Definition of Data Aggregation

- Data Aggregation (Data Fusion) means combining data from different sources enroute to the destination in order to remove redundancy.
Problem Definition

- Minimize data communication among sensors. Implications:
  - Reduced energy consumption and hence longer lifetime of sensor nodes
  - Reduced collisions in the network
LEACH - Low Energy Adaptive Clustering Hierarchy [HC00]
Motivation

- Energy constraints in sending all data to the base station
- Avoid information overload
- Added advantage is that data aggregation (data fusion), can combine several unreliable data measurements to produce a more accurate signal by enhancing the common signal and reducing the uncorrelated noise.
First Order Radio Model

- Transmit Electronics: $E_{elec} \times k$ 
  - $E_{Tx}(d)$ 
- Tx Amplifier: $e_{amp} \times k \times d^2$ 
- Receive Electronics: $E_{elec} \times k$ 
- $E_{Rx}$ 
- $k$ bit packet

Distance: $d$
Assuming $r^2$ energy loss due to channel transmission:

- Energy required to transmit a $k$-bit message to a distance $d$ is given by

$$E_{Tx}(k, d) = E_{Tx-\text{elec}}(k) + E_{Tx-\text{amp}}(k, d)$$

$$E_{Tx}(k, d) = E_{\text{elec}} \times k + \epsilon_{\text{amp}} \times k \times d^2$$

- Energy required to receive $k$-bits is

$$E_{Rx}(k) = E_{Rx-\text{elec}}(k)$$

$$E_{Rx}(k) = E_{\text{elec}} \times k$$

- The two equations show that transmit energy required is proportional to the distance of transmission and that energy is expended for reception too. Thus minimizing number of transmits/receives helps save energy.
Assumptions

- The base station is fixed and located far from the sensors.
- All nodes in the network are homogeneous and energy constrained.
Algorithm

- Cluster based algorithm in which cluster-heads are selected randomly for every “cycle”.
- Sensors belonging to a particular cluster are assigned transmit time-slots by the cluster head in which they can send data to the cluster heads.
- Data is aggregated at the cluster head and sent to the sink.
- After one cycle, the sensors that have not become cluster heads elect themselves as the cluster head for the coming cycle with a particular probability.
Cluster Head Formation

- At every “round”, each node independently decides whether it wants to become cluster head for that round.
- This decision is based on the suggested percentage of cluster heads for the network (determined a priori) and the number of times the node has been a cluster-head so far.
- Node selects a random number between 0 and 1. If this is less than a threshold $T(n)$, it becomes cluster-head for that round.
Cluster Head Formation (contd)

- Threshold $T(n)$:

$$T(n) = \begin{cases} \frac{P}{1 - P \times (r \mod \frac{1}{P})} & \text{if } n \in G \\ 0 & \text{otherwise} \end{cases}$$

where $P =$ the desired percentage of cluster-heads (e.g., $P = 0.05$), $r =$ the current round, and $G$ is the set of nodes that have not been cluster-heads in the last $1/P$ rounds.

- Each node becomes cluster-head at some point in $1/P$ rounds.
Cluster Set-up Phase

- Each node decides which cluster it wants to subscribe to and sends a message to that cluster-head using CSMA MAC protocol.
- Decision as to which cluster-head a node wants to belong to may be based on the distance to the cluster-head, energy required to transmit to the cluster-head and so on.
- In case of a tie, random cluster-head is chosen.
Schedule Creation

- Based on the number of nodes in the cluster, the cluster-head creates a TDMA schedule for each node to transmit the data to it.
- The schedule is broadcasted to the nodes in the cluster.
Cluster Formation Algorithm

Node i cluster-head?

Yes

Announce cluster-head status
Wait for Join-Request messages
Create TDMA schedule and send to cluster members t = 0

No

Wait for cluster-head announcements
Send Join-Request message to chosen cluster-head
Wait for schedule from cluster-Head t = 0

Steady-state operation for t = T_{round} seconds
Data Transmission

- The nodes turn on their transmitters only during their respective TDMA time slots and transmit the data.
- After all the data has been aggregated for a particular round, the cluster-head performs signal processing on the aggregated data and compresses it.
- The composite (compressed) data is then transmitted to the base-station.
Multiple Clusters

- Each cluster uses a different CDMA code for intra-cluster communication. The spreading code is chosen randomly by the cluster-head from a list.
- This reduces interference in neighboring clusters.
Cluster head talk to “super cluster-head” nodes and so on until the top layer of the hierarchy at which point the data will be sent to the base station.
Features

- Localized coordination and control for cluster set-up and operation.
- Randomized rotation of the cluster “base stations” or “cluster-heads” and the corresponding clusters.
- Local compression to reduce global communication
Lifetimes for different types of communications

<table>
<thead>
<tr>
<th>Energy (J/node)</th>
<th>Protocol</th>
<th>Round first node dies</th>
<th>Round last node dies</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>Direct</td>
<td>55</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>MTE</td>
<td>5</td>
<td>221</td>
</tr>
<tr>
<td></td>
<td>Static Clustering</td>
<td>41</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>LEACH</td>
<td>394</td>
<td>665</td>
</tr>
<tr>
<td>0.5</td>
<td>Direct</td>
<td>109</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>MTE</td>
<td>8</td>
<td>429</td>
</tr>
<tr>
<td></td>
<td>Static Clustering</td>
<td>80</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>LEACH</td>
<td>932</td>
<td>1312</td>
</tr>
<tr>
<td>1</td>
<td>Direct</td>
<td>217</td>
<td>468</td>
</tr>
<tr>
<td></td>
<td>MTE</td>
<td>15</td>
<td>843</td>
</tr>
<tr>
<td></td>
<td>Static Clustering</td>
<td>106</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>LEACH</td>
<td>1848</td>
<td>2608</td>
</tr>
</tbody>
</table>
Comparison with MTE (Minimum Transmit Energy) and other algorithms

- LEACH can improve (doubles) system lifetime compared with MTE routing.
- Factor of 8 reduction in communication energy compared to conventional routing protocols. Irrespective of the number of nodes, the time taken for the first node to die in LEACH is 8 times longer than the time taken for the first node to die in other protocols.
Directed Diffusion [IG]
Terms Used

- **Task originator**
  - Sensor which queries for information.

- **Codebook Entry**
  - A table which maps the sensed waveform to an “event” description.

- **Data Naming**
  - Naming the task by attribute-value pairs that describe it.

- **Gradients**
  - Rate at which the data must be sent back to the source node. Different paths may have different “gradients” as requested by the source node.
Design Challenges

- Scale
- Robustness
- Energy Efficiency
Figure shows the event being monitored by distributed sensors.

The path highlighted in orange shows the path with highest gradient and the diffusion of data from the event to the sensor sink.

The aggregation of data takes place at the point where the two paths from the sensor sources meet.
Concepts

- Data Centric – All communication for named data
- All nodes are application aware
- Actuators send query of the form – “Tell me the direction of the vehicle in region Y”
- Nodes with data send attribute-value pairs back towards the originator
- Collaboration to obtain relevant data and improve precision
Data Naming

Interest:
- Type: four-legged animal  // Type of data
- Interval: 20ms // for every 20ms
- Duration: 10 seconds // over a period of 10 seconds
- Location: [-100 100 200 -200]  // Area of interest
Data Naming

Data returned:

- Type: four-legged animal
- Instance: elephant
- Intensity: 0.6
- Confidence: 0.8
- Timestamp: 01:20:40
- Location: [150 200]
Forming Reinforced Paths

- Exploratory interest propagation in the direction of the event from the sink
Gradients are set up towards the node from which the exploratory interest packets were obtained.
The sink now reinforces the optimal path by requesting a higher data rate from that path.

Exploratory paths still exist for use in case of faults in the optimal path.
Link Failure is detected when the data rate decreases or when packets are not received.

In this case, the best path among the exploratory paths is chosen.

Loops are avoided by negatively reinforcing paths from where delayed packets arrive.
# Design Choices

<table>
<thead>
<tr>
<th>Diffusion element</th>
<th>Design Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest Propagation</td>
<td>• Flooding&lt;br&gt; • Constrained or directional flooding based on location&lt;br&gt; • Directional propagation based on previously cached data</td>
</tr>
<tr>
<td>Data Propagation</td>
<td>• Reinforcement to single path delivery&lt;br&gt; • Multipath delivery with selective quality along different paths&lt;br&gt; • Multipath delivery with probabilistic forwarding</td>
</tr>
<tr>
<td>Data caching and aggregation</td>
<td>• For robust data delivery in the face of node failure&lt;br&gt; • For coordinated sensing and data reduction&lt;br&gt; • For directing interests</td>
</tr>
<tr>
<td>Reinforcement</td>
<td>• Rules for deciding when to reinforce&lt;br&gt; • Rules for how many neighbors to reinforce&lt;br&gt; • Negative reinforcement mechanisms and rules</td>
</tr>
</tbody>
</table>
Metrics for Analysis

- **Average Dissipated Energy:**
  - Ratio of total dissipated energy in the network to the number of distinct events seen by the sink
  - Denotes Average work done by the nodes in delivering information to the sink
  - Indicates overall lifetime of the network

- **Average Delay:**
  - Average one-way latency observed between transmitting an event and receiving it at each sink
Average energy dissipated is less for diffusion as compared to flooding since only optimal paths are used for data transmission and there are no redundant data transmissions as in flooding. The energy calculation does not take into account the energy dissipated during the exploratory phase.
Omniscient Multicast finds least delay paths and Directed Diffusion comes very close to this. Flooding has very high delay due to collisions and high contention for the medium.
Impact of Node failures on Average Dissipated Energy

- Graceful degradation of the energy performance with node failure due to conservative negative reinforcements.
Impact of Node Failures on Delay

- Again, the graceful degradation in delay performance is expected since the next best path is chosen on failure of a node in the optimal path.
Impact of Node Failures on Event Delivery Ratio

- In the event of the failure of the nodes in the optimal link, the next best path from the exploratory paths is chosen and this ensures good event delivery ratio even in adverse conditions.
Other Methods for Data Aggregation

- Data Funneling[PS03]
  - If small number of sinks, makes assumption that nodes in a particular area will transmit information to the same sink
  - Instead of sending small messages, border nodes in the cluster aggregate data by stripping headers and appending data from other nodes
  - Also uses ordering of data from different nodes to achieve compression. (might not work too well for vector data)
Other Methods for Data Aggregation

- Infer [HL05]:
  - Distributed algorithm for putting nodes to sleep and using some inference models to infer the data of inactive sensors.
  - Uses principle of diminishing marginal returns.
  - Increases network lifetime and achieves energy savings of up to 59% with 7.9% accuracy.
Other Methods for Data Aggregation

- SPIN-Sensor network Protocols for Information via Negotiation [AS02]:
  - In a way inverse of directed diffusion. Nodes with data send advertisements and interested nodes request data.
Conclusion

- Application aware sensor networks can effectively compress data by using the data centric approach for routing and aggregating data.
- Dynamic clustering approach for data aggregation results in considerable energy savings.
- Redundancy in data can be exploited by the forwarding node to decrease traffic and hence improve performance of the sensor network.
Future

- Aggregation aware reinforcement strategies for the directed diffusion approach must be studied. This is since the optimal path need not necessarily mean that the amount of data aggregated is maximum.

- Better source coding methods for data compression in sensor networks (these should be tailored to meet the low delay and low computational energy requirements of sensor networks).

- Incorporating error detection and correction during data aggregation. (this may potentially mean that there will not be much of error propagating from one cluster to the “super cluster” in cluster based methods and so on).
References

- [AS02] Ian F. Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci, A Survey on Sensor Networks, IEEE Communications Magazine, August 2002
That’s All Folks!!!