A Dynamic Aggregation Protocol for Energy Efficient Data Fusion in Wireless Sensor Network

Adwitiya Sinha
School of Computer & System Sciences
Jawaharlal Nehru University
New Delhi, India

D. K. Lobiyal
School of Computer & System Sciences
Jawaharlal Nehru University
New Delhi, India

ABSTRACT
Effective data fusion principally prolongs the survival of a Wireless Sensor Network (WSN) and largely determines the degree of its performance in terms of energy utilization. In our research work, we propose a data fusion protocol based on clustering technique. The protocol computes the correlation-dominating set by exploiting spatial and temporal correlation among the data sensed by the sensor nodes in the network. On the basis of the dominating set the network correlation graph is derived, which is further applied to form clusters. Moreover, an efficient energy model is taken into consideration for electing a sensor node from the dominating set as the cluster head. Finally within a cluster, the cluster head aggregates data from the remaining dominating nodes and transmits them to the data processing node. It can be observed that with the application of correlation and aggregation in our protocol, the size of the set of actually transmitting nodes is reduced significantly. We have used Network Simulator (ns-2.34) to simulate our work. The results are obtained in terms of three metrics: energy consumption, success rate and network lifespan. The results are obtained by taking average of five runs, to ensure precision in the experimentation.

General Terms
Wireless Sensor Network, Data Fusion, Data Correlation, Graph Theory.

Keywords
Connected correlation dominating set; network correlation graph; BF-hypergraph; data correlation and covariance; data aggregation.

1. INTRODUCTION
Wireless Sensor Networks (WSNs) are becoming a very significant enabling technology in many sectors. This highly distributed framework of tiny and lightweight devices, called sensors (sensor nodes or nodes), is a quite promising descendant of Mobile Ad Hoc Network. The essential goal of sensor network is to collect and aggregate [1] meaningful information from local raw data gathered by the individual sensor nodes, in an energy saving manner. Therefore, to overcome the problem of energy consumption, we propose a dynamic and energy-efficient data aggregation protocol.

The paper is subdivided into several sections. The first section provides introduction, the next section involves the related work being carried out in the concerned field. The network model is depicted in the third portion. Fourth section presents a detailed description over the proposed aggregation protocol. Fifth section introduces network correlation graph (NCG). Sixth section shows the statistical computation involved in deriving the NCG. This follows with the simulation results in the seventh section. Finally the last subdivision concludes the paper with future scope.

2. RELATED WORK
Abundance of research work is carried out in the field of data correlation and aggregation protocols. Gupta [3] proposed an efficient data-gathering algorithm exploiting the data correlation. In the article, they designed techniques that exploited spatial correlations in sensor data to minimize communication costs (and hence, energy costs) incurred during data gathering in a sensor network. Their main approach was to select a small subset of sensor nodes that may be sufficient to reconstruct data for the entire sensor network. Thus, during data gathering only the selected sensors needed to be involved in communication. However, their algorithm is not based on the clustering technique, and the overhead from selecting the connected correlation-dominating set compromises with the efficiency of the proposed algorithm. In addition, their work does not address the data dynamics. In Temporal in-Network Aggregation (TiNA) protocol [4], the approach is to send the data only when there is a significant change in the data value in the adjacent readings over time. The concept of epoch is also used here for synchronizing the receipts of the packets from the child nodes and sending the aggregate. However, TiNA exploits temporal correlation in sensor data while our proposed protocol DAP (Dynamic Aggregation Protocol) takes advantage of both spatial and temporal correlations in sensor data. Moreover, the memory overhead is also higher as compared to the proposed protocol. The Prediction-based Monitoring (PREMON) protocol [5] provides energy-efficient monitoring based on a clustered architecture. Cluster head nodes in PREMON use a technique similar to the MPEG compression algorithm, and generate prediction models to predict the spatio-temporal data within a cluster. PREMON saves energy by avoiding the transmissions of all the redundant data that can be successfully predicted by the cluster head node. But PREMON assumes that the clusters are already formed using any existing mechanism while DAP forms clusters using real-time sensor values. Further study, in this context, reveals that though several protocols are available, but each of them has to compromise on one or more of the following parameters: result accuracy, data correlation, aggregation efficiency, clustering strategy, memory (storage) overhead, effective power control, energy efficient routing, and network throughput. Therefore, it becomes essential to develop a protocol that addresses all these problems, generally confronted by the protocols developed so far, for fusing data from multiple

sources. In response to this, the design of Dynamic Aggregation Protocol (DAP) with clustering technique is taken as the research initiative.

3. NETWORK MODEL

3.1 Problem Statement

In the current work, we purpose to design and develop a Dynamic Aggregation Protocol (DAP) for sensor network that enables fusion of data before transmission to the sink. The fusion of data further leads to reduction in the rate of energy being expended, and therefore increasing the sensor network life span.

The proposed DAP protocol involves the following:

- It supports adaptive clustering of the sensor nodes, i.e. the clustering parameters are recalculated on the epoch timer expiry and adjusts the cluster formation over time.
- It allows in-network lossless aggregation of data. This means all the data samples gathered in the beacon phase are used to statistically derive the fused data.
- There is provision epoch timer, on the expiration of which the network correlation structure is recomputed, to promote data dynamism.
- It also employs an efficient energy model to monitor the utilization of energy during three states: data transmission, data reception and idle state. It also tracks the residual energy with respect to individual nodes, deployed in the sensor network.
- The usage of a random scheduler to transmit data in a random pattern, in order to avoid frequent rate of packet collision.
- It also designates the routing of the correlated and aggregated data from a node, which has an acceptable energy level and is nearer to the sink node.
- It exploits spatial and temporal correlation, thereby decreasing the size of correlation dominating set, i.e. the *active* sensor nodes actually involved in transmission.

3.2 Working Scenario

We consider a sensor field as highlighted in Fig. 1. The sensing area consists of substantial number of sensors deployed in a random manner.

Initially, all the sensors are in *alive* state. However, during the execution of our protocol, the state of any node (s) can be changed to either *active* (selected for data transmission) or to *dormant* (sleep) state. Every sensor node maintains an initial vector of readings recorded by its neighboring sensors at six different time periods.

4. PROPOSED PROTOCOL

Our proposed protocol in this paper, works over three phases: beacon phase, correlation phase and clustering phase.

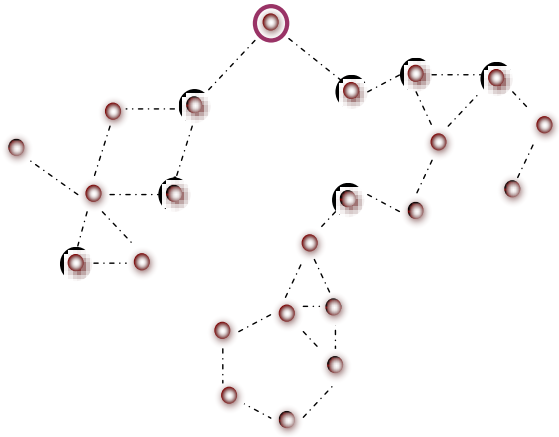

Fig 1: Shows the sensors deployed in the network. The dotted lines represent the communication links.

The protocol commences with the *beacon phase*. This phase initiates the neighborhood discovery for each and every node in the network. The neighborhood set of each sensor is determined with the network simulator [6]. Then, we gather d-hop neighborhood information (data sampling). In our simulation d=2 (neighborhood size), i.e. spatial correlation [7] will be exploited to a degree of two in the proposed network model. The neighborhood span is reflected in Fig. 2.

Next is the *correlation phase*, in which the Network Correlation Graph (NCG) derivation is performed. In this phase, the NCG is described as a BF-hypergraph (Backward-Forward Hypergraph) [8], whose tail set represents the set of the *active* sensors that actually transmit data to the data processing node (also called sink node), and the head set depicts the *dormant* sensors that do not take part in the data transmission. To determine the elements of the tail set with respect to a particular node, its neighborhood set is derived and a neighborhood matrix is formed with the environmental data sensed by the neighboring sensors at six discrete time periods.

The degree of positive correlation is found between the initial vector and the neighborhood matrix with an error tolerable upto 25%, i.e. if two sensors show at least a correlation of 0.5, only then they can be regarded as correlated sensors. This procedure is executed over all the nodes thereby producing a set of few *active* nodes that dominate over all the nodes in the network (i.e. forms a node cover for the sensor network considered for simulation). This set can be regarded as the connected correlation dominating set [9]. The members of the set can be directly or indirectly connected. When *active* sensors are directly connected by 1-hop links (which will be probed in the clustering phase), they form clusters, otherwise they are connected by Steiner nodes [11] or intermediate nodes.

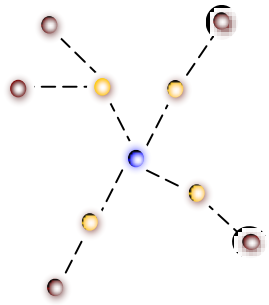

Fig 2: Shows the 1-hop & 2-hop sensors of a particular node (blue) in the network (neighborhood size taken into consideration). The yellow nodes are 1-hop neighbors red ones are the 2-hop neighbors that escape the direct coverage of the concerned (blue) sensor.

The final phase is the *clustering phase*, in which correlation signal is multicasted to all the neighboring sensors of a concerned sensor node. On the basis of the present state of a node, next state transition takes place. A node, whose state is once updated to *active* or *dormant*, cannot be chosen for future state change till the current epoch timer expires. The sensor states and correlation values are recomputed once the timer expires. This supports dynamism without compromising with network stability. The clusters are formed with 1- hop connected *active* nodes along with their respective correlated nodes (*dormant* nodes). The clustering of the nodes is shown in Fig. 3. Once the clusters are formed, nodes satisfying the following conditions are explored:

- The nodes are in *alive active* state.
- The residual energy satisfies an acceptable level.
- The node is in minimum hop distance from the data gathering node.

The data from the *active* sensors are then transferred to the cluster head, where the necessary aggregation of the correlated data takes place. Finally, the correlated and aggregated data (more accurate and precise data) is transferred to the sink node, thereby drastically reducing the size of previously found connected correlation dominating set (set of *active* nodes) to a smaller sized set that are only comprises the cluster heads. More precisely, instead of multiple *active* nodes in a cluster, only one node i.e. the cluster head will participate in the transmission of data to the data gathering node.

5. NETWORK CORRELATION GRAPH

Correlation, one of the data fusion techniques, is considered as a powerful statistical tool for exploiting similarities in the data pattern generated by the sources, with respect to different dimensions. The analysis of the readings from all sensors is required for the layout of the network correlation structure. Furthermore, the network correlation structure depends upon a stable correlation graph. Therefore, derivation of the Network Correlation Graph (NCG) of a given sensor field becomes a prime issue.

5.1 Mathematical Description

A "***Network Correlation Graph***" can be defined as a BF-hypergraph (or BF-graph) NCG=(S,E), where:

$S = \{ s_1, s_2, \ldots, s_n \}$ is the set of sensors, and

$E = \{ E_1, E_2, \ldots, E_m \}$ is the set of all the hyperedges

with:

E_i as the ith hyperedge,

$E_i \subseteq S$ for $i = 1, \ldots, m$, as the set of directed BF-hyperedges.

A hyperedge (E_i) is formed by connecting a number of sensors in the network. Furthermore, each of the directed hyperedge or hyperarc is an ordered pair, $E = (X,Y)$, of disjoint subsets of vertices; X is the *tail* of E while Y is its *head.* The tail and the head of hyperarc E is be denoted by $T(E)$ and $H(E)$, respectively.

A *B-arc* is a hyperarc with $|H(E)| = 1$ and *F-arc* is a hyperarc with $|T(E)| = 1$. A *BF-graph* is one whose hyperarcs are either B-arcs or F-arcs.

5.2 Experimental Interpretation

The interpretation of using a BF-hypergraph for representing the network correlation structure is as follows:

- In the context of our experimentation, T(E) represents a set of correlating (active alive) sensor(s), and H(E) represents a set of correlated (dormant) sensor(s).
- The tail set can be better referred to as correlation set.
- A sensor can be correlated to the data sensed by one or more sensors of a correlation set. The sensors that are uncorrelated need to be explicitly considered during data aggregation process.
- The same correlation set can be used to correlate one or more sensors, i.e. the correlation set of one or more dormant sensor may overlap.

5.3 Significance

The importance of the BF-graph in the context of the network design is illustrated as follows:

- The sensors in the tail set T(E) of the network correlation graph will only be capable to transmit data.
- The tail set includes the minimum possible sensors, which correlate all the remaining dormant nodes. It forms a node cover for the network graph.

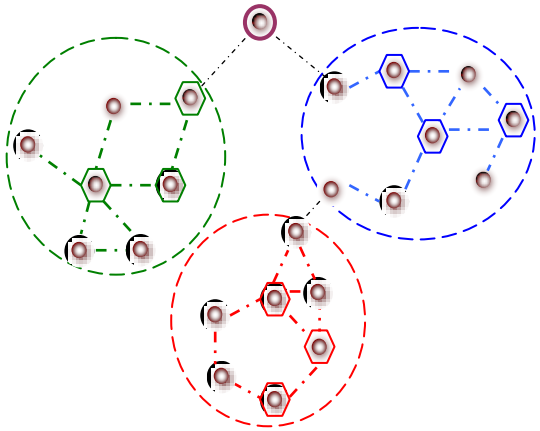

Figure 3. Shows the formation of clusters in the network. The correlation-dominating (CD) nodes are highlighted in hexagonal shapes and the coloured dotted-dashed lines represent the correlated neighbors of the CD nodes.

- T(E) signifies the connected correlation dominating set, in which each of the sensor is either directly connected (within 1-hop range) or connected indirectly via steiner nodes.
- The main aim is to reduce the size of connected correlation dominating set, by:
  - analyzing the list of data generated by each sensor using a multivariate technique
  - exploring the interdependency of the relationship structure among the sensors
  - building a multivariate model to depict the covariance structures
  - estimating the degree of correlation among the set of data sensed by the sensors

6. STATISTICAL COMPUTATION OF NETWORK CORRELATION GRAPH

The **random property** with respect to our sensor model is as follows: "A collection of the number of readings (random samples), recorded by each of the arbitrarily deployed sensor (in a sensing region of 50 sensors) forms the basis of random experiment".

Let the outcome of this random experiment with respect to a sensor (s) be χ. Its neighborhood set of (s), i.e. η_s can be defined as:

$$\eta_s = \{s_i \mid (|s - s_i|.\text{length}) \leq s_{range}\} \quad (1)$$

Each of the random outcomes of s_i in η_s is given by the random variables, γ_i . Let the mathematical expectation with respect to the random variable χ contributed by the sensor (s) be E_s . The respective mean value and variance, in terms of mathematical expectation, are given by:

$$\text{Mean value } (\mu_s) = E_s(\chi) \quad (2)$$

$$\text{Variance } (\sigma_s^2) = E_s[(\chi - \mu_s)^2] \quad (3)$$

Similarly, the mathematical expectation with respect to the random variable γ_i contributed by the sensors (s_i) be Es_i. The relevant mean value and variance for each of $s_i \in \eta_s$, are given by:

$$\text{Mean value } (\mu_{s_i}) = E_{s_i}(\gamma_i) \quad (4)$$

$$\text{Variance } (\sigma_{s_i}^2) = E_{s_i}[(\gamma_i - \mu_{s_i})^2] \quad (5)$$

The aim is to get the measure of correlation between the sensor (s) and each of the neighboring sensors $s_i \varepsilon \eta_s$. To achieve this, covariance structures of a particular sensor's temporal data is to be calculated with regard to each of its neighbors reading collated over time. Using the results from (2-5) covariance structure is calculated as:

$$Cov_{s_i \to \gamma_i,\, s_i \varepsilon \eta_s}(\chi, \gamma_i) = E_{s_i \to \gamma_i,\, s_i \varepsilon \eta_s}[((\chi - \mu_s)(\gamma_i - \mu_{s_i})] \quad (6)$$

Here, '$\rightarrow$' stands for 'contributes to', which means s contributes to the random variable χ, and each of s_i contributes to the random variables γ_i . Finally, the degree of correlation is given by:

$$\rho_{s_i \varepsilon \eta_i}(i) = \left(\frac{Cov_{s_i \to \gamma_i,\, s_i \varepsilon \eta_i}(\chi, \gamma_i)}{\sigma_{s \to \chi}(\chi) * \sigma_{s_i \to \gamma_i,\, s_i \varepsilon \eta_s}(\gamma_i)} \right) \quad (7)$$

Let there be an user specified error threshold ξ, defined by the programmer at the time of the protocol design. A sensor (s) will be correlated to the i^{th} neighboring sensor if the following condition is satisfied:

$$\rho_{s_i \varepsilon \eta_s}(i) \leq \xi \quad (8)$$

As a result of the correlation, the sensor (s) will change its state to *dormant* and the i^{th} neighboring sensor (s_i) will experience a state change to *active* state, and will further form an element of the correlation dominating set.

7. SIMULATIONS AND RESULTS

The network simulator ns-2.34 [10] has been selected for the simulation of our proposed protocol. The initial values required for our simulation are tabularized in Table 1.

The comparison of our proposed protocol is made with Naive and Basic Distributed protocols. The Naive protocol represents a simple protocol in which all the nodes transmit their data to the sink node without exploiting spatial or temporal correlation. It does not involve formation of cluster and therefore none of the sensors ever enter sleep state. The second protocol is the Basic Distributed protocol, which involves only spatial correlation. But it acts as the Naive protocol till the correlation dominating set is formed by exploiting correlation property. Moreover, no aggregation of data or clustering is involved. It resembles the basic distributed protocol introduced in [3].

Table 1. Simulations Parameters Table

S. No.	Variable	Value	Unit
1.	Initial Energy	100	Joules
2.	Transmission Power	4.500	Watts
3.	Reception Power	4.119	Watts
4.	Total Simulation Time	400	Seconds
5.	Propagation Model	TwoRayGround	-
6.	Antenna Type	OmniAntenna	-
7.	Number of Nodes	50	-
8.	Routing Protocol	DSDV	-
9.	Network Dimension	1000 x 1000	$Meter^2$

a. DSDV - Destination Sequence Distance-Vector Routing Protocol

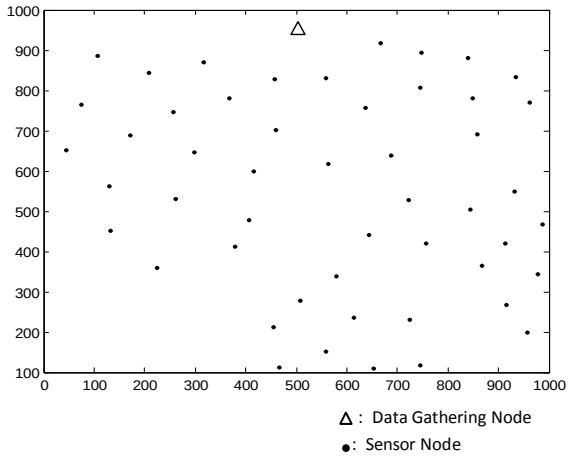

Fig 4: Shows the network layout 1000 x 1000 m^2 with dots indicating the sensors being deployed and the triangular symbol marks the position of the data gathering node.

In Fig. 4, the network scenario selected for simulation is displayed. The sensor field consists of 50 sensors nodes and a data processing node, i.e. the sink node. The clustering of the sensor network is done via the proposed Dynamic Aggregation Protocol (DAP) and is graphically highlighted in Fig. 5, using Matlab. In the following section several graphs are also constructed on the basis of the tracing and log files generated during the simulation of our protocol by the Network Simulator (ns-2.34).

The Fig. 6 shows the average energy expended (in joules) by the sensor network over the simulation time (in seconds). The simulation time is 400 seconds. It is clear from the graph that

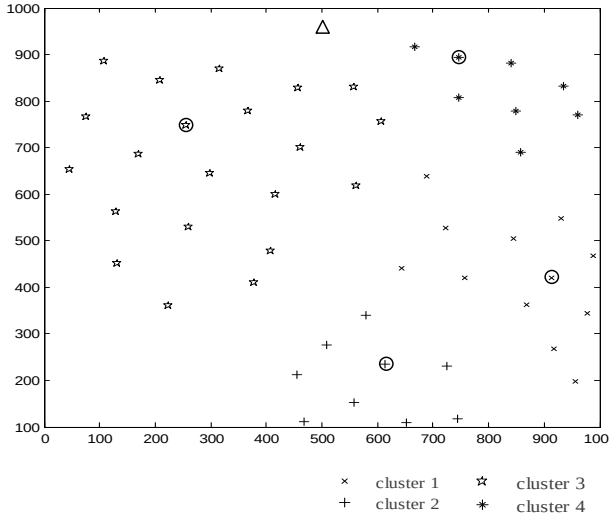

Fig 5: Shows the network structure after clustering. The

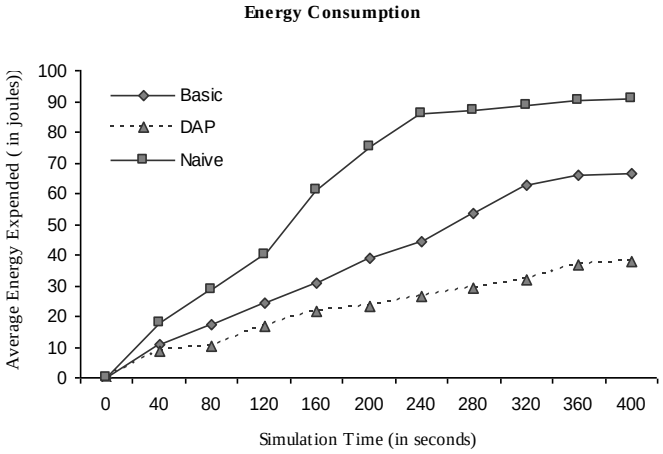

Fig 6: Shows the average energy consumption done by the nodes in the sensor network over simulation time, t=400 seconds

Naive performs the worst. From the time t = 240 second to 400 second, the average energy utilization almost becomes constant. This happens because of the fact that, as the simulation time approaches 400^{th} second, the rate of dead sensors increases at a very rapid rate (Fig. 7). Less the number of alive sensors, minimum will be the energy expended. The Basic protocol, in comparison to the Naive protocol performs relatively better, because of correlating sensors. But our proposed protocol provides the best results. However, at the 100^{th} second, DAP

nears to the Basic protocol (due to the chain of multicast transmissions in the beacon phase), but the situation stabilizes, once DAP does correlation and forms the clusters.

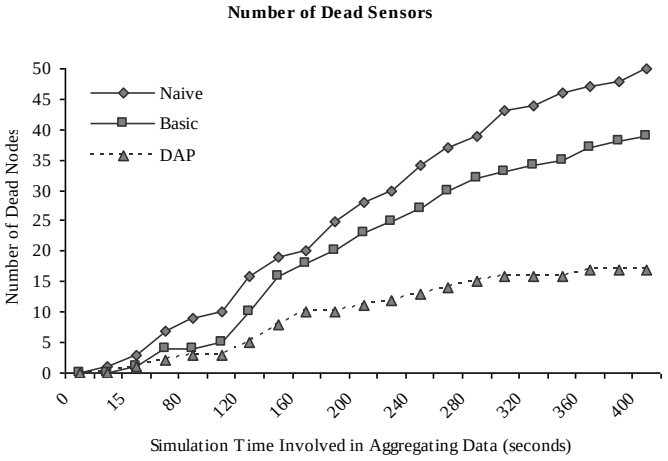

Fig 7: Shows the number of dead sensors in the network over simulation time, t = 400 seconds

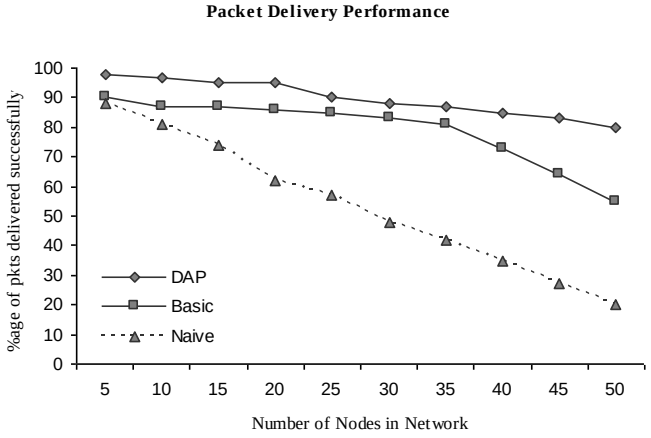

Fig 8: Shows the packet delivery performance over the number of nodes deployed in the network

Figure 7 highlights the formation of number of dead sensors nodes as the simulation proceeds from t = 0 seconds to t = 400 seconds. The Naive protocol results in most number of dead sensors over time. The Basic protocol also produces deteriorating results as compared to the performance of our protocol. The DAP lead to a maximum of 17 dead sensors by the end of the simulation. Since eventually the cluster head transmits data to the sink, the energy used is reduced, which ultimately results in improved network lifespan.

Figure 8 shows the rate of successfully delivered packets with respect to the *active* nodes at a specific instant in the network. In the starting phase, with smaller number of nodes, all the three protocols show a performance of 85% to 95%. The reason is fewer data transmissions are required with small number of sensors, which further implies lesser chances of congestions and therefore better delivery of the data packets to their destinations. However, the actual simulation results can be seen as the number of nodes increases. The result interprets that the Naive protocol suffers the most followed by the Basic Distributed protocol, but the proposed protocol depreciates with a minimal percentage. This further clarifies that data packets are delivered successfully in a comparatively better way. Moreover, the congestion capability of DAP, allows the performance to improve by re-sending the collided, damaged or lost packets.

It is apparent that, in every respect the proposed DAP protocol outperforms the Basic Distributed & the Naive protocol. Therefore, DAP establishes the required goal of our research work.

8. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

We have developed energy efficient fusion protocol (DAP), which is based on clustering technique. The clusters are formed on the basis of Network Correlation Graph, which is constructed with the help of connected correlation dominating set. The dominating set is computed by fusing data from sensors nodes, thereby utilizing spatial and temporal correlation. Moreover, with smaller number of nodes to transfer data, the network traffic is minimized, likeliness of occurrence of congestion is reduced and the network bandwidth is better utilized. Usage of lossless aggregation ensures data accuracy. This further implies that the quality of the fused data is not affected. The proposed protocol outperforms the Naive Protocol and Basic Distributed Protocol [3] in terms of energy consumption, success rate and network lifetime.

However, the research proposal can be further directed towards achieving better performance. In our simulation, all the sensors are assumed to have fixed sensing range and very low degree of mobility. As a scope for future extension, the concerned protocol can be developed for adhoc environment, where the degree of mobility of sensor node is quite high.

9. REFERENCES

[1] Alzaid, H., Foo, E. and Nieto, J. G. 2008. Secure data aggregation in wireless sensor. Information Security Institute, Qeensland University of Technology. ACSC2008 conference, Wollongong, Australia.

[2] Buratti, C., Giorgetti, A. and Verdone, R. 2005. Cross-layer design of an energy-efficient cluster formation algorithm with carrier-sensing multiple access for wireless sensor network. EURASIP Journal on Wireless Communications and Networking.

[3] Gupta, H., Navda, V., Das, S. and Chowdhury, V. 2008. Efficient gathering of correlated data in sensor networks," State University of New York, Stony Brook, ACM Transactions on Sensor Networks, Vol. 4, No. 1, Article 4.

[4] Sharaf, M. A., Beaver, J., Labrindis, A. and Chrysanthis, P. K. 2003. TiNA: A scheme for temporal coherency-aware

in-network aggregation. ACM Workshop on Data Engineering for Wireless& Mobile Access (MobiDe).

[5] Goel, S. and Imielinski, T. 2001. Prediction-based monitoring in sensor networks. ACM Computer Communications Rev. (CCR).

[6] Fall, K. and Varadhan, K. The NS Manual.

[7] Pattem, S., Krishnamachari, B. and Govindan, R. 2008. The impact of spatial correlation on routing with compression in wireless sensor networks. University of Southern California, ACM Transactions on Sensor Networks, Vol. 4, No. 4, Article 24.

[8] Gallo, G., Longo, G. and Nguyen, S. and Pallottino, S. 1992. Direced hypergraphs and applications. National Research Council of Canada.

[9] Li, J., Foh, C. H., Andrew, L. H. L. and Zukerman, M. 2008. Sizes of minimum connected dominating sets of a class of wireless sensor networks. State University of New York, Stony Brook. ACM Trans. on Sensor Networks.

[10] Ros, F. J. and Ruiz, P. M. 2004. Implementing a new manet unicast routing protocol in NS2. Dept. of Information and Communications Engineering, University of Murcia.

[11] Chakraverty, S., Batra, A. and Rathi, A. 2006. Directed convergence heuristic: A fast and novel approach to steiner tree construction", IFIP International Conference on Very Large Scale Integration.