The Impact of Data Aggregation in Wireless Sensor Networks*

Bhaskar Krishnamachari
Cornell University ECE
bhaskar@ece.cornell.edu

Deborah Estrin
UCLA CS
destrin@lecs.cs.ucla.edu

Stephen Wicker
Cornell University ECE
wicker@ece.cornell.edu

Abstract

Sensor networks are distributed event-based systems that differ from traditional communication networks in several ways: sensor networks have severe energy constraints, redundant low-rate data, and many-to-one flows. Data-centric mechanisms that perform in-network aggregation of data are needed in this setting for energy-efficient information flow. In this paper we model data-centric routing and compare its performance with traditional end-to-end routing schemes. We examine the impact of source-destination placement and communication network density on the energy costs and delay associated with data aggregation. We show that data-centric routing offers significant performance gains across a wide range of operational scenarios. We also examine the complexity of optimal data aggregation, showing that although it is an NP-hard problem in general, there exist useful polynomial-time special cases.

1 Introduction

The wireless sensor networks of the near future are envisioned to consist of hundreds to thousands of inexpensive wireless nodes, each with some computational power and sensing capability, operating in an unattended mode. They are intended for a broad range of environmental sensing applications from vehicle tracking to habitat monitoring [1, 8, 10]. The hardware technologies for these networks – low cost processors, miniature sensing and radio modules – are available today, with further improvements in cost and capabilities expected within the next decade [1, 5, 7, 8]. The applications, networking principles and protocols for these systems are just beginning to be developed [2, 4, 8, 11].

Sensor networks are quintessentially event-based systems. A sensor network consists of one or more "sinks" which subscribe to specific data streams by expressing interests or queries. The sensors in the network act as "sources" which detect environmental events and push relevant data to the appropriate subscriber sinks. For example, there may be a sink that is interested in a particular spatio-temporal phenomenon ("does the temperature ever exceed 70 degrees in area A between 10am and 11am ?"). During the given time interval all sensors in the corresponding spatial portion of the network act as event-based publishers. They publish information toward the subscribing sink if and when they detect the indicated phenomenon.

Because of the requirement of unattended operation in remote or even potentially hostile locations, sensor networks are extremely energy-limited. However since various sensor nodes often detect common phenomena, there is likely to be some redundancy in the data the various sources communicate to a particular sink. In-network filtering and processing techniques can help to conserve the scarce energy resources.

Data aggregation has been put forward as an essential paradigm for wireless routing in sensor networks [3, 6]. The idea is to combine the data coming from different sources enroute – eliminating redundancy, minimizing the number of transmissions and thus saving energy. This paradigm shifts the focus from the traditional *address-centric* approaches for networking (finding short routes between pairs of addressable end-nodes) to a more *data-centric* approach (finding routes from multiple sources to a single destination that allows in-network consolidation of redundant data).

In this paper we study the energy savings and the delay tradeoffs involved in data aggregation and how they are affected by factors such as source-sink placements and the density of the network. We also investigate the computational complexity of optimal data aggregation in sensor networks and show that although it is generally NP-hard, there exist polynomial special cases.

2 Routing Models

We focus our attention on a single network flow that is assumed to consist of a single data sink attempting to gather information from a number of data sources. We start with simple models of routing schemes which use data aggregation (which we term data-centric), and schemes which do not (which we term address-centric). In both cases we assume there are some common elements - the sink first sends out a query/interest for data, the sensor nodes that have the appropriate data then respond with the data. They differ in

*This work was supported by the DARPA SensIT program.

the manner the data is sent from the sources to the sink:

Address-centric Protocol (AC): Each source independently sends data along the shortest path to sink (“end-to-end routing”).

Data-centric Protocol (DC): The sources send data to the sink, but routing nodes enroute can look at the content of the data and perform aggregation on multiple input packets. We consider in this paper simple aggregation functions (such as duplicate suppression, min, max) in which multiple input packets can be aggregated into a single output packet.

3 Data Aggregation

3.1 Optimal and Suboptimal Aggregation

Consider k sources, S_1 through S_k and a sink D. Let the network graph $G = (V, E)$ consist of all the nodes V, with E consisting of edges between nodes that can communicate with each other directly. With the assumption that the number of transmissions from any node in the data aggregation tree is exactly one, the following result holds:

Result 1: The optimum number of transmissions required per datum for the DC protocol is equal to the number of edges in the minimum Steiner tree in the network which contains the node set $(S_1, ...S_k, D)$. Hence, assuming an arbitrary placement of sources and a general network graph G, the task of doing DC routing with optimal data aggregation is NP-hard.

We examine three generally suboptimal schemes in this paper:

1. **Center at Nearest Source (CNS)**: In this data aggregation scheme, all sources send their data directly to the source which is nearest the sink which sends the the aggregated information on to the sink.
2. **Shortest Paths Tree (SPT)** : In this data aggregation scheme, each source sends its information to the sink along the shortest path between the two, and overlapping paths are combined to form the aggregation tree.
3. **Greedy Incremental Tree (GIT)** : This is a sequential scheme: at the first step the aggregation tree consists of only the shortest path between the sink and the nearest source. At each step after that the next source closest to the current tree is connected to the tree[1].

3.2 Sensor Network Models

We focus primarily on two performance measures in exploring the gains and tradeoffs involved in data-centric protocols – energy savings due to aggregation in terms of the number of transmissions, and the aggregation latency.

The chief factors that can affect the performance of data aggregation methods are the positions of the sources in the network, the number of sources, and the communication network topology. In order to investigate these factors, we study two models of source placement, the *event-radius* (ER) model, and the *random sources* (RS) model. In both models, we generate a sensor network by scattering n sensor nodes , one of which is a sink node, in a unit square. In the ER model all sources are located within a distance S of a randomly chosen “event” location, whereas in the RS model k random nodes are chosen to be sources. All nodes are assumed to be able to communicate with any other nodes that are within a communication range R.

[1]The GIT scheme is known to have an approximation ratio of 2 [9].

4 Energy Savings due to Data Aggregation

4.1 Theoretical Results

We now give some analytical bounds on the energy costs and savings that can be obtained with data aggregation, based on the distances between the sources and the sink, and the inter-distances among the sources. The main point of this section is that the greatest gains due to data aggregation are obtained when the sources are close together and far away from the sink.

Let d_i be the distance (in terms of number of hops) of the shortest path from source S_i to the sink in the graph. Per datum, the total number of transmissions required for the optimal AC protocol in this case (call it N_A) is:

$$N_A = d_1 + d_2 + ...d_k = sum(d_i) \quad (1)$$

Let the number of transmissions required for the optimal DC protocol be N_D.

Definition : The “diameter” X of a set of nodes S in a graph G is the maximum of the pairwise shortest paths between these nodes $X = max_{i,j \in S} SP(i, j)$ where $SP(i, j)$ is the shortest number of hops needed to go from node i to j in G.

Result 2: If the source nodes $S_1, S_2, \ldots S_k$ have a diameter $X \geq 1$, the total number of transmissions (N_D) required for the optimal DC protocol satisfies the following bounds:

$$N_D \leq (k-1)X + min(d_i) \quad (2)$$

$$N_D \geq min(d_i) + (k-1) \quad (3)$$

Corollary: If the diameter $X < min(d_i)$, then $N_D < N_A$. In other words, the optimum data-centric protocol will perform strictly better than the AC protocol in terms of the total number of transmissions.

Result 3: Assume X and k are fixed, then as $min(d_i)$ tends to infinity (i.e. as the sink is farther and farther away from the sources):

$$lim_{d \to \infty} \frac{N_D}{N_A} = \frac{1}{k}. \quad (4)$$

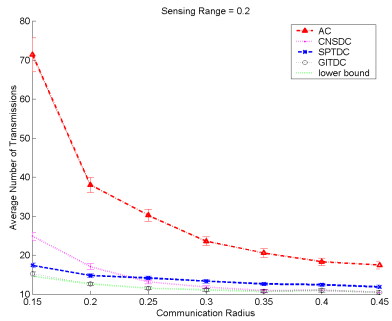

Figure 1. Comparison of energy costs versus R in the ER model

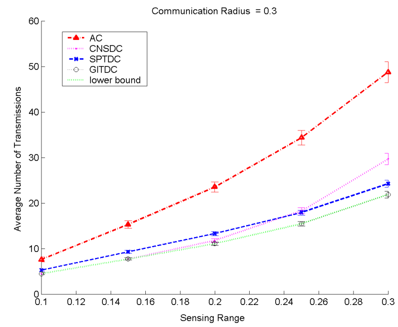

Figure 3. Comparison of energy costs versus S in the ER model

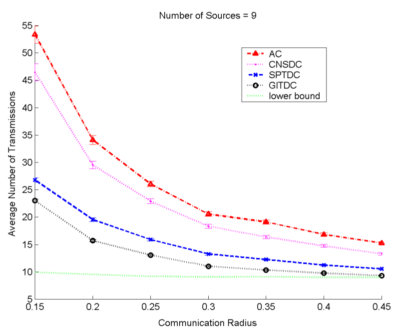

Figure 2. Comparison of energy costs versus R in the RS model

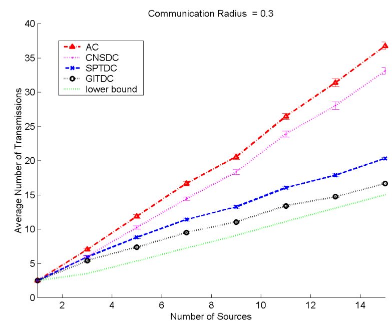

Figure 4. Comparison of energy costs versus k in the RS model

Result 4: If the subgraph G' of the communication graph G induced by the set of source nodes $(S_1, \ldots S_k)$ is connected, the optimal data aggregation tree can be formed in polynomial time.
Corollary: In the ER model, when $R > 2S$, the optimal data aggregation tree can be formed in polynomial time.

4.2 Simulation Results

We now present our simulation results showing the energy costs of AC and DC protocols for both the ER and RS source placement models. The experimental setup is as follows: $n = 100$ nodes are placed in a square area of unit size; for the ER model, the sensing range S is varied from 0.1 to 0.3, and for the RS model the number of sources k is varied from 1 to 15; in both cases the communication radius R is varied from 0.15 to 0.45. For each combination of S or k and R 100 simulations were run. Any runs resulting in unconnected graphs or no sources, which can happen occasionally when the values of S or R are low, were not taken into account. The error-bars shown in the plots represent the standard error in the mean.

Figure 1 compares the transmission energy costs of the various protocols as the communication range is varied In this figure it can be seen that the GITDC seems to coincide with the lower bound (from relation (3)) throughout. This is because the necessary conditions for result 4 holds with high probability in this setting. Also, the performance of the CNSDC approaches the optimal as R increases, as per the corollary to result 4. In all cases there is a $50-80\%$ savings compared to the AC protocol. One thing to note in figure 2 for the RS model is that the lower bound is no longer tight, since the the sources are unlikely to be within one hop of each other except when R is very high. Intuitively, CNSDC performs poorly in the RS model since the sources can be far apart and it is not always beneficial to aggregate at the source nearest to the sink.

Figures 3 and 4 both show that while the absolute transmission costs may increase, the relative gains due to a good data aggregation technique like GITDC can be very significant when the number of sources is high.

To summarize, our experiments show that the energy gains due to data aggregation can be quite significant particularly when there are a lot of sources (large S or large k) that are many hops from the sink (small R).

5 Delay due to Data Aggregation

Although data aggregation results in fewer transmissions, there is a tradeoff – potentially greater delay in

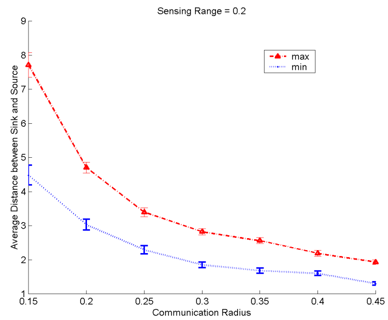

Figure 5. $max(d_i)$ **and** $min(d_i)$ **versus** R **in the ER model**

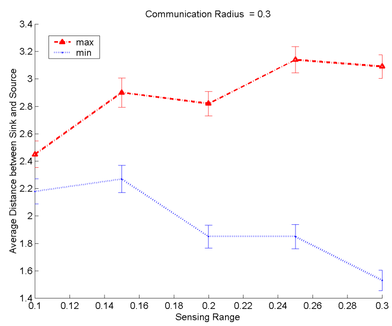

Figure 6. $max(d_i)$ **and** $min(d_i)$ **versus** S **in the ER model**

the case of some aggregation functions because data from nearer sources may have to be held back at an intermediate node in order to be aggregated with data coming from sources that are farther away. In the worst case, the latency due to aggregation will be proportional to the number of hops between the sink and the farthest source. One way to quantify the effect of aggregation delay is to examine the difference $max(d_i)$ and $min(d_i)$. This is shown in figures 5 and 6. Similar figures obtain for the random sources model as well. The experimental setup is the same as discussed in section 4.2. The upper curve in all these figures is representative of the latency delay in DC schemes with non-trivial aggregation functions and the lower curve is representative of the latency delay in AC schemes. The difference between these curves is greatest in both models when the communication radius is low, and the number of sources is high.

6 Conclusions

Wireless sensor networks are an important type of resource-constrained distributed event-based system. We have modelled and analyzed the performance of data aggregation in such networks. We identified and investigated some of the factors affecting performance, such as the number of placement of sources, and the communication network topology. The formation of an optimal data aggregation tree is generally NP-hard. We presented some suboptimal data aggregation tree generation heuristics and showed the existence of polynomial special cases.

The modelling tells us that whether the sources are clustered near each other or located randomly, significant energy gains are possible with data aggregation. These gains are greatest when the number of sources is large, and when the sources are located relatively close to each other and far from the sink. The modelling, though, also seems to suggest that aggregation latency could be non-negligible and should be taken into consideration during the design process. Data-centric architectures such as directed diffusion should support a Type of Service (TOS) facility that would permit applications to effect desired tradeoffs between latency and energy.

In-system processing of data is useful to avoid overwhelming the consumer of data notification, be it a person or a program. Thus the results we have presented in this paper for a resource-constrained event-based system might well hold important design lessons for scalable event-based systems, even if they are less constrained.

References

[1] A. Cerpa *et al.*, “Habitat monitoring: Application driver for wireless communications technology,” *2001 ACM SIGCOMM Workshop on Data Communications in Latin America and the Caribbean*, Costa Rica, April 2001.

[2] D. Estrin *et al.*“Next Century Challenges: Scalable Coordination in Sensor Networks,” MobiCom 1999.

[3] J. Heidemann *et al.*, “Building Efficient Wireless Sensor Networks with Low-Level Naming,” *18th ACM Symposium on Operating Systems Principles*, October 21-24, 2001.

[4] W.R. Heinzelman, J. Kulik, and H. Balakrishnan “Adaptive Protocols for Information Dissemination in Wireless Sensor Networks,” MobiCom 1999.

[5] J. Hill *et al.*, “System Architecture Directions for Networked Sensors,” *ASPLOS*, 2000.

[6] C. Intanagonwiwat, R. Govindan and D. Estrin, “Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks,” MobiCom 2000.

[7] J. M. Kahn, R. H. Katz and K. S. J. Pister, “Mobile Networking for Smart Dust”, MobiCom 1999.

[8] G.J. Pottie, W.J. Kaiser, “Wireless Integrated Network Sensors,” *Communications of the ACM*, vol. 43, no. 5, pp. 551-8, May 2000.

[9] H. Takahashi and A. Matsuyama, “An approximate solution for the steiner problem in Graphs,” *Math. Japonica*, vol. 24, no. 6, pp. 573-577, 1980.

[10] J. Warrior, “Smart Sensor Networks of the Future,” *Sensors Magazine*, March 1997.

[11] A. J. Goldsmith and S. B. Wicker, “Design Challenges for Energy-constrained Ad Hoc Wireless Networks,” *to appear in IEEE Communications Magazine*.