iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4037

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Efficient Hardware Approach for Clustering Technique in Data Analytics

Bangarraju P[1], Vijaya Rama Raju M[2]
[1]*M.Tech Student, Department of ECE, SRKR Engineering College, Bhimavaram*
[2]*Associate Professor, Department of ECE, SRKR Engineering College, Bhimavaram*

***Abstract:** As from the late 2000's the usage of data has been increased tremendously, so there is a need for analysis of the growing data. In analyzing large amounts of data the K-Means clustering is one of the most popular algorithms which is an unsupervised learning algorithm. As the software algorithms are incapable of analyzing large amounts of data fastly, it is imperative to provide some hardware support for analyzing such a big data efficiently and effectively. In this paper for analyzing large amounts of data an efficient hardware implementation of the K-Means algorithm is proposed for two-dimensional data.*
***Keywords:** K-Means, FPGA, IRIS data set, C++, HDL.*

I. INTRODUCTION

Data acquisition and the storage of data have evolved from the late 2000's in almost every field like communication, medical, social networks etc. Analysis and processing of such an enormous amount of data pose a serious challenge. Big data is a field associated with the ways to finding techniques for analysis of enormous amount of data and large data sets. Big data analytics involves many important data mining tasks and the most important one among them is clustering and classification. The purpose of clustering and classification process is to categorize the data into meaningful clusters or groups depending on the similarity or dissimilarity between its data points or objects. Classification is a form of supervised learning technique whereas clustering is an unsupervised learning technique. Classification needs prior assumptions and directions from a human user, which includes assigning labels or classes to the sub-groups by sieving through the data set. But on the other hand, clustering makes its own assumptions and automatically categorizes the data into sub-groups on the basis of the similarity or dissimilarity between its objects or data points.

Existing clustering and classification algorithms are typically processor based (software) algorithms which are incapable of analyzing and processing enormous amounts of data effectively. In order to satisfy the requirements associated with big data analytics, there is a need to provide some special purpose hardware support to it.

II. RECENT WORK

Authors in [1] proposed a highly parameterizable architecture for the implementation of K-Means clustering algorithm called Lloyd's algorithm. The proposed architecture in [1] can be tuned according to the parameters of the input data i.e., width of the data, number of dimensions of the data, number of centroids to be formed, centroid based parallelism degree and the dimension based parallelism degree. Authors in [3] presented the design and implementation of highly parameterizable FPGA core of K-Means clustering. The implementation outperformed equivalent GPP and GPU implementations in terms of speed (two orders and one order of magnitude respectively). In addition the FPGA implementation was more energy efficient than both GPP (615x) and GPU (31x). Authors in [6] proposed a hardware FPGA design for Microarray data mining, which would be power efficient and useful for server solutions. The results provided shows a promising speed-up potentials of FPGA's in Microarray data analysis. Authors in [4] proposed a FPGA implementation of the K-Means clustering algorithm which was fast and area efficient for clustering of one dimensional data, in this implementation the equations for the centroid update are modified for recursive calculation of new centroids. The division operation was replaced with a shift operation which results in a more area efficient hardware implementation without altering the quality of clustering results. Authors in [5] said that the K-Means algorithm is an important clustering algorithm in the field of pattern recognition and data mining. To make this algorithm feasible for multimedia applications with large cluster numbers, a flexible HK-Means hardware architecture was proposed. Authors in [7] proposed a real time hardware realization of a machine learning based K-Means clustering spectrum sensing technique which was more robust and efficient in terms of noise uncertainty as well as computational complexity as compared to the conventional spectrum sensing algorithms. Authors in [2] proposed that the real time k-means clustering can be realized for large size color images and large number of clusters by generating kd-trees dynamically on FPGA and by reducing the amount of computation of squared distances using kd-trees. By reusing units to calculate squared distances for generating kd-trees and for assigning pixels to one of the K clusters, the whole circuit can be

implemented on one FPGA. With the current largest FPGA with DDR memory interface, it will be possible to improve the performance by processing more pixels in parallel. The performance of the system is more than 30fps in average, when the number of pixels in an image is less than 300K. This performance is fast enough for video sequences.

III. CLUSTERING

Machine learning is the study of algorithms and statistical models that computer systems use to efficiently perform a specific task without using explicit instructions, relying on patterns and inference instead. Machine leaning can be considered as a part of artificial intelligence. Machine learning algorithms build a mathematical model of sample data known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are most widely used in a numerous applications like pattern recognition, computer vision etc. Data mining is a field of study within machine learning and focuses on exploratory data analysis through unsupervised learning. In its applications across business problems, machine learning also referred to as predictive analysis.

1) The types of machine learning algorithms differ in their approach, the type of data they input and output and the type of task or problem that they are intended to solve. The three types of learning algorithms are:

a. Supervised or semi supervised learning algorithms
b. Unsupervised learning algorithms
c. Reinforcement learning algorithms

Clustering also called as cluster analysis is the task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar compared to those in other groups or clusters. It is a main task of exploratory data mining and a common technique for statistical data analysis used in many fields including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression and computer graphics etc.

2) The types of clustering algorithms are:

a) Centroid based clustering
b) Connectivity based clustering
c) Distribution based clustering
d) Density based clustering

K-Means clustering algorithm is a centroid based clustering algorithm that is popular for cluster analysis in data mining. It aims at partitioning the "n" observations or data points in a data set into "k" clusters in which each observation or data point belongs to the cluster with the nearest mean or centroid.

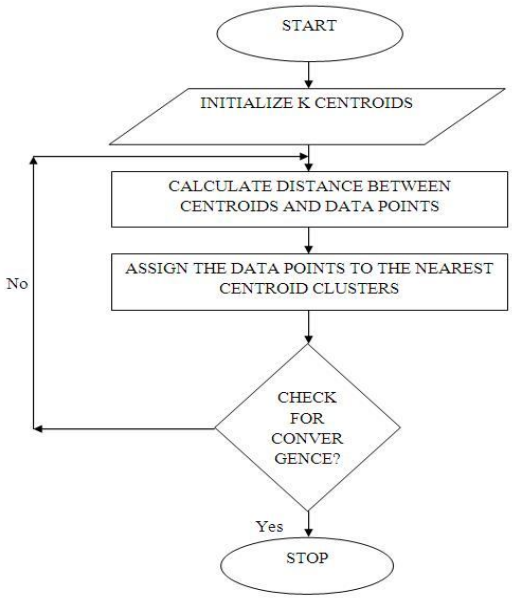


Fig. 1 K-Means algorithm flowchart

3) K-Means clustering algorithm steps:

1) Step 1: Random initialization of k centroids from the n data points in a given data set.

2) Step 2: Distance calculation between the initial centroids and the data points.

3) Step 3: Assigning data points to the nearest centroid cluster.

4) Step 4: Calculating the new centroids of the clusters formed.

5) Step 5: Checking for convergence that is if the previous centroids and the new centroids after step 4 are same then they are said to be converged.

6) Step 6: If the centroids converge then stop the process otherwise go to step 2.

IV. PROPOSED METHODOLOGY

The methodology proposed is to implement the K-Means clustering algorithm on a FPGA based hardware device to make it faster while analyzing large data sets. FPGA's are field programmable gate arrays that can be programmed or reprogrammed after manufacturing. So the K-Means algorithm is developed in C++ software programming language and observed the results in a C++ compiler. The program code is analyzed and constrained architecturally and then trans-compiled into a RTL design in a hardware description language (HDL). Then the HDL program is synthesized to the gate level logic using a synthesis tool.

The block diagram of the proposed hardware architecture of the K-Means algorithm is shown in the fig.2. The centroid buffer contains the centroid values. The pseudo random generator generates the initial centroids randomly by using the seed input. The data buffer is loaded with the input data. The control unit controls the operations of all other units depending on the input values of number of data points, number of clusters to be formed and the number of iterations to be carried out. The distance calculator measures the distance between the data points and the centroids. The adder module finds the shortest distance between the data points and the centroids, the sum updating engine accumulates the sum of all the data points belonging to every cluster. Then the divider module finds the new centroids by taking the mean of the data points in each cluster. The convergence checker checks whether the old centroids before distance calculation and new centroids that formed after distance calculation and updatation of the clusters. If both old centroids and new centroids are same then they are said to be converged. If the centroids are not converged then the process of distance calculation between the centroids and the data points, finding shortest distance between the centroids and the data points and updating the sum of the data points in each cluster, finding the new centroids and the converge check continues until convergence of the centroids occurs.

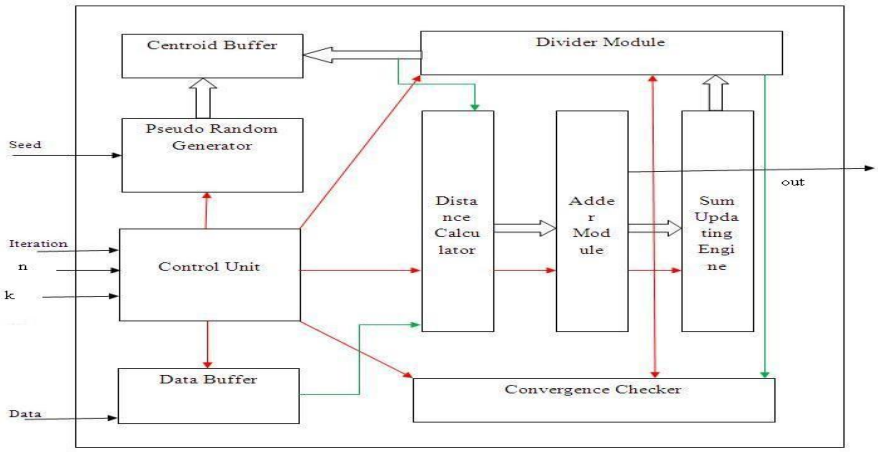


Fig. 2 Block diagram of the proposed K-Means architecture.

The state diagram of the proposed K-Means architecture is shown in fig.3. The control unit used in the design is a finite state machine (FSM). When the pseudo random generator provides the initial centroids based on the input, the process starts with the Go signal from the control unit. On finding the distance between the centroids and the data points and the shortest distance between them the sum updating engine accumulates the data into the corresponding clusters. Then the division process starts and the new centroids are found and updated. Then the convergence check process starts, if the centroids converge then the process stops and gives the output, if convergence is not achieved then the process goes to the iteration state and the process continuous until convergence of the centroids.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

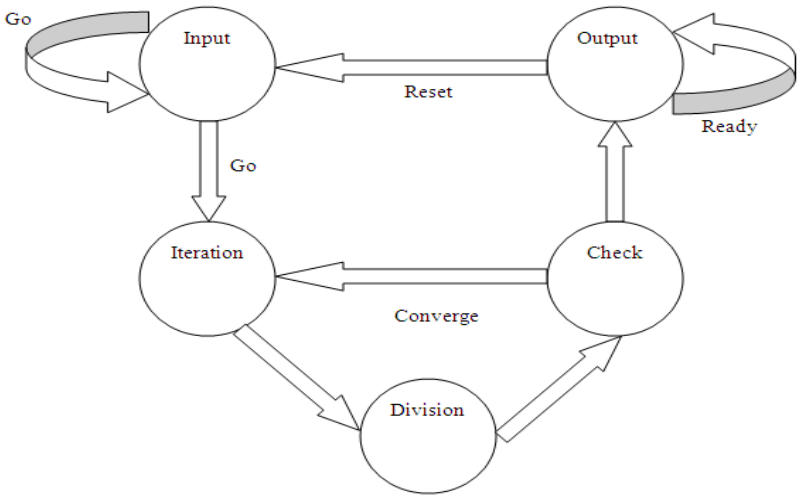


Fig. 3 State diagram of the proposed K-Means architecture

V. EXPERIMENT RESULTS

The universal Iris flower data set consisting of 150 samples containing three different types of flowers iris setosa, iris virginica and iris versicolor is used. The petal width attribute of IRIS flower data set consisting of two dimensional data is taken as the input to the algorithm and then implemented in C++ software programming code and in RTL design using HDL.

```
5.7     2.1
6       1.8
4.8     1.8
4.9     1.8
5.6     2.1
5.8     1.6
6.1     1.9
6.4     2
5.6     2.2
5.1     1.5
5.6     1.4
6.1     2.3
5.6     2.4
5.5     1.8
4.8     1.8
5.4     2.1
5.6     2.4
5.1     2.3
5.1     1.9
5.9     2.3
5.7     2.5
5.2     2.3
5       1.9
5.2     2
5.4     2.3
5.1     1.8

Enter the number of clusters you want to create : 3

Enter the number of objects: 150
```

Fig. 4 Input data with inputs k=3 and n=150

The petal width attribute of the IRIS flower data set containing 150 values is provided as the input i.e., the number of data points are 150 and the number of clusters taken are three since the data set contains three different types of flowers i.e., number of clusters to be formed are 3.

The process takes seven iterations to converge and then stopped. The three clusters formed for versicolor, virginica and setosa with cluster centers (4.2692, 1.3423), (5.5958, 2.0375) and (1.464, 0.244) respectively. The process takes 6.439s to complete using C++ program code. The results obtained are as follows:

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

```
(1.9,0.2)
(1.6,0.2)
(1.6,0.4)
(1.5,0.2)
(1.4,0.2)
(1.6,0.2)
(1.6,0.2)
(1.5,0.4)
(1.5,0.1)
(1.4,0.2)
(1.5,0.1)
(1.2,0.2)
(1.3,0.2)
(1.5,0.1)
(1.3,0.2)
(1.5,0.2)
(1.3,0.3)
(1.3,0.3)
(1.3,0.2)
(1.6,0.6)
(1.9,0.4)
(1.4,0.3)
(1.6,0.2)
(1.4,0.2)
(1.5,0.2)
(1.4,0.2)

--------------------------------
Process exited after 6.439 seconds with return value 0
Press any key to continue . . .
```

Fig. 5 Runtime of C++ Program

Then the HDL program is synthesized on a synthesis tool that produced combinational path delay of 5.934ns.
On comparing the software implementation results and the hardware implementation results, software implementation takes 6.439s whereas the hardware implementation takes only 5.943ns to complete the process. So, there is more speedup in the process when implemented in hardware than the software.

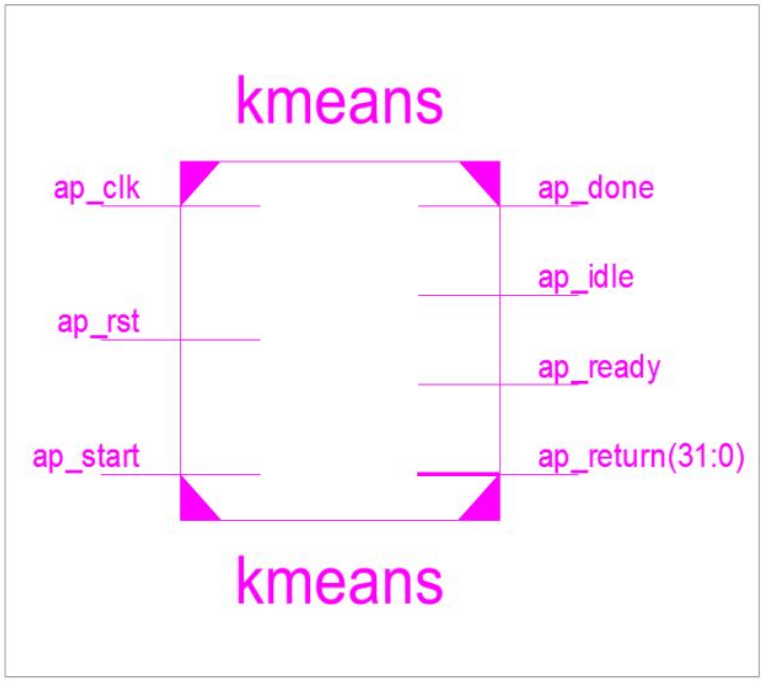


Fig. 6 RTL Schematic at top level

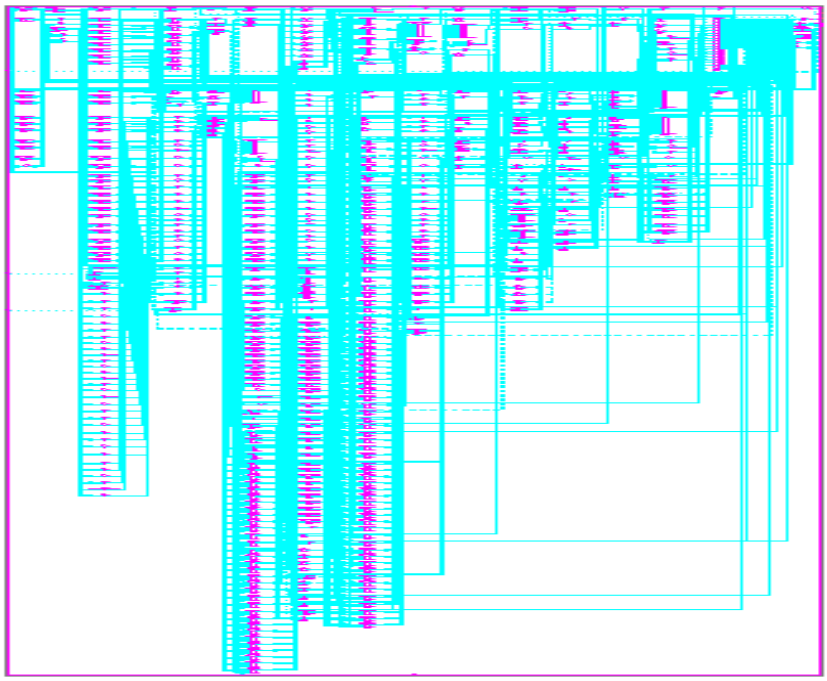


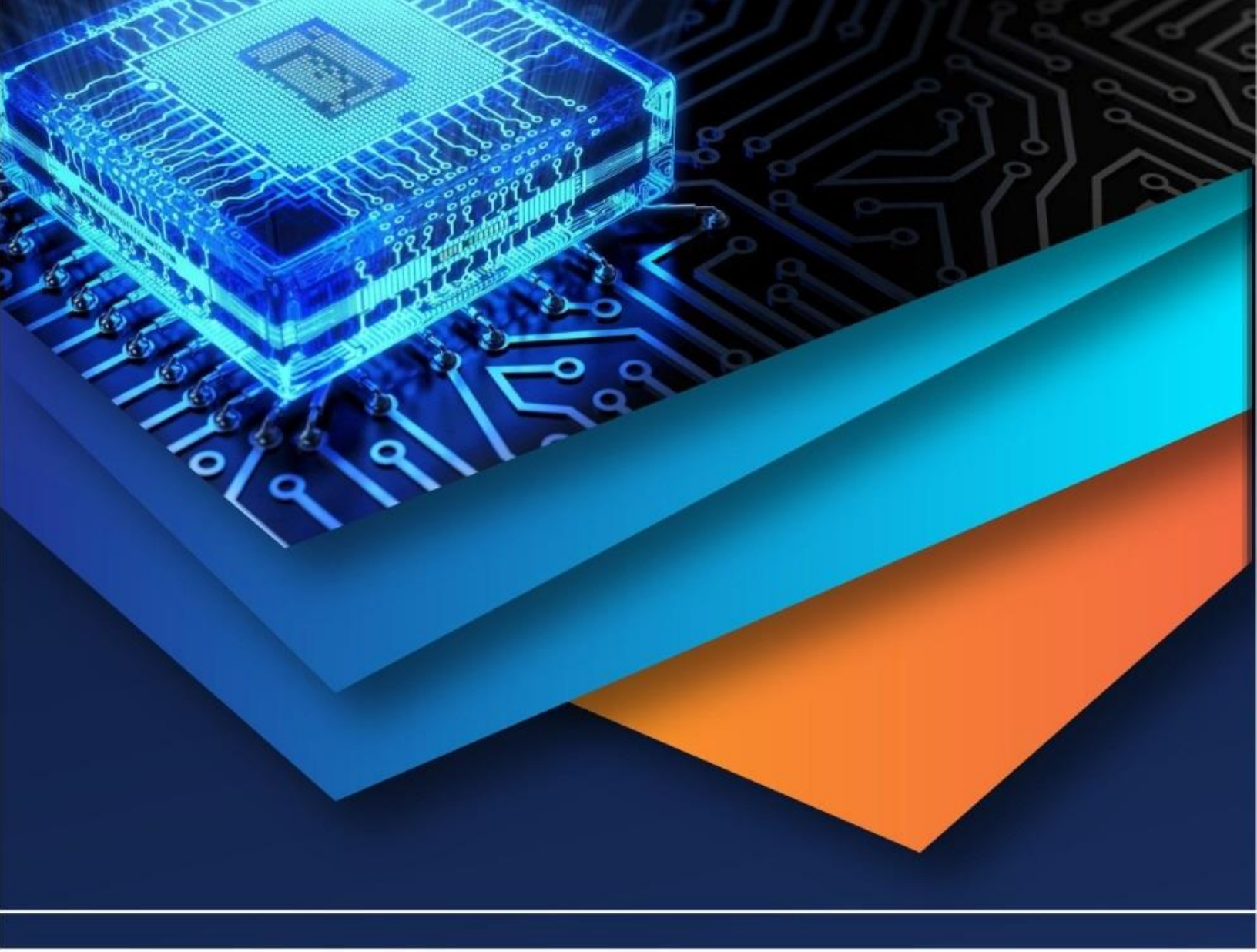
Fig. 7 RTL Schematic at gate level

VI. CONCLUSION

From the observed results the hardware implementation of K-Means clustering algorithm based on FPGA produces faster results than the software based C++ algorithm. Hence, for analyzing large amounts of data the use of FPGA based hardware architectures will provide faster results instead of using general purpose processing devices.

REFERENCES

[1] A. Amaricai, "Design Trade-offs in Configurable FPGA Architectures for K-Means Clustering", Studies in Informatics and Control, Vol. 26, No. 1, pp. 43 – 48, March 2017.
[2] Takashi Saegusa Tsutomu Mauyama, "An FPGA Implementation of Real-Time K-Means Clustering for colour images", Journal on Real-Time Image Processing, Springer, pp.309 – 318, 2007
[3] H. M. Hussain, et al, "Novel Dynamic Partial Reconfigurable Implementation of K-Means Clustering on FPGAs: Comparative Results with GPPs and GPUs", Hindawi Publishing Corporation International Journal of Reconfigurable Computing, Volume 2012, Article ID 135926.
[4] Awos Kanan, F. Gebali, A. Ibrahim, "Fast and Area-Efficient Hardware Implementation of the K-Means Clustering Algorithm", WSEAS TRANSACTIONS on CIRCUITS and SYSTEMS, Volume 15, pp. 133 – 142, 2016.
[5] Tse-Wei Chen, Shao-Yi Chein, "Flexible Hardware Architecture of Hierarchical K-Means Clustering for Large Cluster Number", IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, Volume 19, No. 8, pp. 1336 -1345, August 2011.
[6] H. M. Hussain et al, "FPGA implementation of K-Means Algorithm for Bioinformatics Application: An Accelerated Approach to Clustering Microarray Data", NASA/ESA Conference on Adaptive Hardware and Systems (AHS-2011), pp. 248 – 255.
[7] Anirudh Agarwal, Ranjan Gangopadhyay, "Hardware Implementation of K-Means Clustering Based Spectrum Sensing Using USRP in a Cognitive Radio System", "International Conference on Advances in Computing, Communications and Informatics (ICACCI)", pp. 1772 – 1777, 2017.
[8] R.woods and J. Mc Allister, FPGA – based implementation of signal processing systems: John Wiley & Sons, 2008.
[9] F. C. J. Allaire, FPGA implementation of an unmanned aerospace vehicle path planning genetic algorithm. Royal Military College: Libraries and Archives, Canada, 2007.
[10] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Prentice Hall Advanced Reference Series, 1998.
[11] D. Singh and C. Reddy, "A Survey on platforms for big data analytics", Journal of Big Data, 2014.
[12] M. Naghmash, M. F. Ain and C.Y. Hui, "FPGA Implementation of Software Defined Radio Model Based 16QAM", European Journal of Scientific Research, vol. 35, pp. 301 – 310, 2009.
[13] Z. Ge, Y . Jinghua, L. Qian and Y. Chao, "A real – time speech recognition system based on the implementation of FPGA", Cross Strait Quad – Regional Radio Science and Wireless Technology Conference (CSQRWC), pp. 1375 – 1378, 2011.
[14] S. A. Kadir, A. Sasongko and M. Zulkifli, "Simple power analysis attack against elliptic curve cryptography processor on FPGA implementation", International Conference on Electrical Engineering and Informatics (ICEEI), pp. 1 -4, 2011.
[15] Y. Abhyankar, C. Sajish, P. Kulkarni and C. R. Subrahmanya, "Design of a FPGA based data acquisition system for radio astronomy applications", The 16th International Conference on Microelectronics (ICM), pp. 248 – 255, 2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY