An Implementation of Grouping Nodes in Wireless Sensor Network Based on Distance by Using k-Means Clustering

Rizqi Fauzil Azhar[1], Ahmad Zainudin[2], Prima Kristalina[3], Bagas Mardiasyah Prakoso[4], and Ni'am Tamami[5]
[1−5]Department of Electronic Engineering, Politeknik Elektronika Negeri Surabaya
Surabaya 60111, Indonesia
Email: [1]fauzilrizqi@gmail.com, [2]zai@pens.ac.id, [3]prima@pens.ac.id, [4]bagasmardiasyah@gmail.com, [5]niam@pens.ac.id

***Abstract*—Wireless Sensor Network (WSN) is a network consisting of several sensor nodes that communicate with each other and work together to collect data from the surrounding environment. One of the WSN problems is the limited available power. Therefore, nodes on WSN need to communicate by using a cluster-based routing protocol. To solve this, the researchers propose a node grouping based on distance by using k-means clustering with a hardware implementation. Cluster formation and member node selection are performed based on the nearest device of the sensor node to the cluster head. The k-means algorithm utilizes Euclidean distance as the main grouping nodes parameter obtained from the conversion of the Received Signal Strength Indication (RSSI) into the distance estimation between nodes. RSSI as the parameter of nearest neighbor nodes uses log-normal shadowing channel modeling method that can be used to get the path loss exponent in an observation area. The estimated distance in the observation area has 27.9% error. The average time required for grouping is 58.54 s. Meanwhile, the average time used to retrieve coordinate data on each cluster to the database is 45.54 s. In the system, the most time-consuming process is the PAN ID change process with an average time of ±14.20 s for each change of PAN ID. The grouping nodes in WSN using k-means clustering algorithm can improve the power efficiency by 6.5%.**

***Index Terms*—Wireless Sensor Network, Cluster-Based, k-Nearest Neighbour**

I. Introduction

WIRELESS Sensor Network (WSN) is a network of devices consisting of a large number of node sensors, computing or data processing tools, and communication tool to send and receive data. Generally, WSN consists of three important parts including sensor nodes on the network used for sensing, base station or sink placed outside the observation area, and user as a data organizer on the sink. The communication and data transmission on each sensor node use radio waves with a specific frequency. In the application, the distance between sensors is not too far. Meanwhile, the distance between the sink with sensor nodes can be far apart.

Every node on WSN can collect data and establish communication with another node sensor. For some WSN implementation, the main problem is the power efficiency of the sensor node [1]. One way that can be used to solve the problem is by applying a cluster-based protocol. The cluster-based protocol can perform grouping of multiple sensor nodes with various parameters [2]. WSN sensor nodes, which have a shorter distance to the nearest cluster head, are in the same group. The main task is to collect the data from the target area and pass them to the cluster head. After that, the cluster head forwards the collected information towards the base station [3, 4]. By using k-means algorithm with Euclidean distance as the main grouping parameter derived from the Received Signal Strength Indication (RSSI), the conversion can group nodes with a close range into the same cluster. Meanwhile, the sensor nodes with a far distance apart will be grouped into the other clusters.

Reference [5] analyzed the power effect on the WSN using k-means clustering-based routing protocol by considering an optimal fixed packet size based on radio parameter and channel condition. The k-means method had been used to create some of the clusters. The radio parameter and channel condition determined the energy consumption. The calculation of energy consumption and the average distance of the corresponding cluster nodes toward base station had been defined as cluster

Received: Apr. 27, 2018; received in revised form: Sept. 11, 2018; accepted: Oct. 3, 2018; available online: Dec. 5, 2018.

head selection criteria for that particular round.

Moreover, Ref. [6] proposed a LEACH-Centralized protocol to minimize the disadvantages of the LEACH by applying the k-means algorithm. The proposed system was implemented by partitioning the network into several clusters in the first step using k-means method. In the second step, the cluster which was produced in previous work applied the LEACH-C protocol to extend the lifespan of the WSN. The performances of the system such as average end-to-end delay, packet delivery ratio, average energy consumption, average throughput, and control routing overhead had been evaluated by NS2 simulator. The result showed that the LEACH-C with k-means could achieve better performance.

Reference [7] evaluated the performance of an enhancement cluster-based routing protocol in WSN. They proposed two routing protocol schemes and compared them. The first way was applying the cluster-based k-means routing scheme. For the second way, they used an improved k-means approach with the generation of balanced clusters and selected the nearest cluster member for its corresponding cluster. The distribution of the load equitably could reduce the energy consumption among the cluster head. Moreover, the main objective by applying of the balanced cluster could improve the network lifetime that was shown in the simulation results.

In this research, the researchers propose a grouping node based on distance using k-means clustering with a hardware implementation. The implementation uses an Arduino MCU board with XBee RF and GPS modules to get the position information of the node. The researchers create the clusters in which the nodes with a shorter distance to the nearest cluster head are in the same group. The cluster head has been determined before. After the cluster is formed, the cluster head will arrange data transmission from all nodes in the cluster to the base station. The sent cluster head in the collected data uses LEACH protocol continually. The selection of the new cluster head is conducted periodically when the residual energy is in the minimum threshold. By applying k-means in LEACH protocol, it can increase device performance and efficiency of energy consumption.

II. Cluster-Based Protocol with k-Means Approach

A. Cluster-Based Protocol in Wireless Sensor Network (WSN)

In the cluster-based protocol, nodes are grouped into clusters. Every cluster has a selection of cluster head which is based on different algorithms. The cluster heads are used for higher-level communication and reducing the traffic overhead. The cluster-based protocol can save energy on node devices. This protocol uses Energy Aware Distributed Clustering (EADC) to select a cluster head by determining the ratio between the average residual energy of neighbor nodes and the residual energy of the node itself. Thus, the cluster-based protocol can decrease power consumption on the end nodes [8]. The researchers propose a cluster-based routing protocol with the multi-hop system that the cluster heads select neighbor nodes, collect node data and report them to sink. Then, end nodes will report data to each cluster head and are forwarded to sink.

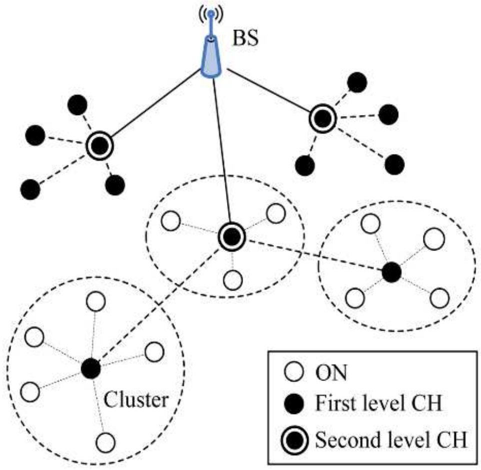

Fig. 1. Data transmission on the cluster-based protocol.

Clustering may be extended to more than two level communications that have the same concepts of communication at every level. There are many advantages in routing hierarchy usage. It can reduce the size of routing tables by providing better scalability. The cluster head is used as the main communication node in the cluster, so it has the highest number of traffic. However, by using cluster-based protocol, the overhead traffic will be decreased. The cluster head takes responsibility for coordinating activities within the cluster and forwarding data between clusters. It is shown on Fig. 1.

In the scheme in Fig. 1, cluster-based protocol reduces energy consumption and extends the lifetime of the network. Clustering is increased two levels and has a communication hierarchy. It has the high delivery ratio and scalability and can balance the energy consumption. The nodes around the base station or cluster head will deplete their energy sources faster than the other nodes. To build a cluster, it can implement k-means algorithm to classify the end node to the cluster head based on the nearest distance [9].

B. k-Means Technique Based on Distance Clustering

In WSN, clustering has been widely observed by the research community to solve the energy and lifetime problems. WSN has indeterminate distribution nodes characteristics. It is required to use the unsupervised classification of nodes to create a cluster. k-means is one the unsupervised learning to address clustering issue [6]. The k-means use a predefined number of cluster (assume k clusters) in the first step. The nodes send residual energy information to the base station and select the cluster head with maximum residual energy for each cluster.

The cluster members are the closest nodes from the cluster head. The neighbor distance is determined by using the Euclidian distance method. The Euclidean distance tests the size that can be used as an interpretation of the proximity of the distance between two objects. The Euclidean distance formula shows on Eq. (1) with X_{ik} as X on data training, X_{jk} as X on data testing, and m as maximum number of data. The Eq. (1) is as follows:

$$d = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}. \tag{1}$$

In Eq. (1), if the formula produces a large value, it will further extend the level of similarity between two objects. However, if the formula produces small value, it will be closer to the level of similarity between objects. The application of the k-means algorithm in WSN system is shown in the flowchart in Fig. 2.

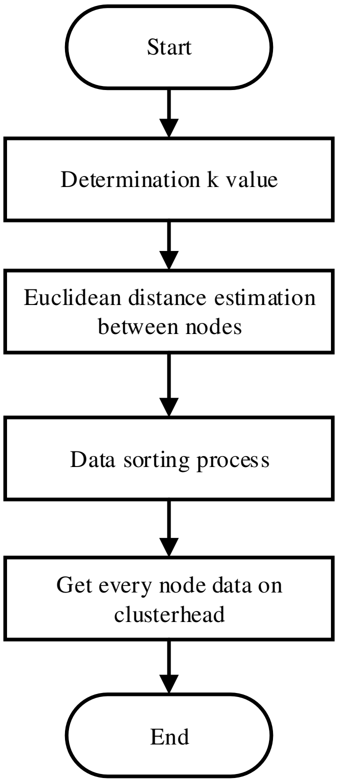
Start
Determination k value
Euclidean distance estimation between nodes
Data sorting process
Get every node data on clusterhead
End

Fig. 2. The procedure of k-means clustering in WSN system.

k-means can increase energy efficiency, scalability, and robustness to dynamic network topologies. Specifically, an efficient k-means algorithm should minimize the number of sensor nodes involved with query execution. It is because the radio operations dominate the energy consumption of sensor networks, minimize the total amount of data transmitted, and distribute the responsibility of a query execution to all involved sensor nodes evenly [8].

C. Distance Estimation Using Received Signal Strength Indication (RSSI)

RSSI is one of the most commonly used approaches for estimating the distance between nodes. It is because almost every node the can analyze the signal strength of received messages. To use RSSI as a distance calculation approach, it is necessary to calculate exponent path loss value in an environment. The free space path loss equation computes PL (d_0). The d_0 value is the reference distance that determines 1 meter. It should be in the far-field of the transmitting antenna, although the small distances are used according to any practical distance in mobile communication channels. Path loss exponent determination using a log-normal shadowing model is in Eqs. (2)–(4) [9]. The equations are as follows:

$$P_L = P_{L0} + 10 \cdot n \cdot \log \frac{d}{d_0} + X_\sigma, \tag{2}$$

$$-P_{RX} = -P_{RX0} + 10 \cdot n \cdot \log \frac{d}{d_0} + X_\sigma, \tag{3}$$

$$n = \frac{P_{RX0} - P_{RX}}{10 \log \frac{d}{d_0}} - X_\sigma. \tag{4}$$

To estimate the distance between nodes, path loss coefficient obtained from Eqs. (2) and (4). It can be performed with substitute Eqs. (2) and (4) with Eq. (3) given by Eqs. (5) and (6). The equations can be seen as follows:

$$-P_{RX} = -P_{RX0} + 10 \cdot n \cdot \log \frac{d}{d_0}, \quad \text{and} \tag{5}$$

$$d = d_0 \cdot 10^{\left(\frac{P_{RX0} - P_{R_X} - X_\sigma}{10n}\right)}, \tag{6}$$

where n denotes the path loss coefficient, P denotes the path loss (dB), P_{RX0} denotes the received power on

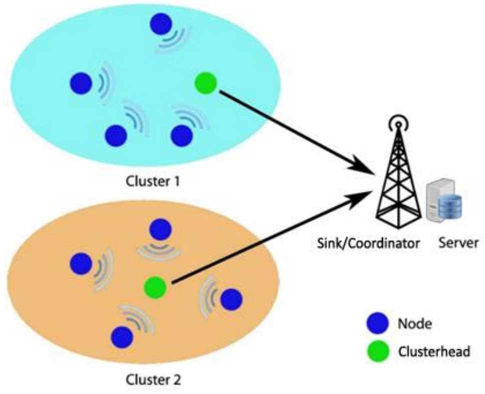

Fig. 3. Nodes communication scheme.

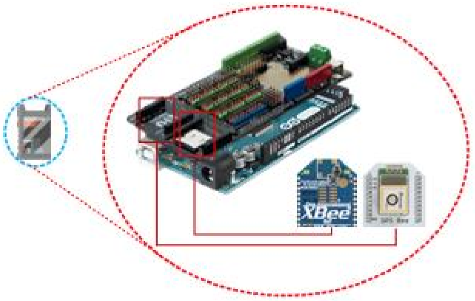

Fig. 4. Hardware design.

TABLE I
THE EXAMPLES OF RECOGNIZED USER INPUT VARIATIONS IN THE CHATBOT’S LOG.

Message	Tx	Rx	Function
STARTCH1	Sink	CH01	Start Clustering
TXDATACH1	Sink	CH01	Request cluster data
TXDATACH2	Sink	CH02	Request cluster data
MSGCH01	CH	End node	Request distance estimation
SLATLON	CH	End node	Request position data

1	2	3	4	5
#	*Node* ID	#	Distance	;

Fig. 5. The distance of estimation frame.

1	2	3	4	5	6	7
#	*Node* ID	#	Latitude	#	Longitude	;

Fig. 6. The position of the data frame.

reference distance d_0 (dBm), P_{RX} denotes the received power on d (dBm), D denotes the distance between nodes (m), d_0 denotes the reference distance (m), and X_σ denotes the standarized Gaussian distributed random parameter (dB).

D. Node Design

WSN nodes are classified to end nodes and cluster heads placed in the observation area. Meanwhile, the sink section is on the other side that has control over the node in the coverage area. It is by sending a message to build cluster and request node data as illustrated in Fig. 3

The node is designed using Arduino mega 2560 as a microcontroller. Moreover, an XBee pro RF module is used as a wireless module for communication with other nodes. Then, GPS Bee parses latitude and longitude. In the implementation stage, every node will be placed on the climbers as illustrated in Fig. 4.

E. Data Frame Design

The data frame design manages communication between the sink, cluster head, and end nodes. Commonly, the message is designed as an instruction message and data frame message. The entire component will work the next task when it receives an instruction message and sends data frames as the feedback. Instruction message designed in this research is shown in Table I.

The instruction message “MSGCH01” used by CH01 to request the connected node distance data. The feedback message contains the distance of each node to CH01. The distance data messages are shown in Fig. 5. The “SLATLON” instruction messages are used by CH01 and CH02 to request the position data for each connected end node to each cluster head. Then, it will generate a data message containing the latitude and longitude of each node. The coordinate data message is shown in Fig. 6.

III. RESULTS AND DISCUSSION

A. Received Signal Strength Indication (RSSI) Measurement

RSSI measurement is performed to obtain path loss exponent characteristic on the communication channel in various environments. Data collection is performed using two XBee nodes with 9600 bps baud rate and free space loss environment. To collect RSSI data, the researchers need to observe P_{RX0} and P_{RX} data on the observed environment. Measurement of P_{RX0} is performed by splitting two XBee nodes with 1 meter of distance on 3 locations in the observed environment. On P_{RX} measurement, distances between XBee are variated from 2100 meters with 2 meters of increasement. The P_{RX0} and P_{RX} measurement methods are shown in Fig. 7.

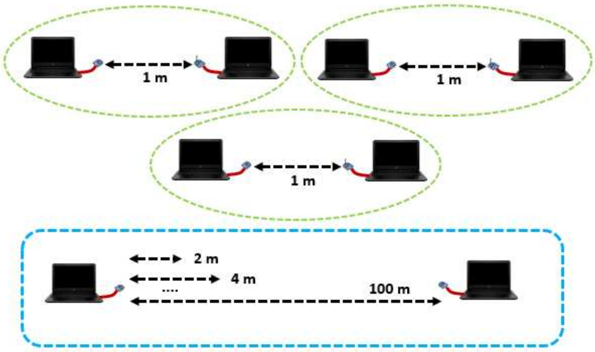

Fig. 7. The P_{RX0} and P_{RX} measurement method.

TABLE II
THE RESULT OF P_{RX0} MEASUREMENT.

Test	RSII		
	1	2	3
	(dBm)	(dBm)	(dBm)
1	−35	−37	−35
2	−35	−37	−35
3	−37	−37	−35
4	−37	−37	−35
5	−36	−37	−35
6	−35	−37	−35
7	−35	−37	−35
8	−35	−37	−35
9	−37	−37	−35
10	−36	−37	−35
P_{RX0}	−35.8	−37	−35
Average P_{RX0}		−35.9333	
$X\sigma$	0.1333	−1.0667	0.9333
Average $X\sigma$		2.36848×10^{-15}	

P_{RX0} measurement results in 1 meter of reference as shown in Table II and P_{RX} measurement is shown in Fig. 7. In Table II, the average measurements of reference signal strength (P_{RX0}) are −35.8 dBm. In addition, Table II shows that the average of normalized values is 2.368×10^{-15}. The average normalized value is the variable value of zero. It means that Gaussian standard deviation ($X\sigma$) is derived from the different reference signal strength value (P_{RX0}) and the average reference signal strength value (P_{RX0} average). Since the value is very small (close to zero) and has no effect in calculating the signal strength, the standard deviation value in the observation area can be eliminated.

Figure 8 shows that when the distance is 2 to 20 meters, the P_{RX} value has a continuous decline, but it has generally greater distance generating lower RSSI value. At the next distance, the measurement results show a fluctuating value even though the received signal strength (P_{RX}) value tends to decrease. This is caused by environmental conditions that there are many trees, which have the potential to scatter the signal. In addition, the level of precision of the device in reading the signal strength also affects the measurement results.

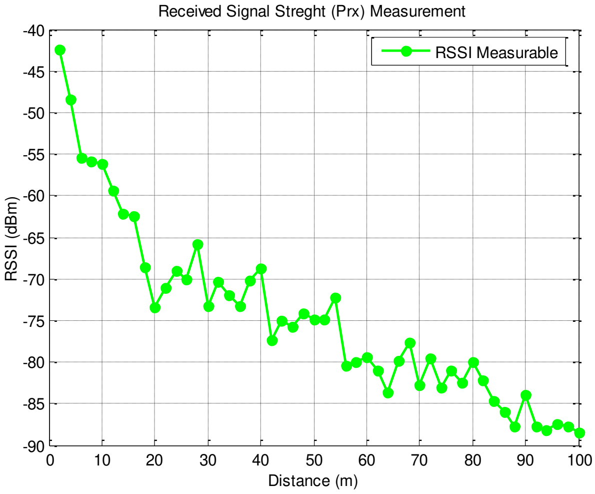

Fig. 8. The result of P_{RX} measurement.

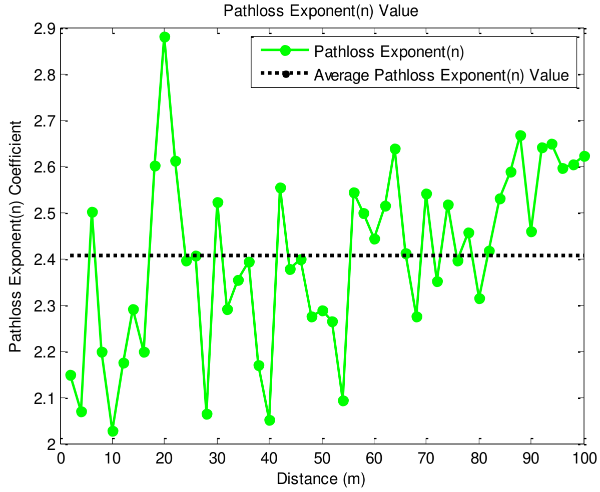

Fig. 9. The distribution chart of path loss exponent.

B. Path Loss Exponent (n)

Path loss exponent is used to estimate the distance between nodes with Eq. (4). Path loss exponent value is influenced by P_{RX0}, P_{RX}, and zero-mean Gaussian which is distributed $X\sigma$ randomly. Path loss exponent in various distances fluctuates the value. The average value of the overall exponent path loss value is 2.405. The distribution chart of path loss exponent is shown in Fig. 9.

From the observation in Fig. 9, the result of path loss exponent average value is not affected by the distance between nodes. However, it is affected by the comparison of P_{RX0} and P_{RX} on the entire distance test. Due to the fluctuating value, the path loss exponent data on

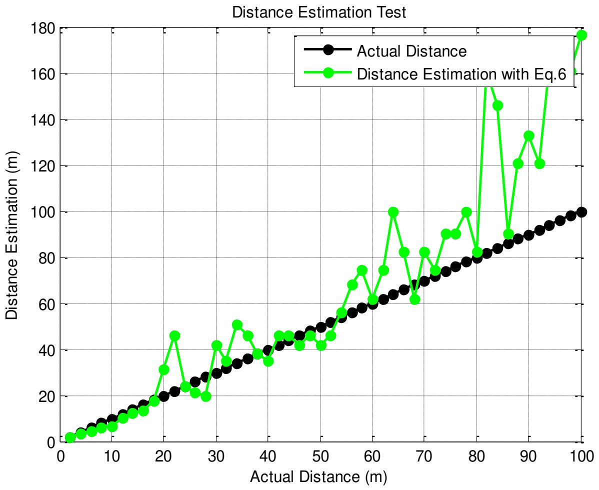

Fig. 10. The result of distance estimation.

an environment is represented by 2.405 as the average value of the observation area.

C. Distance Estimation

The distance estimation is performed by entering the calculation program on Eq. (6) to every node. This equation converts the RSSI response into distance estimation. This scenario allows the cluster head to get information about the distance between the other end node effects for distance estimation process. The results of the distance estimation process are contained in Fig. 10.

From the observation in Fig. 10, it is shown that distance estimation value is below 20 meters and has a small error percentage compared to other distances. It is caused by P_{RX} measurement values which are above 20 meters and have high deviation value. In Fig. 6, in the distance estimation process, the smallest error is 38 meters with 0.62%, and the worst error is 22 meters with 104%. Thus, the result of distance estimation has 27.9% of average error. The deviation value is affected by XBee factor which has a maximum range of 100 meter of transmission on the urban environment. As a result, the greater distance from that coverage will have unstable estimation value.

D. k-Means Time Computation to Build Cluster

Algorithm time test is performed by calculating the specific time of node to build cluster and transmit nodes data to the server. The time is required for build cluster. The required time for node data transmission can be shown in Tables III and IV

TABLE III
THE RESULT OF BUILD CLUSTER TIME.

Test	Time (ms)
1	58550
2	58543
3	58539
4	58542
5	58540
6	58548
7	58538
8	58532
9	58560
10	58542
Average	58543

TABLE IV
THE RESULT OF DATA TRANSMISSION TIME.

Test	Time (ms)
1	44382
2	46077
3	45910
4	45475
5	45087
6	42610
7	43333
8	41457
9	45766
10	44347
Average	45541

In Tables III and IV, the average total time used for building cluster is 58543 ms. To get coordinate data from GPS until the data is received on the server, it needs 45540 ms averagely. The data transmission process has faster time compared to building cluster process because there are two times in the PAN ID change process. The change takes 14000 ms, so it is very influential on the system.

E. Quality of Service (QoS) Test

Quality of service (QoS) allows the system to calculate throughput and Packet Loss Ratio (PLR) on WSN data transmission. QoS investigates the rates of successful message delivery over a network. Meanwhile, PLR is for the lost message. The distance estimation is performed by sending messages with 10 bytes, 55 bytes, and 135 bytes of packet data with standard IEEE header 802.15.4 protocol. The throughput calculation on various packets is shown in Fig. 11. The researchers conclude that the greater packet has greater throughput.

PLR is performed by sending packets on the transmitter for receiver node and observing the number of packet loss on the entire process. The results of the PLR test on WSN nodes are contained in Table V. It shows that there is no packet loss on frame I and frame II with 128 bits and 448 bits. However, on frame III, it is with 1128 bits. Packet losses have occurred in 3 out of 10 data experiment. Packet loss happens when the sent packets have big size over a maximum bit. It can be sent by XBee pro. If there are packet losses of

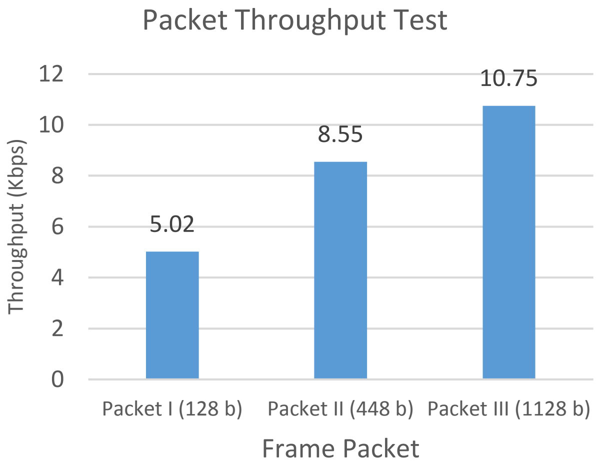

Fig. 11. Throughput on frame transmission.

TABLE V
THE RESULT OF DATA TRANSMISSION TIME.

Test	Packet I (128 b) PLR(%)	Packet II (448 b) PLR(%)	Packet III (1128 b) PLR(%)
1	0	0	0
2	0	0	0
3	0	0	21.2
4	0	0	0
5	0	0	21.2
6	0	0	0
7	0	0	0
8	0	0	21.2
9	0	0	0
10	0	0	0
Average	0	0	6.36

TABLE VI
THE RESULT OF DATA TRANSMISSION TIME.

Scenario	Voltage (V)	Current (mA)	Power (W)
Cluster	5.321846	197.1388	1.085686
Non-Cluster	5.55418	203.4893	1.161908
Cluster Efficiency	4.1%	3.1%	6.5%

data transmission, the next data transmission will have packet loss too.

F. Power Efficiency with k-Means LEACH Protocol

Power efficiency testing is to analyze battery power requirements of the node. It consists of Arduino mega, XBee pro, and GPS. The power efficiency test compares power efficiency on a cluster node with not on a cluster node. Power efficiency is performed by getting battery voltage and battery current by using INA 219 sensor. The results of the power efficiency test are contained in Table VI. The graphs of current, voltage and power measurement results are shown in Figs. 12–14

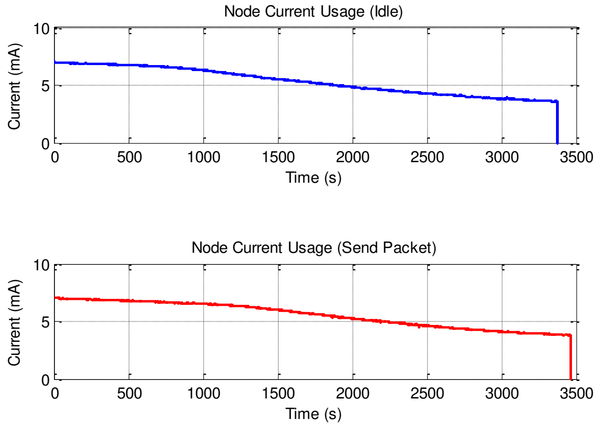

Fig. 12. Nodes current usage.

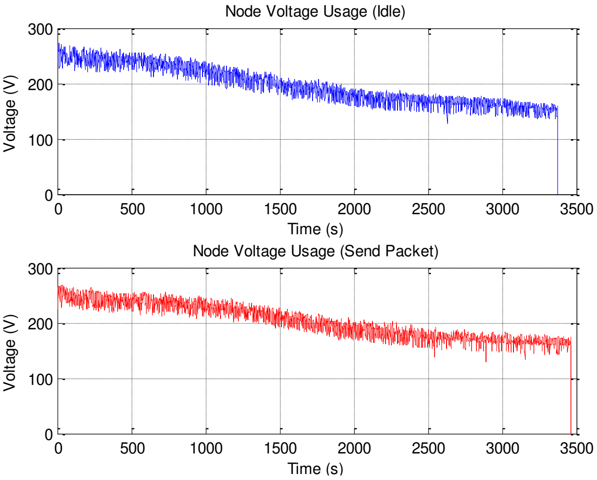

Fig. 13. Nodes voltage usage.

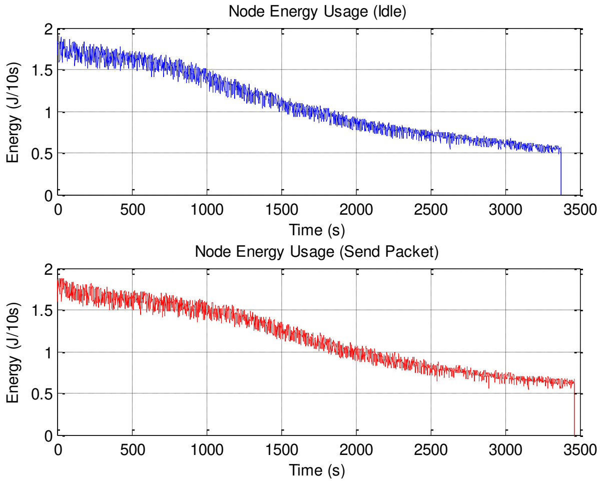

Fig. 14. Nodes power usage.

The voltage on the battery is reduced from average 0.001 V and average 0.037 mAs in current reduction. The serial monitor on the Arduino will be disconnected if the power shield does not longer provide sufficient power to the node. Without cluster, nodes are disconnected when the voltage and current values are reduced to a minimum value of 3.8 V and 175.6 mA. It means the required minimum power for not cluster node is 0.67 W. However, with cluster method, nodes are disconnected when the voltage and current are 3.8 V and 175.6 mA with minimum 0.67 W required power for cluster nodes. Therefore, by using the cluster-based method on WSN nodes, it can improve 6.5% of power efficiency.

IV. Conclusion

The researches have presented a WSN with the cluster-based protocol. Cluster formation and member node selection are performed based on k-means algorithm. It utilizes Euclidean distance from the conversion of RSSI into distance estimation. The result of distance estimation in the observation area has 27.9% error. The average time required for cluster formation is 58.54 s, while the average time used to retrieve coordinate data on each cluster until it is sent to the database is 45.54 s. By using the cluster-based method on WSN nodes, the method can improve the power efficiency by 6.5%.

Acknowledgement

This work is partially supported by Penelitian Dosen Pemula (PDP) 2018, Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

References

[1] K. Kaur, R. Sethi, and A. Kaur, "Power efficiency in agriculture using wireless sensor network," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 3, pp. 3791–3793, 2014.

[2] N. Pachori and V. Suryawanshi, "Cluster head selection prediction in wireless sensor networks," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1033–1035, 2015.

[3] H. Taheri, P. Neamatollahi, O. M. Younis, S. Naghibzadeh, and M. H. Yaghmaee, "An energy aware distributed clustering protocol in wireless sensor networks using fuzzy logic," *Ad Hoc Networks*, vol. 10, no. 7, pp. 1469–1481, 2012.

[4] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 660–670, 2002.

[5] M. Razzaq, D. D. Ningombam, and S. Shin, "Energy efficient k-means clustering-based routing protocol for WSN using optimal packet size," in *International Conference on Information Networking (ICOIN)*. Chiang Mai, Thailand: IEEE, Jan. 10–12, 2018, pp. 632–635.

[6] H. Echoukairi, A. Kada, K. Bouragba, and M. Ouzzif, "A novel centralized clustering approach based on k-means algorithm for wireless sensor network," in *Computing Conference*. London, UK: IEEE, July 18–20, 2017, pp. 1259–1262.

[7] M. Lehsaini and M. B. Benmahdi, "An improved k-means cluster-based routing scheme for wireless sensor networks," in *International Symposium on Programming and Systems (ISPS)*. Algiers, Algeria: IEEE, April 24–26, 2018, pp. 1–6.

[8] J. Yu, Y. Qi, G. Wang, and X. Gu, "A cluster-based routing protocol for wireless sensor networks with nonuniform node distribution," *AEU – International Journal of Electronics and Communications*, vol. 66, no. 1, pp. 54–61, 2012.

[9] R. Patil and V. V. Kohir, "Energy efficient flat and hierarchical routing protocols in wireless sensor networks: A survey," *IOSR Journal of Electronics and Communication Engineering (IOSR–JECE)*, vol. 11, no. 6, pp. 24–32, 2016.