VOL. 13, NO. 5, MARCH 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CLUSTERING TECHNIQUES OF WIRELESS SENSOR NETWORKS FOR INTERNET OF THINGS

Anagha Rajput and Vinoth Babu K.
School of Electronics Engineering, VIT University, Vellore, Tamil Nadu, India
E-Mail: anagharajput@gmail.com

ABSTRACT

Internet of Things (IoT) is one of the emerging and disruptive technology. IoT based systems are used in environment surveillance, e-healthcare, automation for industry, etc. Wireless sensor networks (WSN) is a predominant option to realize such systems because WSN though resource constrained but can connect cyber network and the physical real environment. The present day requirement of IoT is densely deployed sensor nodes over a large area. In this paper, clustering protocols for scalability is premeditated to clip the suitability with respect to IoT applications. The study of progressive algorithms gives a detail explanation of mathematical models used and their effect on WSN. It also aimed to present a comparison of performance metrics for single and multiple hop clustering protocols. Thus, exploring challenges fronting to refrain WSN for IoT applications.

Keywords: internet of things (IoT), wireless sensor networks (WSN), clustering, scalability.

1. INTRODUCTION AND ENERGY MODEL

The future of Internet/web is seen to be a large network where people, objects or anything will be connected at any time. This is termed as Internet of Things. It is popularly abbreviated as IoT. As per [1], by 2020, a target of connecting billions of things to internet will be realized. The things are of different variety in terms of size, mobility and their deployment in the environment. Here, things can be electronic gadget, animals, packing stuffs, living or non-living but are physical and real in existence. Thus, realization of IoT based systems will result to a disruptive technology [2]. IoT communication will allow autonomous exchange of information between physical quantity measuring device and the end users at remote place. This can be achieved by the wireless technologies like Radio-Frequency Identification (RFID) and WSNs [3] and [4]. Among the above two technologies, WSN is considered as an important part of IoT monitoring infrastructure as it gives larger coverage range than RFID. Thus, primarily, WSN is been used commercially to connect physical things to the internet through network devices like base stations, gateways, etc. [5]. This paper explores the WSN cluster architecture and reviews several clustering protocols in the perspective of scalable and energy saving schemes to increase node density and network lifetime.

WSN is a network where spatially scattered sensor nodes measure various parameters like temperature, voice activity, motion, pressure, air and water pollutants, etc. according to the application scenario. To overview an IoT based WSN, an example of temperature monitoring of a particular area is explained. A number of smart temperature sensor nodes are dispersed throughout the monitoring area. The temperature is sensed by sensor nodes and transmitted to base station or sink through wireless channel. In this paper, we use sink notation for convenience. Sink will further route the data to gateway. This sink to gateway communication might be wired or wireless channel. Gateway are the devices that bridges one network architecture with another network architecture usually web architecture. Thus, gateway converts the data packets coming from sink into Ethernet packets and forward to the user through Internet. Such many wireless sensor networks deployed for variety of applications can be connected to Internet. A WSN based IoT network scenario can be visualized as shown in Figure-1.

As sensor nodes are smart nodes, the network can be designed as time driven, query driven or event driven network. In time driven network, end user will be receiving data from sensor nodes periodically. In query driven network, a remote end user can access the data at any time instant whenever required. In event driven network, alarming signals are sent to the control system and required action is taken.

A wireless sensor node is a tiny embedded processor device with sensors interfaced to it. It has three functionalities. They are sensing, signal processing and wireless communication. These nodes are battery driven, which limits their operative life. Among above mentioned functionalities, a node utilizes more energy during wireless communication than the other functionalities [6]. Thus, many researchers have concentrated to estimate energy consumption due to radio communication by designing various communication models. Communication model consists of transmitter and receiver placed at some distance D. Every circuit operation needs some battery power. These are known as energy factors and are considered to estimate energy spent by the device. Most popularly used communication model is explained further. An energy calculation for one time data transmission is also illustrated.

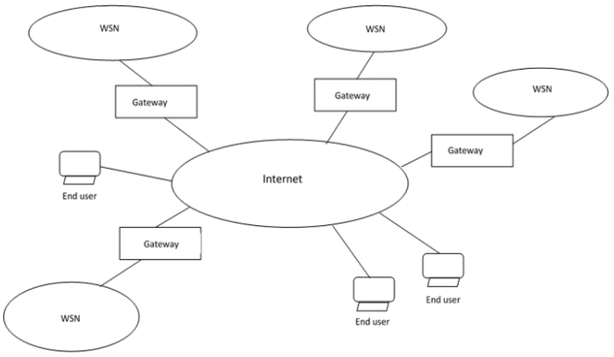

Figure-1. WSN based IoT network scenario.

Sensor node's energy utilization in WSN is concerned with network operation. A radio communication model determines energy needed to send a packet of Z bits from transmitter to receiver at a distance of D. One time data transmission by all the nodes in the network towards sink is called as a round. This data transmission towards sink might take place by direct transmission or by clustering method. In direct transmission methods, all nodes individually transmit data to sink. In this method, nodes far away from sink depletes more energy because energy dissipation is directly related to transmission distance. In clustering, group of near nodes are formed. Each group is called as one cluster. A leader is chosen among the group called as cluster head node (CHN). Other nodes in the group are called non-cluster head nodes (NCHN). Sensed data from NCHNs in a group are transmitted to CHN. CHN will further transmit data collectively to sink [7-10]. In [11], a cluster formation technique was popularly introduced which reduces sensor node's energy consumption by reducing number of transmissions towards sink. An illustration of radio interface communication model to calculate total energy spent for each round of the clustering algorithm is as follows. The typical radio interface communication model is shown in Figure-2.

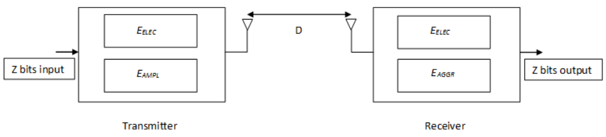

Figure-2. Radio communication energy model.

The energy to transmit Z bits by a sensor node is denoted by $E_{Tx}(Z, D)$.

$$E_{Tx}(Z, D) = Z(E_{ELEC} + E_{AMPL}) \tag{1}$$

where E_{ELEC} is energy term that takes into account transmitter and receiver electronic circuits like digital signal processing. $E_{AMPL}$is distance dependent term that takes into account radiated power necessary to transmit one bit over a distance of D, which is calculated as.

$$E_{AMPL} = \varepsilon\, D^{\alpha} \tag{2}$$

Here α is path loss that ranges from 2 to 5. $\varepsilon$is a term that characterizes transceiver and is measured in J/bit m^{α}. It is formulated as,

$$\varepsilon = \frac{SNR\ NF_{Rx} N_o\ B \left(\frac{4\pi}{\lambda}\right)^{\gamma}}{Gain_{ant}\ \eta\ R_b} \tag{3}$$

where SNR is signal to noise ratio, NF_{Rx} is receiver noise figure, $N_o$is noise power spectral density, B is channel noise bandwidth, λ is wavelength, γ is a constant, $Gain_{ant}$ is antenna gain, η is transmission efficiency, R_b is channel bit rate.

Sensor network size is usually in terms of meters. Thus, many researchers prefer free space propagation model for short distance communication and multipath fading channel model for long distance communication. For free space propagation model, E_{AMPL} is calculated using Eq. (4).

$$E_{AMPL} = \varepsilon_{fs} D^2 \tag{4}$$

Here D is distance less than a predetermined threshold distance D_0. $\varepsilon_{fs}$is calculated from Eq. (3) for free space model and is usually a constant. Similarly, for multipath fading channel, E_{AMPL} is calculated using Eq. (5).

$$E_{AMPL} = \varepsilon_{mp} D^4 \tag{5}$$

Here D is distance greater than predetermined threshold distance D_0. $\varepsilon_{mp}$is calculated from Eq. (3). The D_0 value is assumed by researchers as per the application for which the protocol is designed. The energy to receive Z bits by CHN is denoted by $E_{Rx}(Z)$, which is given by

$$E_{Rx}(Z) = Z(E_{ELEC} + E_{AGGR}) \tag{6}$$

where E_{AGGR} represents energy consumed by CHN for data aggregation to form one data packet. Calculation of total energy consumed for one round can be further elaborated as follows. Let, a cluster has Ʊ number of NCHNs and one CHN. Energy consumed by Ʊ number of NCHNs to transmit Z bits of data to CHN is given as,

$$E_{NCHN} = \text{Ʊ} E_{Tx}(Z, D_{toCHN}) \tag{7}$$

where D_{toCHN} is the distance between sensor node and CHN. Energy consumed by CHN to receive data from Ʊ nodes and transmit it to sink is given as,

$$E_{CHN} = \text{Ʊ} E_{Rx}(Z) + E_{Tx}(Z, D_{toS}) \tag{8}$$

where D_{toS} is the distance between CHN and sink. For a cluster, total energy consumption is calculated using Eqs. (7) and (8).

$$E_{Cluster} = E_{CHN} + E_{NCHN} \tag{9}$$

Total energy consumption for one round is then given as,

$$E_{Round} = \sum_{i=1}^{C} E_{Cluster_i} \tag{10}$$

where C is the total number of clusters in the network. Above energy model terminologies are commonly used in the explanation of protocols in section 3.

The energy conservation or reduction in energy consumption is one of the key and essential requirement for WSN to sustain network life for long time. In IoT applications, number of sensors used are on large scale. As number of nodes increases, energy cost also increases. Thus, WSN architecture designing is an influencing factor wherein researchers have many aspects and parameters, which can be analyzed and redeveloped for IoT systems. The active area in WSNs which still can be explored for energy efficient and omnipresent network are optimal data routing, data aggregation, scalability, fault tolerance and sensor measurement validation and maintenance. In clustering, group of sensor nodes form cluster and transmit data collectively by selecting one node among them for each transmission. Densely deployed sensor nodes over a large area is the present day requirement for the applications based on IoT. As sensor nodes are energy constraint devices, effective utilization of sensor node's energy becomes crucial as well as developing an algorithm for scalable network have also turn out to be one important requirement [12] and [13].

Rest of the paper is organized as follows: Section 2 illustrates classification of WSN clustering techniques. Section 3 contains detail explanation and comparison of clustering protocols. Finally, conclusion with challenging issues are discussed in section 4.

2. CLASSIFICATION

In this section, WSN is studied within a perspective of clustering methodology. A cluster in WSN has two parts. One is cluster architecture properties that decides the deployment strategy of WSN. These can be selected based on the application requirements. Another part is cluster performance attributes based on the measuring metrics like CHN distribution, node density, node death rate etc. Figure-3 elaborates the classification of cluster architecture properties.

2.1 Cluster architecture properties

Using cluster architecture properties, clustering protocols can be designed as per the required application. A brief explanation of these properties is discussed further.

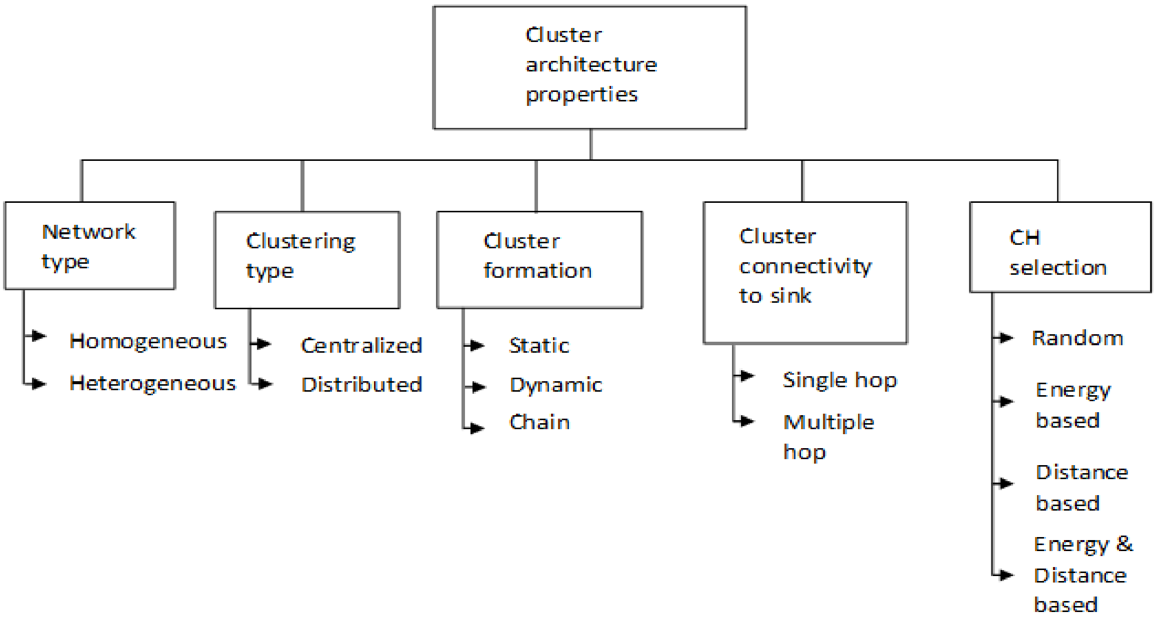

Figure-3. Classification of cluster architecture properties.

2.1.1 Network type

Network type is classified as homogeneous or heterogeneous network. In homogeneous network, all nodes are assumed to have same characteristics and perform in the similar manner. In context of energy, all nodes are assumed to have same initial energy. In heterogeneous network, nodes with dissimilar properties are considered to be deployed in the same monitoring area. Industrial automation is the good example of the heterogeneous network infrastructure. In many of the WSN protocols, mobility property of nodes is popularly considered to define heterogeneity of the network.

2.1.2 Cluster type

Clustering protocols can be basically approached in two ways - Centralized and Distributed clustering. In centralized clustering, a sink forms clusters and also selects CHNs. Sensor nodes have to only sense the environment and obey the protocol instruction given by the sink. The energy of nodes can be saved tremendously in centralized clustering schemes. But sink will get over burden as it has to handle both way operations towards sensor network and also towards gateway. In distributed clustering, each sensor nodes themselves decide whether to become a CHN or not. It is also called as self-organizing network. In distributed clustering, energy saving is achieved by reducing number of long distance communications towards sink.

2.1.3 Cluster formation

The deployment of nodes and its formation into groups called clusters are used to classify clusters as static, dynamic or chain formation. In static cluster formation, the group of nodes formed initially as clusters do not change throughout the process. Only the CHN rotation is distributed among themselves for every round. In dynamic cluster formation, new group of nodes are formed as clusters for every round or in between the process conditionally. Dynamicity may depend on number of cluster members, dead nodes in the past cluster and undoubtedly average energy of the network. Another class of cluster is chain formation. All the nodes in the network are connected to only its neighbor to forward data packet.

2.1.4 Cluster connectivity

A CHN is elected to transmit collective data of cluster towards sink. This can be done in two ways - single hop and multiple hop. In single hop, every CHN directly transmit data to sink. In multiple hop, data is transferred through intermediate nodes. These intermediator might be another CHN node or various techniques are used to find strong NCHNs to act as intermediators.

2.1.5 CHN Selection

The CHN is a node that depletes more energy than other sensor nodes as it performs additional operations like data reception, aggregation, etc. Thus, a CHN chosen in the algorithm for data transmission has impact on the overall performance of the network. A node which is energy content and has long lifetime can be selected as CHN. CHN can be selected stochastically or non-stochastically based on random selection, energy, distance and combination of energy and distance parameters.

2.2 Cluster performance attributes

The operation of protocols can be analyzed with performance attributes as shown in Figure-4. There can be many more attributes. Few of them are discussed here.

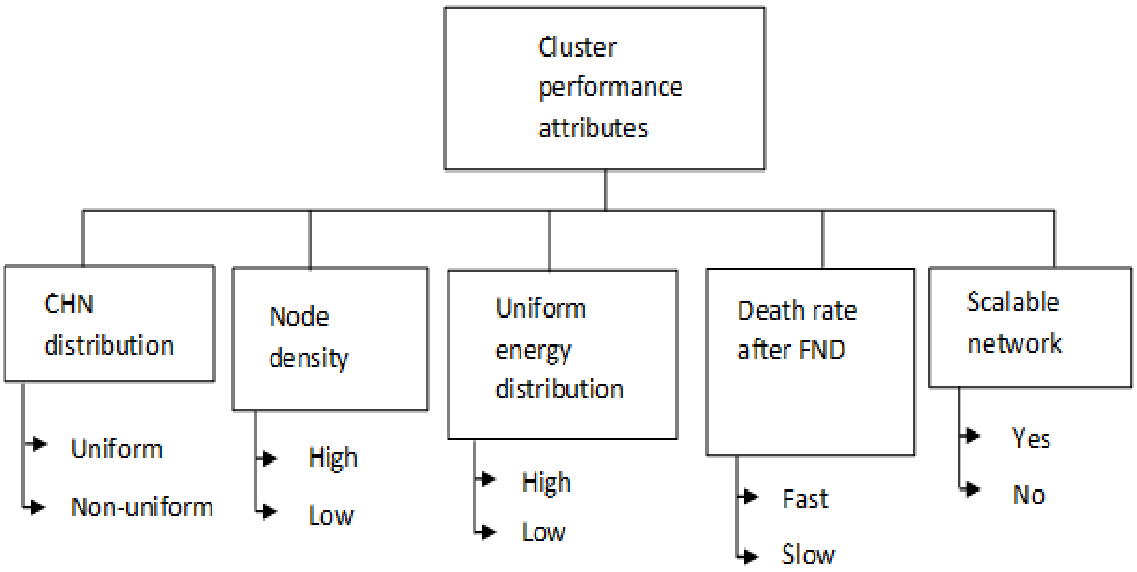

Figure-4. Cluster performance attributes.

2.2.1 CHN Distribution

A cluster is usually associated with one CHN. The CHNs selected in the network must have following features:

High energy for extra processing compared to other nodes.

a) All CHs must cover overall monitoring area.
b) None of the two nearest nodes must be selected as CHN.

CHN distribution is uniform, if all CHNs in the network satisfy above features. Otherwise, issues like hot spot, lack of connectivity, node isolation will occur.

2.2.2 Node density

The sensor nodes are deployed by dispersing them throughout the monitoring area. WSNs are personal area networks with few meter square area coverage. Node density is calculated as number of nodes present per unit area. For instance, in an area of 100x100 meter square, if 100 nodes are dispersed randomly, then node density is given as,

$$Node\ density = \frac{100}{100\ x\ 100} = 0.01 \qquad (11)$$

This metric is useful for analysis of node mobility [10] and [31]. Alive nodes can also be represented in terms of node density for densely populated network.

2.2.3 Uniform energy load distribution

Sensor nodes are energy constraint devices. Thus, allocating tasks to these nodes becomes a crucial decision. Among all, CHN consumes more energy. If a node is selected as CHN frequently, it will die out early. In the network, almost all nodes should sustain maximum rounds. First node dead (FND) and last node dead (LND) are the measure of energy load distribution [15]. FND occurs too early because energy load distribution is non-uniform. Other parameters affecting load is number of trans-receptions of query packets within cluster.

2.2.4 Death rate after first node dies

The network shuts down when all nodes are dead in the network. Death rate is said to be fast when number of rounds between FND and LND are few hundred. When number of rounds between FND and LND are large in thousands, it implies network dies slowly. This death rate of node might be fast or slow depending on algorithm design. The fast death rate is due to relative dependency of nodes on each other. Fast death rate must be avoided as once network is dead, the area remains uncovered. Some applications like military surveillance, need few nodes alive to collect information from the network for more time.

2.2.5 Scalability

This attribute states whether an algorithm can perform consistently if number of nodes are increased within the given monitoring area. Most of the algorithms fail to sustain large number of nodes due to lack of proper clustering process. This attribute is the main concern as IoT environment is heterogeneous with large number of nodes in the network.

3. CLUSTERING ALGORITHMS

Considering the IoT platform, key requirement is scalability of the network with energy efficiency. Scalable is considered in context of increase in node density and coverage area. Thus, in this paper protocols are distinguished on the basis of single hop and multiple hop clustering algorithms.

3.1 Single hop clustering algorithms

In single hop clustering algorithms, data is transmitted to final sink destination through one hop. This means there is single intermediator between source node and destination node. The typical single hop clustering protocols are explained in this subsection.

LEACH [11]

This algorithm is popularly abbreviated as LEACH (Low Energy Adaptive Clustering Hierarchy). It is a self-organizing, adaptive clustering protocol, which aims to equalize energy load distribution among the sensor nodes in the WSN. This is implemented by forming clusters of sensor nodes into the network. One time data transmission from sensor network to sink is called as one round. Every round has two phases – Set-up phase and Steady state phase. These two phases work out consecutively and continuously for environment monitoring. The operation of LEACH is explained in following steps.

a) Before the start of first round, a percentage of CHN to be selected in the network is fixed. The authors simulated using 5% (i.e. p = 0.05) of nodes as CHN from the total of 100 nodes. Here the probability of being a CHN is indicated by p.
b) In set-up phase, a node is checked whether it was a CHN for past $\frac{1}{p}$ rounds. If not, a random number is generated within a range of 0 to 1 and compared with TH value. TH is the threshold on which a node decides its eligibility to become CHN. If random number generated is less than TH then the node becomes a CHN. TH is given as follows,

$$TH = \frac{1}{1-p*\left(r,mod\left(\frac{1}{p}\right)\right)}, \quad \vartheta \in Q \qquad (12)$$

where r is the current number of round, Q denotes the set of total number of nodes that have not been CHN for past $\frac{1}{p}$ rounds.

c) Once CHNs are decided, these nodes broadcast 'REQUEST' message for cluster formation. The NCHNs select their CHN based on received request packet. By sending 'REQUESTACK' message, nodes join particular CHN node. Thus, clusters are formed depending on the number of CHs formed in every round.
d) In steady state phase, every CHN node forms a TDMA (Time Division Multiple Access) schedule and send it to all its cluster members. According to TDMA schedule, every NCHN sends its sensed data in scheduled time slot. Data gathered at CHN is processed for redundancy and send to sink using CDMA (Code Division Multiple Access).

To prolong the lifetime of the network, LEACH includes randomized rotation of the high-energy CHN and performs local data fusion to transmit the amount of data being sent from the CHNs to the sink. As sink is far away from the network, the energy of only CHNs will be affected as these are directly communicating with the sink. Other nodes energy is thus saved. In every round, dynamic clusters are formed. As random selection of nodes is done, chance of near nodes getting selected as CHN cannot be avoided. This will result non-uniform distribution of CHN in the network.

Despite the good performance, LEACH also has some flaws. Because of probabilistic criterion, good CHN distribution cannot be guaranteed and some nodes will not have any CHN in their range. The CHNs are assumed to have a long communication range. This may not always be a realistic assumption. Additionally, LEACH is applicable in one-hop topology, thus it cannot be used effectively on large scale networks.

PEGASIS [14]

This algorithm is popularly abbreviated as PEGASIS (Power Efficient GAthering in Sensor Information Systems). It is a chain formation protocol. The algorithm concentrates to still minimize number of transmissions towards sink than LEACH [11]. Each node communicates only with a close neighbour and takes turns to become leader to transmit fused data to sink using Greedy algorithm. This will reduce amount of energy spent per round for CHN selection.

Greedy algorithm is used to find the optimal choice within the given set of choices for the next instant input set. Advantage of this algorithm is that chain formation is not required for every round. This increases lifetime of the network. Following is the methodology for finding the next neighbour.

a) All the nodes are placed randomly in the given monitoring area.
b) An assumption is made that all nodes have global knowledge of the network. A node is selected randomly as a CHN to initiate chain formation before start of first round. The CHN selected passes the control token till the ends of the chain as shown in Figure-5.

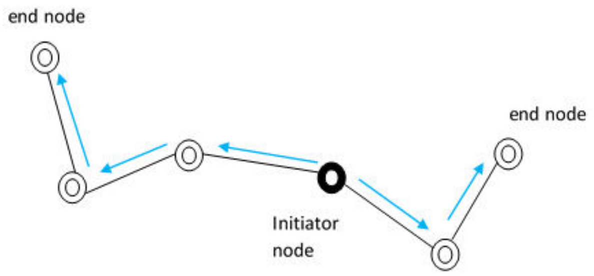

Figure-5. Token passing from initiator towards ends of chain.

c) After token passing process, data passing starts from the end nodes of the chain. Every node will fuse its data with incoming packet and passes further except end nodes as shown in Figure-6.

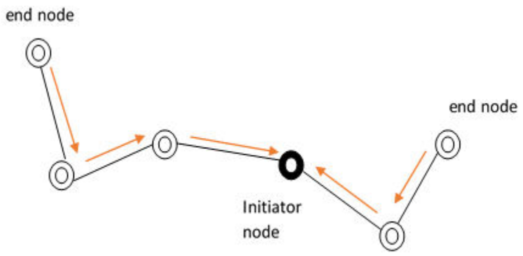

Figure-6. Fused data passing towards initiator.

Neighbour distance will increase gradually as nodes already on chain will not respond to incoming token. In greedy algorithm, paths are formed instantly. No criteria is used to find nearest neighbour. Also, tie cases for neighbours at equidistant is not considered. Due to this, few nodes will have long route and more energy dissipation compared to other nodes. As number of nodes increases, chain formed is more complex and time consuming.

LEACH-E [15]

This is the extended and modified version of [11] abbreviated as LEACH-E (LEACH-Extended). This algorithm reduces the randomness in CHN selection by adding energy ratio to Eq. (12). Following are the steps of algorithm.

a) The threshold depends on current residual and initial energy of the node and is given as below:

$$TH = \frac{p}{1-p\left(r \; mod\left(\frac{1}{p}\right)\right)} \frac{E_{residual}}{E_{initial}} \qquad (13)$$

where $E_{initial}$ and $E_{residual}$ and are initial and current residual energies of the node respectively.

b) The threshold is compared with randomly generated number. The initial energy value remains greater as number of rounds increases. This yields very low value. This algorithm fails to maintain proper threshold value, because often it yields very low *TH* value. The value is lesser than the randomly generated number, which may not allow eligible nodes to become CHN for the current round.
c) This situation is sorted out by further adjusting Eq. (13) as given below

$$\mathrm{TH} = \frac{\mathrm{p}}{1-\mathrm{p}\left(\mathrm{r\;mod}\left(\frac{1}{\mathrm{p}}\right)\right)}\left[\frac{\mathrm{E_{residual}}}{\mathrm{E_{initial}}} + \mathrm{r_c div}\left(\frac{1}{\mathrm{p}}\right)\left(1-\frac{\mathrm{E_{residual}}}{\mathrm{E_{initial}}}\right)\right]$$
(14)

where r_c is number of consecutive rounds in which a node has not been a CHN. This adjustment is temporary solution and authors lack to explain the network stuck up when same situation might occur while using the secondary adjustment Eq. (14).

d) The maximum value of r_c is $\frac{1}{p}$. When r_c reaches to its maximum value, the corresponding term equates to zero. This yields a value same as Eq. (13). This implies that above adjustment is done to increase the value of *TH* so that CHN selection process must not halt in between the protocol execution.

During ideal operation, the deterministic CHN selection avoids the issue of hot spot. But after few rounds, problem of electing CHN arises because $E_{residual}$ value keep decreasing.

LEACH-C [16]

This algorithm is popularly abbreviated as LEACH-C (LEACH-Centralized). The authors tried to overcome limitation of LEACH by designing centralized clustering algorithm. The total energy of network is considered for CHN selection process. Following are the steps of algorithm.

a) A new parameter to predefine number of clusters is introduced. For a given sensor network, number of optimum clusters $C_{optimum}$ is given as,

$$C_{optimum} = \frac{\sqrt{N\varepsilon_{fs}}}{\sqrt{2\pi\varepsilon_{mp}}} \frac{w}{D_{toS}} \tag{15}$$

where N is total number of nodes in network. $w$is side of square area in meters. The simulated results state that optimal number of clusters range for 100x100 square meters area with 100 nodes is in between 3 to 5.

b) Initially, for first r rounds, threshold probability of becoming a CHN is calculated depending on $C_{optimum}$ using

$$TH = \begin{cases} \frac{C_{optimum}}{N - C_{optimum}\left(r \bmod \left(\frac{N}{C_{optimum}}\right)\right)}, & Y_i(t) = 1 \\ 0 & , Y_i(t) = 0 \end{cases} \tag{16}$$

where N is total number of nodes in network. $Y_i(t)$isa function used to indicate whether a node was CHN or not in past $\frac{N}{C_{optimum}}$ rounds. This ratio count yields same value compared to $\frac{1}{p}$ in [11] because $C_{optimum}$ and N are kept constant. Initially, alive nodes are N but when nodes die out, alive nodes decreases. Here authors do not consider number of alive nodes for determining number of clusters to be formed in later rounds.

c) The CHN rotation among the nodes is illustrated further. Expected number of nodes not been CHN in first r rounds is $(r\ C_{optimum})$. Thus, for $\left[\frac{N}{C_{optimum}}\right]^{th}$ round, number of nodes not been CHN will be,

$$N - \left(\frac{N}{C_{optimum}} C_{optimum}\right) = 0 \tag{17}$$

Thus, every round, $(r\ C_{optimum})$ number of nodes get reduced from non-CHN list.

d) The authors also considers energy factor in event driven case for CHN selection. Depending on energy level of nodes, threshold probability is calculated using

$$TH = min\left\{\frac{E_{residual}}{E_{total}} C_{optimum}, 1\right\} \tag{18}$$

The total residual energy of the network is calculated using

$$E_{total} = \sum_{i=1}^{N} E_{residual_i} \tag{19}$$

The consideration of total network energy instead of initial node energy as in [11] selects higher energy node often as CHN. This maintains approximate equal lifetime of all the nodes. In this algorithm, all nodes must be intimated the network's residual energy. Also to calculate threshold probabilities, each node must know $C_{optimum}$ and N values. This implies that algorithm is not completely autonomous and requires a routing protocol for global parameters.

DEEC [18]

This algorithm is abbreviated as DEEC (Distributed Energy Efficient Clustering) and implemented for distributed heterogeneous network where global knowledge is not required for network operation. A categorization of nodes into advanced and normal is done for proper energy utilization. An assumption that advanced node will always have φ times more energy than normal node's energy makes the network heterogeneous. Following is the operation flow of DEEC.

a) Nodes deployed in the monitoring area are considered to have energies as follows:

$$E_{normal} = E_{initial}$$
$$E_{advanced} = \varphi\ E_{initial} \tag{20}$$

Advanced node's energy is increased by a constant φ than normal node. From total nodes, a fraction of u nodes are considered as advanced nodes. As node energies are different, the factors for calculating probabilities is also considered differently.

b) The threshold is same as mentioned by LEACH protocol. The p value in Eq. (12) is made dependent on two factors. Every node calculate its probability for becoming CHN using energy based relation which is given as,

$$p = \begin{cases} \frac{C_{optimum}E_{residual}}{(1+\varphi u)E_{total}}, & for\ normal\ node \\ \frac{C_{optimum}(1+\varphi)E_{residual}}{(1+\varphi u)E_{total}}, & for\ advanced\ node \end{cases} \tag{21}$$

c) To achieve optimal number of CHNs in every round, a conditional Eq. (22) is always checked. This limits on number of clusters in the network.

$$\sum_{i=1}^{N} P_i = N\ C_{optimum} \tag{22}$$

d) A probability ratio is multiplied by $\frac{1}{1+\varphi u}$ for normal nodes and $\frac{1+\varphi}{1+\varphi u}$ for advanced nodes. These will change the probability to higher value for advanced nodes and lower value for normal nodes. Thus, selecting higher energy nodes as CHNs.

e) A randomly generated number by CHN candidate is compared with p. If generated number is less than p, nodes declares itself as CHN. Further steady state phase is carried out similar to LEACH.

This algorithm has weighted probability value for each node. This concept is used to enhance multi-level heterogeneous environment. The results state that this protocol has large stable time than [11], [15] and [17]. It outperforms by 20% and 15% than the conventional algorithms [17] and [15] respectively in terms of FND of the network.

DDEEC [19]

This algorithm is the modified version of [18] and abbreviated as DDEEC (Developed DEEC). The authors introduce an additional condition to calculate threshold probability. In the conventional algorithm [18], the criteria for CHN selection was depending only on residual energy and optimum number of CHN in the network. In this extension, the authors have used an additional threshold on residual energy so that very low energy nodes are not considered for CHN selection. The probability for CHN is calculated as per conventional method only when residual energy of node is above a threshold value. This threshold depends on energy dissipated by node during past round.

$$TH_{dis} = E_{initial}\left(1 + \frac{\varphi\, E_{disnn}}{E_{disnn} - E_{disan}}\right) \tag{23}$$

Here E_{disnn} and E_{disan} are the energies dissipated by normal and advanced nodes respectively. The threshold, TH_{dis} depends on initial energy and energy dissipated by the nodes in the past round. The denominator term denotes difference between energy dissipated by normal and advanced node. This energy dissipation ratio ranges from 0 to 1. If the bracket term is equal to zero, the algorithm operates like [18]. As, the bracket value increases towards 1, CHN selection in the network varies abruptly.

If the residual energy drops down below threshold value, probability ratio is multiplied by a constant Ψ. Ψ is the real positive variable that affects number of rounds before FND. The author simulated with different Ψ values and concluded that, for lower values of Ψ (around 0.02), FND is better. This is because if Ψ is increased, probability value of many nodes to become CHN increases. This allows increased number of transmissions. Ultimately, more nodes deplete their energy causing early FND. The simulated results state that this algorithm performs 15% better than [18] with respect to FND of the network. The death rate is not steady as nodes are distinguished in terms of initial energy. This helps in increasing network lifetime and gives a real time implication of heterogeneity.

ALEACH [20]

This algorithm is popularly known as ALEACH (Advanced LEACH) and is aimed to select appropriate CHNs to steadily distribute energy throughout the network. This algorithm uses Eq. (24) for CHN selection. The probabilistic threshold depends on sum of two terms – General node probability (GEN_p) and Current State Node probability (CSN_p). It is given as,

$$TH = GEN_p + CSN_p (24)$$

$GEN_p$is probability based on number of rounds, a node has not become a CHN. $CSN_p$is probability based on current energy of the node. Both GEN_p and CSN_p are calculated as follows,

$$GEN_p = \frac{C_{optimum}}{N - C_{optimum}\left(r \bmod \left(\frac{N}{C_{optimum}}\right)\right)} \tag{25}$$

$$CSN_p = \frac{E_{residual}}{E_{total}} \frac{C_{optimum}}{N} \tag{26}$$

Following are the steps of operation:

a) For every round, alive nodes are considered for CHN selection process. If a node was CHN in past $\frac{C_{optimum}}{N}$ rounds, its threshold value is terminated to zero and node will not participate in CHN selection process.
b) Remaining nodes will calculate threshold value TH and generate a random number between 0 and 1. Unlike LEACH, Eq. (24) yields greater value heading towards 1. This is because both terms GEN_p and CSN_p are calculated with respect to total number of nodes and total residual energy of the network. To maintain least number of CHNs, author select nodes that generate random number greater than TH.
c) If TH is greater than generated random number, that node declare itself as CHN and broadcast the 'REQUEST' message.
d) Other nodes within the communication range and receiving 'REQUEST' message becomes cluster member by sending 'REQUESTACK' message to respective CHN. After, cluster formations, a TDMA schedule is send by each CHN node to its cluster members.
e) In steady state phase, data transmission takes place as per the schedule.

CHNs are selected depending on their current energy level relative to network's energy, which reduces the chances of very low energy node getting selected as CHN. This enhances the lifetime of the network with more number of rounds than the conventional LEACH.

Though number of CHN are limited by $C_{optimum}$ value, cluster members joining a CHN node does not have specific criteria. Depending upon firstly received request, a node joins a CHN. Thus, some CHNs may have many members while some may have very few. This will deplete more energy of the CHNs that have more members. This non-uniform cluster formation fails to reduce transmission distance of nodes.

NCACM [21]

This algorithm is as NCACM (New Clustering Algorithm with bounded Cluster Members) and is a centralized clustering protocol, which concentrates on energy efficiency by better CHN selection. It is implemented for heterogeneous WSN. The heterogeneity is with respect to mobility of nodes. Few nodes are considered to be mobile while most of the nodes are static. An attempt to precise CHN selection is made by considering many factors like battery power, distance towards CHN, etc.

a) CHN selection threshold TH is based on number of rounds and energy ratio as given below. The nodes current residual and its initial energy ratio decreases the TH, as node drains its energy every round.

$$TH = \frac{p}{1-p\left(r \bmod \left(\frac{1}{p}\right)\right)} \frac{E_{residual}}{E_{initial}} \qquad (27)$$

b) After nodes are selected as CHN, the node calculates confidence value ($Cval$). This is used for cluster formation. It indicates the capability of a node to support cluster members with maximum throughput. It is calculated as,

$$Cval = \begin{cases} 0, & B_{CHN} < T(B_{CHN}) \\ \frac{B_{CHN}}{\partial D_{toS}} + \frac{B_{NCHN}}{D_{toCHN}}, & B_{CHN} > T(B_{CHN}) \end{cases} \qquad (28)$$

where B_{CHN} is the battery power of given CHN. $\partial$is number of nodes already a member of given CHN in past round. $D_{toS}$is the distance between CHN and sink. $D_{toCHN}$is the distance between CHN and NCHN. $B_{NCHN}$is battery power of NCHN. $T(B_{CHN})$is the threshold of battery power that can support a limited number of nodes. If the battery power of CHN is less than $T(B_{CHN})$, it implies that CHN has cluster members to its fullest limit. Thus, no more nodes can join such CHN. So the $Cval$ is assigned to zero.

c) If the battery power is greater than $T(B_{CHN})$, CHN calculates its $Cval$ which indicates that it can still support nodes more than∂.
d) This $Cval$ value is broadcasted by CHN along with 'REQUEST' message. A NCHN may receive $Cval$ from two or more CHNs. In this case, a node joins a CHN which have highest $Cval$ value. At NCHN, this $Cval$ value is checked and then 'REQUESTACK' message is sent. If the present battery power is less than $T(B_{CHN})$, additional members will not be added to the given CHN node.

The consideration of $Cval$ term enabled the protocol to distribute network load among most energetic nodes. This leads to increase the network lifetime.

SECA [22]

This algorithm is abbreviated as SECA (Saving Energy Clustering Algorithm). The aim of SECA is to gain a uniform cluster formation to achieve balanced load distribution. Here distance based CHN selection method is introduced. A centralized controller sink update all the nodes with cluster information and CHN selection throughout the process. Initially, sink receives location information and residual energy for each sensor node in network. Then sink calculates average residual energy of network. The algorithm is explained in following steps.

a) In the set up phase, the centre location of the monitoring area is calculated as,

$$L = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (29)$$

where X_i is the coordinate of ith nodes.

b) The average distance between L and all nodes is then calculated as,

$$L_{avg} = \frac{\sum_{i=1}^{N} |X_i - L|}{N} \qquad (30)$$

c) From the value of L, at radius of R, a virtual circumference is considered. Few mean points are determined on that circumference. The number of mean points is equal to $C_{optimum}$. It is calculated as,

$$\mathrm{mp_i} = \mathrm{R}\cos\left(\frac{360}{\mathrm{C_{optimum}}}(\mathrm{i}-1)\frac{\pi}{180}\right) + \mathrm{L}, \mathrm{i} = 1 \text{ to } \mathrm{C_{optimum}} \qquad (31)$$

where mp_i is the coordinate of ith mean point. Thus, monitoring area is partitioned into parabolic sections. This process of setting initial mean points reduces the iteration time required for cluster formation.

d) Further for every mean point, distance between all mean points and all nodes is calculated. Every node joins to a mean point with minimum distance. Thus, N nodes are partitioned into $C_{optimum}$ number of clusters. This is achieved using following k-means Equation (32).

$$avg = min \sum_{i=1}^{K} \sum_{X_j \in H_i} |X_j - mp_i|^2 \qquad (32)$$

where H_i is ith cluster, X_j is coordinates of jth node and mp_i is coordinates of ith mean point.

e) After clusters are created, new mean point is again calculated for every cluster in t iterations using Equation (33).

$$mp_i^{t+1} = \frac{1}{H_i^t} \sum_{X_j \in H_i^t} X_j \tag{33}$$

where H_i^t is the ith cluster executing tth iteration.

f) Steps 4 and 5 are repeated until an approximate centre point of cluster is found with minimum average distance between mean points and cluster nodes.
g) After final cluster is formed, CHN is selected within the cluster nodes by two conditions.
(a) Node with minimum distance from final mean point. This mean point is virtual point on the area and not the node location.
(b) Node having $E_{residual}$ greater than average energy of all nodes.
h) Sink sends all above information to all nodes. Nodes update their routing tables as per received information. Nodes that are declared as CHNs, schedules slots for cluster members. In steady state phase, data transmission takes place as per schedule.
i) Step 2 to 6 is executed only once initially. For later set up phases till a node dies, algorithm follows directly from step 7, where condition (a) is neglected and CHNs will be selected based on only energy criteria.

If a node dies, then all steps are re-executed with new cluster formation. The disadvantage of this process is centre space remains open and transmission is crowded at outer space. CHN nodes will be always from outer area of the region. Another disadvantage is after FND, for every node death, above time consuming calculations are executed leading to increase in computation cost. This algorithm fails to perform consistently for large number of nodes because large computation utilizes time that leads to latency and unreliable system.

Clustering for IoT [9]

This algorithm runs in two phase. It is a query based protocol, where in nodes are updated with recent information of the node locations without any central authority controller. Thus, author has declared this process as fully autonomous protocol. This algorithm uses Non-probabilistic criteria for CHN selection. The steps of operation are as follows:

a) At set up phase, every node starts the counter with random number. The counter is decremented every unit time. Sensing node that has counter value zero is eligible to form cluster.
b) Such cluster originator nodes send 'REQUEST' message to neighbouring nodes within the sensing range. In the network, following are the possible neighbouring nodes to each eligible CHN.

(a) Type I- all are non-cluster members
(b) Type II- all are cluster members
(c) Type III- few nodes are of Type I and few are of Type II.

c) Clustering for Type I: Nodes that receive 'REQUEST' message, acknowledges to originator. In the acknowledgement packet, node send its residual energy and location information. All the nodes that have send 'REQUESTACK' message and the originator together will form a cluster.
d) The originator then calculates approximate cluster centre point (CCP) of the cluster and the selection factor (F) for each node using Equations. (34) and (35).

$$CCP_i = \frac{\sum_{i=1}^{C} \sum_{j=1}^{n} L_j}{C} \tag{34}$$

where CCP_i is center point of ith cluster. $L_j$is the location of the jth node. $n$is number of nodes in the ith cluster.

$$F = \frac{E_{residual}}{E_{average} d_{CCP_{i,j}}} \tag{35}$$

where$F$ is CHN selection factor of each node. $E_{average}$is average residual energy of all nodes in the cluster including originator node. $d_{CCP_{i,j}}$is the distance between ith CCP and jth node.

e) The F values of all nodes in the cluster are compared and node with highest F value is elected as CHN of that cluster. Originator may or may not become CHN depending on its F value.
f) Clustering for Type II: As nodes are already members of other cluster, they send 'JOIN' message as a response to the originator. This message packet contains CCP value and cluster ID of the member node. The originator receives these 'JOIN' messages from nodes that are cluster members of same or different clusters. The originator calculates new CCP values using Equation (34) for all requested clusters. The CCP value is calculated considering originator node as cluster member.
g) Then the distances from originator to all CCP values are calculated and compared. Finally, the originator joins the cluster from whom CCP distance is minimum. Further steps 4 and 5 are executed to select CHN.
h) Clustering for Type III: All Type II nodes are neglected and cluster is formed using steps 3, 4 and 5.
i) After cluster formation, as per TDMA scheduled by CHN, data transmission from nodes to CHN takes place.

The simulated results achieve better network lifetime than the conventional schemes [11] and [21]. This is because CHN selected is the node that is at near centre location. Also the centre point technique reduces the transmission distance of overall nodes. Since the nodes getting zero count start forming clusters, this algorithm do not limit the number of clusters formed. The case where more than required CHNs might form clusters at same time instant. This increases energy consumption in the network, decaying the lifetime.

Dynamic CHN [23]

Maximum network coverage and low energy consumption by participating nodes is the main aim of this algorithm [23]. It builds static clusters using Voronoi diagram. The frequent cluster formation is a wasteful process. The concept of static cluster formation saves energy of network on large scale. In this work, the author concentrated on CHN selection process. The redundant nodes which have average energy level are selected as first kind of CHN. Such nodes prioritize to early death. After the death of first kind of CHs, second class of CHN are selected based on energy criteria. Following are the steps of operation.

a) In the initial set up phase, monitoring area is divided into hexagonal clusters using Voronoi diagram. Each partition is called as Voronoi cell. The advantage of this cell is mean point calculation is not required unlike [9] and [22]. It is mathematically represented as,

$$A_k = \{o \in O; d(o, M_k) \leq d(o, M_j)\}; \quad j \neq k \tag{36}$$

where A_k is the k^th cell. $O$is metric space. $o$is the instantaneous point considered in the O space. $d$is the distance between instantaneous point and the mean point of the cell.

b) The distance of each node from mean point is calculated as,

$$d(n_i, M_j) = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \tag{37}$$

where x and y are coordinates of the ith node and mean point jth mean point.

c) Every node has a relative proximity over a space it is deployed. This space is also called as perceived area of node. The authors considered this perception quality of node and calculated perceived probability of each node as shown in Eq. (38).

$$p = \begin{cases} 1 & d(n_i, M_j) < A_s - \omega \\ e^{-\theta d_i} & A_s - \omega < d(n_i, M_j) < A_s + \omega \\ 0 & d(n_i, M_j) > A_s + \omega \end{cases} \tag{38}$$

where $d_i = d(n_i, M_j) - (A_s - \omega)$, A_s is the sensing range of the node and ω and θ are constants that assumes uncertainty of the device in sensing process and communication path loss.

d) Further, a set of nodes having probability greater than a threshold is considered as CHN eligible nodes. To select first kind CHNs, nodes with average distance probability are considered. Also perception area with respect to each mean point is also calculated using Equation (39).

$$NP(M) = \sum_{i=1}^{N} p(n_i, M) = \sum_{i=1}^{N} e^{-\theta d_i} \tag{39}$$

where $NP(M)$ is the perception area covered by Mth mean point.

e) To select a CHN with higher area proximity, every CHN eligible node calculates perception contribution as,

$$PC(n_i, M) = \frac{p(n_i, M)}{NP(M)} \tag{40}$$

where $PC(n_i, M)$ is the perception ability of ith node with respect to Mth mean point. The above relation provide minimum ratio for a node with higher probability value. Thus, node with minimum PC value is selected as CHN of that cell.

f) Second stage of CHN is selected based on survival time estimation of the nodes when first kind of CHN dies out. A probability based on energy ratio is calculated as,

$$p = p(n_i, M)\left[1 - \frac{E_{average} - E_{residual}}{E_{average}}\right] \tag{41}$$

g) The steady state phase for data transmission is same as LEACH protocol.

This algorithm assumes uniform distribution of nodes in the network. Thus, achieves better network lifetime then compared conventional algorithms [11] and [18]. Drawback of this method is every cell fails to elect a CHN as number of alive nodes decreases. This leads many clusters uncovered and uncommunicated with the sink. Table-1 compares typical single hop clustering protocols. The notations used in Table-1 are as follows:

HO - homogeneous, HE - heterogeneous, DIS - distributed, CEN - centralized, NU - non uniform, U - uniform.

Table-1. Comparison of single hop clustering protocols.

Protocol	**Network type**	**Clust-ering type**	**Cluster formati-on**	**CHN selection**	**CHN distri-bution**	**Node density** (nodes/m^2)	**Uniform energy distribution**	**Death rate after FND**	**Scalable network**
LEACH [11]	HO	DIS	Dynamic	Random	NU	0.01	Low	Slow	No
PEGASIS [14]	HO	DIS	Chain	Random		0.01	Low	Slow	No
LEACH-E [15]	HO	DIS	Dynamic	Energy		0.005	Low		
LEACH-C [16]	HO	CEN	Dynamic	Energy	U	0.01	High	Slow	No
DEEC [18]	HE	DIS	Dynamic	Energy	U	0.01	Low	Fast	No
DDEEC [19]	HE	DIS	Dynamic	Energy		0.01	Low	Fast	No
ALEACH [20]	HO	DIS	Dynamic	Energy	NU	0.0001	Low	Fast	No
NCACM [21]	HE	DIS	Dynamic	Energy, distance	U	0.002	High	Slow	No
SECA [22]	HO	CEN	Dynamic	Distance	U	0.003	High	Fast	
Clustering for IoT [9]	HO	DIS	Dynamic	Distance	U	0.025	High		
Dynamic CHN [23]	HO	DIS	Static	Energy, distance	U	0.025	High		No

3.2 Multiple hop clustering protocols

Multiple hop is the technique in which data from source is transferred to final destination through one or many intermediator nodes.

HEED [25]

This algorithm is popularly known as HEED (Hybrid Energy Efficient Distributed). It was implemented for two-level hierarchy to transfer data from sensor nodes to sink. It uses inter cluster and intra cluster radius to adjust the transmission distance. A network operation interval is introduced to optimize resource usage according to network density. Number of iterations are limited so that protocol stops in finite number of rounds. Load of the network can be balanced by controlling the network architecture and topology. Following are the steps involved in algorithm.

a) The nodes are assumed to have number of transmission power levels. These levels are set via standard ioctl() system call [24]. The timings required for HEED operation is divided into two. One is clustering interval and other is network operation time interval. Network operation time interval is the interval between end of clustering interval and start of next clustering interval.

b) CHN selection is done probabilistically for initial set of CHNs. When a node falls within the range of more than one CHN, a tie case occurs for a node to select its CHN. The secondary parameter which is the function of cluster power level is used to break the ties. The cluster range is determined by transmission power level used for cluster formation announcement. This level is called cluster power level. The lowest power level with respect to hardware specification of device is chosen as cluster power level. In this algorithm, higher power level are used for inter cluster communication. Thus, higher power level must cover two or more clusters otherwise; this algorithm is inapplicable with respect to power adjustment.

c) Here, two cases are considered. If power level variation is not allowed, then secondary parameter is directly proportional to node degree. Otherwise, it is inversely proportional to node degree. The latter case is suitable for dense clustering. The above direct proportional relation assures that node joins a CHN with minimum degree to distribute CHN load equi-probably. And the otherwise relation assures that node joins with CHN with maximum degree to create dense cluster.

d) This algorithm is implemented using later case where power variation is allowed and secondary parameter is inversely proportional to power level. The minimum power level of nodes in a cluster is denoted by PL. The average minimum power of N nodes is given as,

$$W = \frac{\sum_{i=1}^{N} PL_i}{N} \qquad (42)$$

This value is the minimum average power of cluster required to communicate with its CHN.

e) Clustering is triggered every unit time (clustering interval + network operation interval). At initial, 5% of total nodes are selected as CHNs. Every node sets its probability to become CHN as,

$$TH = \emptyset \frac{E_{residual}}{E_{initial}} \quad (43)$$

Here $\emptyset = 0.05$ and $E_{initial}$ is a constant value. The Varying parameter is $E_{residual}$. As $E_{initial}$ is in denominator, *TH* will not fall below a numerical quantity. In this algorithm, the lower limit of threshold is considered to be 0.0001.

f) All nodes calculate *TH* value.A set of eligible nodes denoted by S{} is formed. This set consist of node's secondary parameter values which, are broadcasted to all nodes. A node selects its CHN from the set S{}, which has lowest cost. If the node itself has the least cost, it will declare itself as eligible CHN.
g) After selecting eligible CHN, every node also checks *TH* value of its elected CHN candidate. If *TH* value is 1, then the node declares that node as final CHN. After completing CHN selection step, node will simply doubles its *TH* value and executes next steps.

The secondary parameter considered for tie cases of CHN selection totally depends on intra cluster communication. As a multiple hop protocol, inter cluster communication's impact is not considered. This hierarchical model does not limit number of clusters formed in the network.

BCDCP [26]

This algorithm is popularly called as BCDCP (Base station Controlled Dynamic Clustering Protocol). A dynamic clustering protocol is introduced which is controlled by sink. The aim of this algorithm is to cover large area by proper routing of data towards sink. This process executes in two phases - set up and data communication phases. Object based addressing scheme is used, where node's location and sensing functionality is known to sink apriori. Set up phase does following operations: cluster formation, CHN selection, CHN to CHN routing and schedule creation for each cluster. The steps of operation are as follows:

a) Initially, sink receives current energy status from all nodes and then it calculates average energy of all nodes. A set of nodes S{} is formed whose energies are greater than average value.
b) A splitting algorithm is executed to split area into clusters such that each cluster should have a CHN node selected from S{}. The steps of splitting algorithm are as follows:

(a) Choose two nodes from S{} with maximum distance within each other.
(b) Group the remaining nodes in the network with selected two nodes of step (a). Grouping is based on closest distance from selected nodes.
(c) Balance two groups for approximately equal number of nodes.
(d) Split the set S{} into independent set S1{} and S2{}.
(e) Repeat steps (a) to (c) separately on newly formed two sets S1{} and S2{}.

c) Finally all CHN are grouped by near node members. This forms clusters in the network. The CHNs are then selected. Using spanning tree algorithm low energy routing paths are then fixed.
d) Sink assigns scheduling ID for all nodes within a cluster. In data communication phase, CHNs gather from nodes. CHN will further aggregate collected data into single packet. This single packet is transmitted to the next hop CHN which is updated in its routing table by sink.

The simple splitting technique increase FND value of the network compared to [11] and [16]. As cluster formation and routing process is handled by sink, nodes perform energy relaxed operations. The use of spanning tree for routing makes this algorithm scalable in terms of area. The simulated results state that BCDCP saves more energy compared to [11], [14] and [16], when network area is increased.

EC [27]

This algorithm is popularly called as EC (Energy-efficient Clustering). In this algorithm, issue of heavy data traffic near sink area is tried to redistribute among nodes so as to overcome early dying of CHN near the sink. The cluster size formation depends on hop distance to sink. The algorithm works in three steps for every round of data transmission.

a) The monitoring area is divided into small regions Awith region width w. Also average region width is calculated as Φ. Nodes are categorized into these regions depending on the hop distance from sink. Nodes from a region are considered to have same hop distance. The node energy level is approximated to a common value at every round so that CHNs nearer to sink will be rotated within small clusters resulting equal rate of energy consumption.
b) In every region, to balance the cluster radius and hop distance, CHN selection probability of each node is made dependent on width of that region and node density (σ). It is calculated by Eq. (44).

$$TH = \frac{\rho_i}{\Phi w \sigma} \quad (44)$$

where ρ_i is number of CHNs in the i^{th} region.

c) The parameter called node energy level is introduced which depends on area width and node density along with the energy of node. It is shown in Eq. (45).

$$\tau(i) = \frac{E_{region}\,\Phi\, w\, \sigma}{E_{diss}(A_i)} \quad (45)$$

where E_{region} is the initial energy of the region in which node is deployed and $E_{diss}(A_i)$ is consumed energy of the region in past (i-1)th round. The node energy level is approximated such that, τ(1) = τ(2) = … = τ(k) for k regions. These values are further approximated to a constant value £ assuming dissipated energy of region is linear in nature. The approximation algorithm steps are as follows:

(a) Initialize lifetime energy level by a value £.
(b) Set $\tau(i)$ = £. This is made dependent on corresponding value of CHN probability TH.
(c) Update value of £ iteratively.
(d) Next round probabilities are calculated depending on current iteration's £ value using logical relationships as follows:

i. Calculate τ(k) = £, for new TH_k value
ii. Calculate τ(k-1) = £, using TH_k for new TH_{k-1} value
iii. Calculate τ(1) = £, using TH_2 for new TH_1 value

The above obtained new set of probabilities TH_1 to TH_k are used for next iteration. Using such approximation method, this algorithm achieves maximum equal energy distribution through the network. This makes the network operative with all nodes alive for long duration.Once a node dies, the network life drop down immediately. This leads to disconnection of a monitoring area suddenly. This makes a complete region uncaptured with not a single sensor alive to send information.

ESRPSDC [28]

This algorithm is abbreviated as ESRPSDC (Efficient Secure Routing Protocol based on Signal to noise ratio based Dynamic Clustering). It modifies Ad Hoc on demand distance vector routing. Following are the steps of execution.

a) Initialization: Sink broadcasts 'REQUEST' message to all nodes. Then sink floods level-1 packets with minimum transmission power level. Nodes receiving these packets, set their level as '1'. In similar way, sink floods level-2 packets in the network with increased transmission power level. Nodes without level, after receiving level packets set their levels accordingly. Thus, nodes are partitioned into levels. Later sink sends 'HELLO' message to all nodes. This message contains highest and lowest level numbers.
b) Energy based CHN selection: Each cluster has its cluster table (CT). It contains - cluster ID, number of active and sleep nodes, current CHN with its energy and next CHN. Initially, CHN's routing table columns are blank. A node having energy greater than threshold becomes CHN candidate. Then within the cluster, node with highest energy is elected as CHN. The next CHN is elected with second highest energy using Eq. (46).

$$TH = \frac{(p*\zeta)(U_i - d_{toS})}{1-p\left(r\ mod\left(\frac{1}{p}\right)\right)(U_i-\mathcal{L}_i)}\left(\frac{E_{residual}}{E_{initial}}\right)^{\varpi} \qquad (46)$$

where p is the desired percentage of CHN. $\zeta$is a constant value within 0 to 1. $U_i$and$\mathcal{L}_i$ are upper and lower limit of level i. d_{toS} is distance between node and sink. $\varpi$is a constant value used within the range 0 to 3.

c) Mediator node selection based on SNR: The nodes away from the CHN cannot become cluster members. Such nodes select a mediator node towards CHN based on the SNR value of the received packet. The nodes receiving 'CHN_STATUS' message from CHN sets their intra hop-count as One-hop. These nodes further broadcast their 'ONE-HOP_STATUS' message into the region. An isolated node receiving this message will set its hop count as two-hop.
d) This algorithm gives solution for inter cluster routing. Data flow starts from last level. Data packets from last level are flooded to consecutive next level until it reaches to sink. This algorithm also implements invader detection. Sink floods ID of affected nodes initially. These are further flooded in the network. Sink analyses the route patterns periodically. The repeated patterns towards a node will declare that node as an invader.

This algorithm uses flooding technique for message sending. Flooding is a wasteful method with respect to energy and bandwidth. The data of last level is broadcasted to next level by CHNs. The next level nodes receiving these packets will further aggregate the data and send to their respective CHNs. Due to this process; most of the NCHN also spent their precious energy in aggregating last level's data with self-data. This will also produce duplication of data in the network. This algorithm is unsuitable for real time application, where nodes are energy and bandwidth constraint.

SEECH [29]

This algorithm is popularly called as SEECH (Scalable Energy Efficient Clustering Hierarchy). Sensing nodes are divided into three types- CHN, NCHN and relays. The burden of CHN is reduced by selecting another energetic set of nodes called as relay nodes. These both categories are separately considered to distribute energy consumption of nodes equally to prolong the network lifetime. CHN and relays are selected based on number of neighbouring nodes and residual energy. The algorithm is executed in three phases. Start phase before first round; Set up phase at initial of every round followed by steady state phase.

A. In start phase, every node does following three operations:

(a) Calculates its distance from the sink and number of neighbours $n_i$within a radius R.
(b) Send its information to other nodes.
(c) Calculates its degree using following ratio,

$$deg_i = \frac{n_i}{\max\{n_1, n_2, \ldots, n_N\}} \qquad (47)$$

B. In set up phase, intra-cluster communication cost is avoided by considering location of each node. The suitability of a node to be a CHN is analysed from the number of neighbours. From number of neighbours, it can be analysed how a node is suitable for being CHN. The threshold probability is formulated as in Eq. (48).

$$TH = \frac{E_{residual} deg_i}{P_{ctot}}, \quad E_{residual} \geq E_{average}(1-\zeta) \qquad (48)$$

$$P_{ctot} = \frac{E_{average}\sum_{i=1}^{N} deg_i}{2C_{optimum}} \qquad (49)$$

Here $E_{average}$ value is broadcasted by CHN in previous round. $\zeta$is a number between 0 to 1. In above relation, a lower energy threshold is used so that low energy nodes will not get any chance to become CH. To select more CHN candidates, the ζ value is kept higher. Higher the value of ζ, lower is the energy threshold, $E_{average}(1-\zeta)$. So many nodes become eligible for CHN selection process.

C. If a random number generated between 0 and 1 by node is smaller than TH value of that node then the node declares itself as CHN candidate.
D. An elimination algorithm is implemented on the newly formed set of CHN candidates to select the final CHNs for that round.

a) The total number of CHN candidates are denoted by variable count. This is compared with $C_{optimum}$.
b) If count is greater than $C_{optimum}$, a node with smallest residual energy is eliminated from the list.
c) The count is decremented by one.
d) Steps (a), (b) and (c) are repeated till count is equal to $C_{optimum}$.

E. The eliminated nodes in step 4 can become relays. They were selected for CHN depending on deg value. Thus, for relay selection same deg cannot be used, as again they will get eliminated. Thus, authors considered the value $(1-deg)$ for selecting relay nodes. Eqs. (48) and (49) are used by replacing deg with $(1-deg)$. Relay selection is done by repeating steps 2 to 5.
F. Routing path is decided based on distance between CHNs and relay nodes. A CHN finds the minimum distant relay node for its data transmission.

The algorithm results in increased network lifetime. Consideration of neighbour nodes within limited distance form clusters with average transmission distance of NCHNs. Here, every CHN forwards data to relay and not to sink. This reduces transmission distance of CHN. In a particular case, CHN transmits data to its nearest relay. But the distance between that relay and sink is greater than the distance between CHN and sink. Such case leads more energy dissipation of relay due to large transmission distance. So, energy conservation of CHN is achieved but not considered for relay nodes.

FSC [30]

This algorithm is popularly called as FSC (Fan Shaped Clustering) and proposes clustering protocol for large scale. Clusters formed are fan shaped within a circular monitoring area. This algorithm also categories nodes into three types depending on its operation. They are CHN, NCHN and relay nodes. The sink is considered inside the region and at the centre.

The area is partitioned into concentric rings. The width of the layers is considered to be a constant value ϕ. The width of first layer is kept greater than ϕ to avoid hotspot issue. Every layer is then again divided angularly into fan shaped clusters. The number of clusters formed in a layer is decided by $ɳ(2i-1)$ where, i is layer number and ɳ is a natural number. ɳis selected in such a way that formed clusters are approximately squared shaped. The cluster length is calculated as in Equation (50).

$$l = \frac{R}{\sqrt{8}} \qquad (50)$$

where R is node's sensing radius. The ratio considered in Eq. (50) ensures that a node wherever it is located in the cluster can always communicate with node in neighbour cluster. As the polar coordinate system is considered, the longest distance between the nodes of the neighbouring clusters can be calculated diagonally as shown in Figure-7.

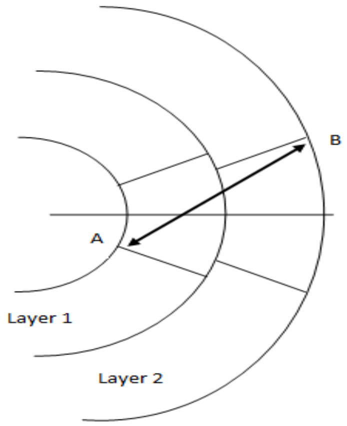

Figure-7. Fan shaped cluster with maximum distance [30].

Here points A and B are the corner points of the two neighbour clusters. From second layer onwards, the expected distance of CHN from cluster members is calculated as,

$$d_{A-B} = \sqrt{[r(i-1)]^2 + [r(i+1)]^2 - 2r^2(i-1)(i+1)\cos\left(\frac{i\pi}{i^2-1}\right)} \qquad (51)$$

The algorithm operates in following five steps:

a) Partition: Sink broadcasts partition message within the range of R. The message contains sink location, first layer's radius and other layer's width r. Each node receiving this message calculates to which cluster it belongs based on its position information and received information.

b) CHN selection: For each cluster, a circular area is defined at the central region. The nodes from this area can only become CHN. Initially, nodes from this region and with energy greater than threshold sets a back off timer T_b. A node is selected as CHN of which T_b value firstly becomes zero. Then node sends 'CHN_SELECT' message to acknowledge other nodes about selected CHN and to halt their T_b.
c) Relay selection: After CHN selection, the anew formed CHN broadcasts 'RELAY_REQUEST' message to all its cluster member. Nodes from outside of circular area are only considered for relay operation. Such nodes after receiving 'RELAY_REQUEST' message set a back off timer and wait till T_b gets to zero. The process of relay selection is same as CHN selection.
d) If the CHN drains off his energy less than threshold, a 'NEXT_CHN' message is broadcasted to select another CHN from the circular area. When the last node within circular area is CHN and its energy drains below threshold, it broadcast 'NEXT_CHN' message. After certain time interval, as it do not receive any 'CHN_SELECT' message, that CHN will now flood 'RE-CLUSTERING' message to all the nodes of the layer.
e) Re-clustering is done only on each layer and not the whole network. Each cluster is shifted by predefined angle in one specific angular direction. An angle is chosen such that rotation of cluster will not overlap previous cluster formation.

In this scheme, clustering process is not performed often. This saves energy on large scale. Shifting of clusters will allow relay nodes to become central nodes of cluster and vice versa. The newly formed outside area nodes are already low at energy. Thus, such low energy nodes might not be able to transmit data packets due to higher communication cost.

Fuzzy LEACH [31]

In this algorithm, LEACH protocol is modified by introducing multiple hop towards sink. A super CHN node is selected among all CHN to send overall network data to the sink. To accomplish super CHN selection, fuzzy inference system is used.

In this algorithm, nodes are static and sink is kept mobile. Based on the distance calculated from received signal, centrality of a CHN is determined. The centrality focuses on the location of super CHN. It must be approximately in the communication region of all the CHNs. The node with more centrality and better residual energy is chosen as super CHN.

A. In every round, a node generates random number between 0 and 1. It also calculates threshold probability of becoming CHN using Eq. (12). If random number is less than threshold then node declares itself as CHN.
B. Among all CHNs for a round, a super CHN is selected using three fuzzy descriptors

(a) Residual battery power
(b) Centrality
(c) Mobility of sink

C. The above three inputs are given to fuzzifier. Fuzzification is the process, where inputs are given some crisp (hard) values and changed into a fuzzy set. In this algorithm, Mamdani's method is used.
D. Every input is defined using three membership functions. For all inputs, two membership functions are trapezoidal and one is triangular. The output is defined with seven membership functions and all are triangular.
E. Fuzzy rules contain If-Then rules. These are defined and stored in inference engine. Total 27 rules are designed by the authors for selecting a proper node as super CHN.

$$\mathrm{SuperCH} = (\mathrm{E}_{\mathrm{residual}} - 1) + \mathrm{Mobility} + \mathrm{Centrality} \quad (52)$$

F. The inference engine applies rules on inputs and corresponding output inference is obtained. The defuzzifier converts the output fuzzy set into crisp value.

This algorithm obtains better performance due to implementing complex process of node selection using fuzzy logic. The network scenario implemented is for sink mobility. Moving sink information is typical application of precision agriculture.

JCR [10]

This algorithm is popularly called as JCR (Joint Clustering and Routing) and addresses connectivity issue. A gradient routing algorithm is proposed by the authors to achieve energy efficiency. The gradient of each node is calculated and shared among other nodes. This parameter is used to select CHN and also to route the data. The gradient is the minimum hop distance towards the sink. In this algorithm, two distances are defined. R_t is the transmission range and R_c is the cluster range. Following are the steps of operation.

a) Initially, sink floods gradient packets with $g = 0$ within R_t. The nodes receiving gradient packet, will update their gradient by incrementing the received value by one.
b) This updated gradient packet is further flooded to next farther nodes. Nodes also keep record of forward nodes and backward nodes with respect to itself. The node from which the gradient packet is received is recorded as forward node. The node from which gradient packet acknowledgement is recorded as backward node. This process determines hop distances for each node.
c) All nodes calculate a back off timer T_b as follows,

$$T_b = \begin{cases} (G-g)+\propto\left(\frac{E_{initial}-E_{residual}}{E_{initial}}\right)+(1-\propto)\left(\frac{N_F-N_f}{N_F}\right), & g>1 \\ (G-g)+\propto\left(\frac{E_{initial}-E_{residual}}{E_{initial}}\right)+(1-\propto)\left(\frac{N_b}{N_B}\right), & g=1 \end{cases} \quad (53)$$

The ratio of forward nodes results in smaller T_b value. This increases the chance of node to become CHN. $\propto$is used as constant factor to adjust T_b. G is the highest gradient in the network. $N_f$is number of forward nodes of each node. $N_b$is number of backward nodes of each node. $N_F$andN_B are the maximum number of forward and backward nodes respectively.

d) CHN selection: At start of each round, every node will set its T_b and start decrementing by one. Here either of the two situations will occur until T_b becomes zero. In first situation, a node may receive 'JOIN' message and in second situation, it may not receive any message. If former situation occurs then that node acknowledges and becomes cluster member. For later one, the node will declare itself as CHN and broadcasts 'JOIN' packets within its R_c.

e) Route selection: A node receiving 'JOIN' packet will check a flag bit and the gradient value. This flag bit indicates forward route or backward route from the node. If the gradient of receiving node is greater than the transmitting node's gradient, it will set flag value to one otherwise zero. If flag of a node is zero, that node's ID is stored in routing table. This node acts as the forward node to pass the data towards sink.

CHN selection process chooses forward nodes as its next hop node. But it does not check for current energy level. The occurrence of imbalanced energy consumption among CHNs is not considered. The factor $\propto$ plays additional role without technical correlation with performance parameter. Since the nodes are static, their gradient can be calculated. For mobile network, this algorithm fails because gradient keeps on varying instantly. For large scale network, flooding is not an ideal solution as it will result congestion in network and might lead to high data packet loss.

Table-2. Comparison of multiple hop clustering protocols.

Algorithm	**Network type**	**Clust-ering type**	**Cluster formati-on**	**CHN selection**	**CHN distri-bution**	**Node density** (nodes/m^2)	**Uniform energy distribution**	**Death rate after FND**	**Scalable network**
HEED [25]	HE	DIS	Dynamic	Energy	U	0.0002	Low	Slow	Yes
BCDCP [26]	HO	CEN	Dynamic	Energy	NU	0.05	High	Fast	Yes
EC [27]	HO	DIS	Dynamic	Distance	U	0.025	High	Fast	No
ESRPSDC [28]	HE	DIS	Dynamic	Energy, SNR	NU	0.0005	Low		Yes
SEECH [29]	HO	DIS	Dynamic	Distance	NU	0.025	High		No
FSC [30]	HO	DIS	Dynamic	Random	U	0.0073	High	Slow	Yes
Fuzzy LEACH [31]	HO	DIS	Dynamic	Random	NU	0.004	High		No
JCR [10]	HO	DIS	Dynamic	Energy, distance	U	0.01	High		Yes

4. CONCLUSIONS

This paper reviewed clustering protocols based on scalable techniques. The protocols are classified based on single hop and multiple hop clustering approach. Single hop clustering is energy efficient for homogeneous network whereas multiple hop holds good for both homogeneous and heterogeneous networks. The energy distribution based on distance leads to better clustering and increased node density. But many of the techniques are not scalable though their energy efficiency is improved remarkably.

For IoT applications, a large sensor network with an ability to perceive and aware the monitoring or control system in peak timing is required. Innovative techniques are needed to make network adaptive in nature as nodes may not be static in many applications. Nodes in the field may change their position or sensing parameter as per the realistic situation. To save energy, nodes may form clusters dynamically to reduce transmission distance towards the sink. Most of the algorithms fail to sustain large number of nodes due to lack of proper clustering process. Thus, there is still scope for researchers to improve energy efficiency when node density rises in the heterogeneous environment of IoT. It is inferred that for IoT based system, a protocol must incorporate a proper CHN selection, timely cluster formation and routing the data in large sensor network without latency.

REFERENCES

[1] D. Evans. 2011. The Internet of Things - How the Next Evolution of the Internet Is Changing Everything. White paper-Cisco IBSG.

[2] Disruptive Civil Technologies- Six Technologies with potential impacts on US interests out of 2025.2008. National Intelligence Council, Conference report. United States of America.

[3] L. Atzori, A. Iera and G. Morabito. 2010. The Internet of things: A survey. Computer Networks. 54: 2787-2805.

[4] J. A. Stankovic. 2014. Research directions for the Internet of things. IEEE Internet of Things Journal. 1(1): 3-9.

[5] J. Gubbi, R. Buyya, S. Marusic and M. Palaniswami. 2013. Internet of things (IoT): A vision, architectural elements, and future directions. Journal of Future Generation Computer Systems. 29(7): 1645-1660.

[6] F. Zhao and L. J. Guibas. 2004. Wireless Sensor Networks: information processing approach. Morgan Kaufmann Publishers.

[7] T. S. Rappaport. 2001. Wireless Communication-Principles and Practice. Prentice Hall.

[8] J.Y. Cheng, S. J. Ruan, R. G. Cheng and T.T. Hsu. 2006. PADCP: Power aware dynamic clustering protocol for wireless sensor network. Proceedings of IFIP WOCN. 1-6.

[9] J. Y. Chang. 2014. A distributed cluster computing energy-efficient routing scheme for Internet of things systems. Wireless Personal Communications. 82(2): 757-776.

[10] Z. Xu, L. Chen, C. Chen and X. Guan.2016. Joint clustering and routing design for reliable and efficient data collection in large-scale wireless sensor networks. IEEE Internet of Things Journal. 3(4): 520-532.

[11] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan. 2000. Energy-efficient communication protocol for wireless microsensor networks. Proceedings of the 33rd Hawaii international conference on system science. 1-10.

[12] S. Yinbiao. 2014. Internet of Things: Wireless Sensor Networks. IEC-white paper.

[13] C. Perera, A. Zaslavsky, P. Christen and D. Georgakopoulos. 2014. Sensing as a Service Model for Smart Cities Supported by Internet of Things. Transactions on Emerging Telecommunications Technologies. 25(1): 81-93.

[14] L. Stephanie and C. S. Raghavendra. 2002. PEGASIS: Power efficient gathering in sensor information systems. Proceedings of IEEE Aerospace conference. 3.

[15] M. J. Handy,M. Haase and D. Timmermann. 2002. Low energy adaptive clustering hierarchy with deterministic cluster-head selection. Proceedings of fourth IEEE conference on mobile and wireless communications network. pp. 368-372.

[16] W. R. Heinzelman, P. Chandrakasan and H. Balakrishnan. 2002. An application-specific protocol architecture for wireless microsensor networks. IEEE Transactions on Wireless Communications. 1: 660-670.

[17] G. Smaragdakis, I. Matta and A. Bestavros. 2004. SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks. Second International Workshop on Sensor and Actor Network Protocols and Applications.

[18] L. Qing, Q. Zhu and M. Wang. 2006. Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. Computer Communications.

[19] A. Rathee, I. Kashyap and K. Choudhary. 2015. Developed distributed energy-efficient clustering algorithm based on fuzzy logic approach for optimizing energy management in heterogeneous WSNs. *International Journal of Computer Applications*. 115(17): 14-19.

[20] Md. S. Ali, T.Dey and R. Biswas. 2008. ALEACH: Advanced LEACH Routing Protocol for Wireless Microsensor Networks. Fifth International Conference on Electrical and Computer Engineering. 20-22.

[21] S. Babaie,A. K. Zadeh and M. G. Amiri. 2010. The new clustering algorithm with cluster members bounds for energy dissipation avoidance in wireless sensor network. Proceedings of computer design and applications. 613-617.

[22] J. Y. Chang and P. H.Ju. 2012. An efficient cluster-based power saving scheme for wireless sensor networks. EURASIP Journal on Wireless Communications and Networking. pp. 1-10.

[23] D. Jia,H. Zhu, S. Zou and P. Hu. 2016. Dynamic cluster head selection method for wireless sensor network. IEEE Sensors Journal. 16(8): 2746-2754.

[24] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D.E. Culler and K.S.J. Pister. 2000. System Architecture Directions for Networked Sensors. Architectural Support for Programming Languages and Operating Systems. pp. 93-104.

[25] O. Younis and S. Fahmy. 2004. HEED: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Transactions on Mobile Computing. 3(4): 366-379.

[26] S. D. Muruganathan, D. C. F. Ma, R. I. Bhasin and A. O. Fapojuwo. 2005. A centralized energy-efficient routing protocol for wireless sensor networks. IEEE Communications Magazine. 43(3): 8-13.

[27] D. Wei, Y. Jin, S. Vural, K. Moessner and R. Tafazolli. 2011. An energy-efficient clustering solution for wireless sensor networks. IEEE Transactions on Wireless Communications. 10(11): 3973-3983.

[28] S. Ganesh and R. Amutha. 2013. Efficient and secure routing protocol for wireless sensor networks through SNR based dynamic clustering mechanisms. Journal of Communications and Networks. 15(4): 422-429.

[29] M. Tarhani, Y. S. Kavian and S. Siavoshi. 2014. SEECH: Scalable energy efficient clustering hierarchy protocol in wireless sensor networks. IEEE Sensors Journal. 14(11): 3944-3954.

[30] H. Lin, L. Wang and R. Kong. 2015. Energy efficient clustering protocol for large-scale sensor networks. IEEE Sensors Journal. 15(12): 7150-7160.

[31] P. Nayak and A. Devulapalli. 2016. A fuzzy logic-based clustering algorithm for WSN to extend the network lifetime. IEEE Sensors Journal. 16(1): 137-144.