of one bin from a neighboring sensor and the addition of thecorresponding bin in the local sensor.

$$T_{MRG} = \sum_{m-1}^{\log N} (\tau_\alpha + 2\tau_{tr})\frac{B}{2} + \tau_{tr} \sum_{m-2}^{\log N - 1} B + 2^{m-2} - 1 \quad (2)$$

Given an *M * M* area, T_{HIST} is the time required to obtain the *N* partial histograms on a sensor network consisting of *N* sensors.

$$T_{HIST} = \frac{M^2}{N}\tau_\alpha \qquad (3)$$

T_{SUPER} is the total time required for the cooperative sensor processing.

$$T_{SUPER} = T_{HIST} + T_{MRG} = \qquad (4)$$

$$\tau_\alpha \left(\frac{M^2}{N} + \frac{B}{2}\log N \right) + \tau_{tr} \left[\frac{N}{4} + (2B-1)\log N - 2B + 1 \right]$$

D. Performance model

Polling is a centralized approach of direct data retrieval from the sensors. A comparison between the polling method and distributed histogramming is discussed in this section. $T_{CENTRAL}$ is the total time required by polling to obtain the histogram of an *M * M* area by use of a *Centralized Sensor*.

$$T_{CENTRAL} = M^2 \cdot \tau_\alpha \qquad (5)$$

The efficiency can be measured by the speedup (*S*) of the distributed cooperative sensor processing with respect to the centralized polling methodology.

$$S = \frac{T_{CENTRAL}}{T_{SUPER}} = \qquad (6)$$

$$\frac{M^2\tau_\alpha}{\tau_\alpha \left[\frac{M^2}{N} + \frac{B}{2}\log N\right] + \tau_{tr}\left[\frac{N}{4} + (2B-1)\log N - 2B + 1\right]}$$

E. Simulation results

Figure 6 shows the communication overhead using a super-sensor as compared to the communication overhead using a centralized polling method for two differently sized geographical areas. Even though the communication overhead is low for smaller sets of sensors, the savings using a super-sensor remains significant even for larger sets of sensors. For 64 sensors over 50 km^2, the super-sensor communication is only 7% of that of the centralized polling method. When the number of sensors is increased, the communication overhead also increases. The percentage of relative overhead for 256 sensors is 30% over 50 km^2 and is 61% over 100 km^2. The larger geographical area has a steeper slope and greater overhead due to the higher amount of energy required to transmit data over larger distances,

and larger amounts of data to be processed. The savings on communication overhead by the super-sensor is derived from the way data merges as it converges onto the sink node, resulting in efficient utilization of the available network resources.

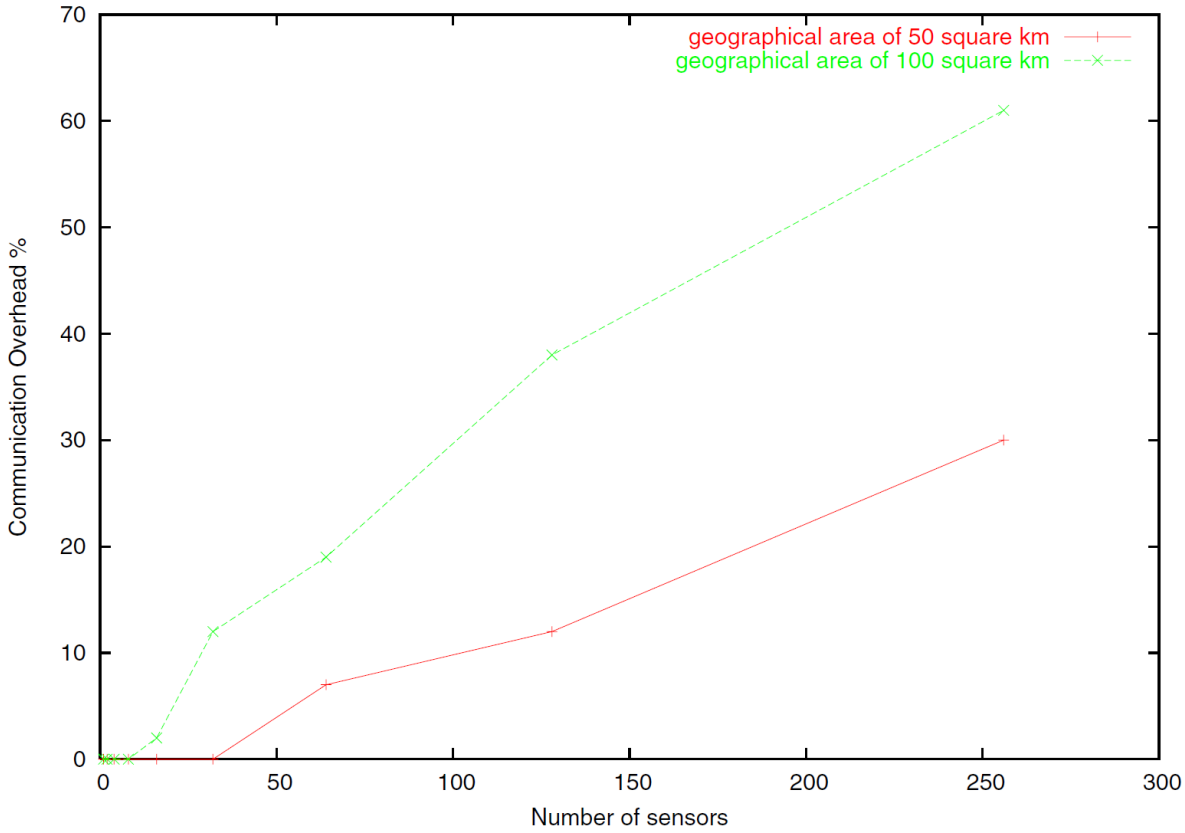

Figure 6. Communication overhead against number of sensors for histogram calculation (1-256 sensors).

V. Conclusion

Environmental sensor networks encompass a set of micro sensors deployed over a geographical area. They are useful in the collection and processing of environmental data from terrains that are otherwise inaccessible. Due to the energy constraints of the sensors, optimal utilization of sensor power is a prime criterion. Distributed data centric modeling is an efficient method of processing and collecting environmental sensor data over large geographic areas. An autonomous temporal data collection methodology, combined with a spatially optimal data fusion wave computation, has been presented in this paper.

The environmental data collected from Lake Lewisville for a year at a sampling interval of every five minutes has been used to analyze the inductive model. Energy savings of 90% at an error rate of 1% is observed for temperature data while an energy savings of 80% can be achieved at an error rate of 5% for the pH data. The data fusion algorithm uses the super-sensor concept and recursive doubling for global collaboration between the sensors. The bucket brigade methodology is exploited for the distributed cooperative processing between the sensors to compose the complete histogram image. The communication overhead for two geographical areas of 50 km^2 and 100 km^2 has been analyzed.

The technique of doubling and halving the sleep times in the data collection algorithm can be improved by tuning to varying degrees of increase or decrease depending on the data dynamics. The cooperative data fusion algorithm can be enhanced in the future by pipelining the transfer of the partial histograms, triggered immediately after the neighboring histograms are merged.

Acknowledgements

The authors would like to thank John Mayes and David Hunter for the water quality data collected from Lake Lewisville [6].

References

1. J. Allen, W. Waller, M. Acevedo, E. Morgan, K. Dickson, and J. Kennedy, *A minimally-invasive technique to monitor valve movement behavior in bivalves*, Environmental Technology, 17:501-507, 1996.
2. J. Allen, W. Waller, J. Kennedy, K. Dickson, M. Acevedo, and L. Ammann, *Real-Time Whole Organisms Biomonitoring - Deployment, Status, and Future*, pp: 187-192, AWRA, Annual Spring Specialty Conference Proceedings, 2001.
3. M. Bhardwaj, T. Garnett, and A. Chandrakasan, *Upper Bounds on the Lifetime of Sensor Networks*, IEEE Intl Conf on Comm, vol. 3, pp. 785-90, 2001.
4. C. Borcea, D. Iyer, P. Kang, A. Saxena, and L. Iftode, *Cooperative Computing for Distributed Embedded Systems*, Proceedings of 22nd International Conference on Distributed Computing Systems, 2002.
5. R.V. Dantu, *A computer vision system for VLSI wafer probing*, Ph.D. Thesis, Concordia University, Montreal, Canada, April 1990.
6. *Ecoplex*, http://www.ecoplex.unt.edu, Web Site.
7. J. Elson and K. Romer, *Wireless Sensor Networks: A New Regime for Time Synchronization*, Proceedings of the First Workshop on Hot Topics in Networks, October 28-29, 2002.
8. D. Estrin, *Embedding the Internet: This Century Challenges*, IPAM Workshop on Massively Distributed Self-organizing Networks, May 2002.
9. U. Fayyad, G. G. Grinstein, and A. Wierse, *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann Publishers, August 2001.
10. D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, September 2000.
11. J. Heidemann, F. Silva, and D. Estrin, *Matching Data Dissemination Algorithms to Application Requirements*, Proceedings of the First ACM Conference on Embedded Networked Sensor Systems, 2003.
12. C. Intangonwiwat, R. Govindan, and D. Estrin, *Directed diffusion: A scalable and robust communication paradigm for sensor networks*, Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking, August 1999, pp. 174-185.
13. B. Krishanamachari, D. Estrin, and S. Wicker, *Modelling Data-Centric Routing in Wireless Sensor Networks*, USC Computer Engineering Technical Report CENG 02-14, 2002.
14. S. Meguerdichian, S. Slijepcevic, V. Karayan and M. Potkonjak, *Sensor networks and energy management: Localized algorithms in wireless ad-hoc networks*, Proceedings of the 2001 ACM International Symposium on Mobile ad hoc networking & computing, Long Beach, CA, USA, October 2001.
15. G. Pottie and W. Kaiser, *Embedding the Internet: wireless integrated network sensors*, Communications of the ACM, vol. 43, no. 5, pp. 51-58, May 2000.
16. S. Ratnasamy, D. Estrin, R. Govindan, B. Karp, S. Shenker, L. Yin, and F. Yu, *Data-centric storage in Sensornets*, First Workshop on Sensor Networks and Applications, September 28, 2002.
17. C. Intanagonwiwat, R. Govindan, and D. Estrin, *Directed diffusion: A scalable and robust communication paradigm for sensor networks*, Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking, Boston, MA, USA, August 2000.
18. M. A. M. Vieira, C. N. Coelho, D. C da Silva, and J. M. da Mata, *Survey on Wireless Sensor Network Devices, Emerging Technologies and Factory Automation*, 2003 Proceedings. September 2003.
19. A. Wang and A. Chandrakasan, *Energy efficient DSPs for Wireless Sensor Networks*, July 2002.
20. H. Wang, D. Estring, and L. Girod, *Preprocessing in a Tiered Sensor Network for Habitat Monitoring*, EURASIP JASP special issue of sensor networks, Vol. 2003, No. 4, pp. 392-401, March 15, 2003.

21. J.S.K Wong, R. Nayar, and A. R. Mikler, *A framework for a World Wide Web-based Data Mining system*, Journal of Network and Computer Applications, 1998 Volume 21 pp.163-185.
22. USEPA, 1991, *Methods for aquatic toxicity identification evaluations. Phase I, Toxicity Characterization Procedures*, 2nd Edition, EPA/600/6- 91/0303, Office of Research and Development, Duluth, MN.
23. W. Ye, J. Heidemann, and D. Estrin, *An Energy-Efficient MAC Protocol for Wireless Sensor Networks*, Proceedings of the 21st International Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002), New York, NY, USA, June 2002.
24. J. Zhao, R. Govindan, and D. Estrin, *Computing Aggregates for Monitoring Wireless Sensor Networks*, Proceedings of First IEEE International Workshop on Sensor Network Protocols and Applications Anchorage, AK, May 2003.