H-RAN: An Approach Toward Cloud-RAN Load Balancing

Byomakesh Mahapatra *, Ashok Ku. Turuk †, Niranjan Ray ‡, and Awaneesh Ku. Yadav §
* † § Dept. of Computer Sc. & Engg.
National Institute of Technology, Rourkela, Odisha, India
‡Dept. Computer Sc. & Engg.
KIIT University, Bhubaneswar, Odisha, India
*Email: [*byomakesh22, †akturuk, ‡rayniranjan, §kumarawani9696] @gmail.com*

***Abstract*—The centralized cloud radio access network (C-RAN), change the BS architecture to a large extent, in term of associated cost and coverage. The main limitations of C-RAN architecture is that it has a centralized processing unit, which increases the network latency by increasing the service waiting time. The proposed Heterogeneous Radio Access Network (H-RAN) consist of some Edge node and local processing unit (LPU) which resides in remote radio head (RRH). This introduction of enhanced RRH (eRRH) node increases the capacity of the network and decrease the network latency, on the other hand its also share the computing and controlling task of the centralized BBU. The proposed LPU unit consists of a number of small VB which generally handle real-time computing and processing task. In the first section of this paper, we have proposed a new RAN architecture and all the load balancing condition are represented through a detail mathematical model. The paper also present a load-aware dynamic BS switching algorithm (LDBSS) for power optimization and load balancing. The simulation work is carried out with the help of CloudSim 3.0.3 simulator by considering equal number of packet arrival for both C-RAN and H-RAN. The simulated result shows that with the introduction of the middle Edge layer and LPU at the RRH reduces the waiting time and blocking probability of the network as compared the C-RAN architecture.**

***Index Terms*—C-RAN, eRRH, H-RAN, LPU, Load Balancing**

I. Introduction

With the development of new wireless technology, the key challenges are to design a wireless system which provides high data-rate, low latency, energy efficient and cost-effective network. C-RAN is a new concept for base station energy optimization and reducing the different cost involved in the BS establishment and operation [1]. The edge computing based small cell is a new technology for creating an energy efficient network which solves many problems present in the traditional cellular network. At the current scenario with the fixed physical resources, it is a challenging task to implement a large number of BS corresponding to the user requirement. C-RAN more how to solve some of this physical constraint by modifying the BS architecture [2]. But still, C-RAN can't be more profitable in a densely populated wireless area, where a large number of user entity (UE) and digital wireless devices (DE) associated with a particular BS. To accommodate this and to get a better quality of services (QoS) the BS should be closer to the mobile devices [3]. Recently develop Edge computing is more helpful to solve this issues by putting the frequently used task to the edge of the base station. The Edge computing may be a part of the remote radio head (RRH), or the wireless access point (AP). As this Edge-RRH is small in size it can be put at any location like a light post, roof-top and tree-trunk which is near to the end user and can be accessed easily. All this small BS, growth over a local cloud known as Edge-node or Fog-node. These nodes are again controlled by a centralized BS to handle a large processing load [4]. Though the use of edge node in small cell solves many computing issues, but it increases the network and BS complexity to a large extent.

In this paper we have proposed a heterogeneous RAN architecture for load balancing and power consumption optimization of C-RAN architecture. This H-RAN consists of two main layers, i.e., a centralized cloud BBU layer and a small distributed edge cloud layer. The incoming traffic load or task are classified and allocated to the different segment of H-RAN platform, i.e., Fog or cloud layer [5]. In this paper, we have highlighted different challenge associated with the present C-RAN architecture and proposed some corresponding solution to overcome this challenges:

A. Challenges and solution

- *Issue:* In C-RAN base station all the base band processing and controlling activity is carried out at the central BBU. This centralization increase processing delay and network latency [6].
- *solution:* The BBU load and network delay is avoided by distributing processing load between local edge cloud and centralized cloud.
- *Issue:* The one to one connection between RRH-BBU needs large number of fronthaul link, which leads to increase in network complexity and capital cost [7].
- *Solution:* In H-RAN a cluster eRRH unit are interconnected (microwave) to each other and some of the eRRH unit work as a relay or master node. This master nodes are only connected to the BBU-pool through optical cable.
- *Issue:* The centralized processing increase load on the fronthaul and C-BBU connection which connect RRH.
- *solution:* In H-RAN some of the regular task are processed at the LPU and rest of task are transfers to the

C-BBU. This distribution of the task decrease loading effect on the fronthaul connection.

B. Objectives and main contribution

The objective of this paper is to propose a distributed cloud-based H-RAN which avoid limitations like centralized load balancing at C-BBU and fronthaul load sharing. The main contributions of this paper are as follows:

- Proposed a small cell using the LPU to decrease the network latency and increase scalability of the network.
- Balances the incoming traffic load across the BBU-pool by performing some local base band processing at the edge node instead at the centralized cloud.
- Implement a load based dynamic switching technique at the local node to optimized the power consumption.

C. Architecture of H-RAN

The H-RAN architecture consists of a data-centric BBU-pool and a number of eRRH connected for baseband processing and load sharing as shown in Fig.1. The different functional unit of the H-RAN are described as follows.

- **Antenna Unit:** The antenna unit consists of some directional sector antenna for transmitting and receiving the RF signal from the cellular UE/DE. Higher data rate and lower bit error rate (BER) can achieve through advances multiple antenna technologies like MIMO and massive MIMO.
- **Enhanced RRH (eRRH):** The eRRH unit mainly consists of some communication and computation block. The communication block performs signal UP/DOWN, amplification and the computational unit performs signal processing as well as signal reconversion.
- **Centralized Base Band Unit (C-BBU**):The C-BBU located at a centralized location performs baseband signal processing and controlling. The C-BBU and eRRH connect with a large bandwidth optical fronthaul or microwave link. A scalable BS platform can be achieved by using cloud computing, and virtualization technology at the C-BBU. The C-BBU contain some Virtual Box (VB) which is controlled by a hypervisor unit as shown in Fig.2. The eRRH data transferred to the C-BBU through the optical link with the help of Common Public Radio Interface (CPRI) protocol [11].

The rest of the paper is organized as follows: Section. II describes, about prior related work. In Section III, the architecture of the H-RAN describes. Section IV, described the system model and mathematically formulated the load condition of H-RAN and compared with the C-RAN. In section V, simulation and experimental result are evaluated and explained. Finally concluding remarks are given in section VI.

II. Prior Related Work

The next generation cellular network services providers mostly focus on two most important thing, i.e., latency and data-rate. A high quality of services (QoS) and low end to end delay are achieved by minimizing the processing and communication delay. On the other hand, by increasing the channel bandwidth, a high data rate can be obtained. Again real-time traffic load management at the BS is another challenging task for the services provider. In 2012, virtualization and cloud computing technique were first introduced by IBM to make the radio access platform more scalable and flexible. Many research works are carried out by different research organization and service providers to solve the load distribution problem for C-RAN architecture. In [8], authors proposed a Fog RAN architecture for local processing of the most frequent arrival task, to reduces the load on central cloud BS. This work introduced a general purpose processor (GPP) based fog node with an integrated radio unit. In [10], [12], Authors described about a scenario of load distribution in the small cell using cloud computing technology for the ultra-dense cellular network. The paper used a low complexity small cell cluster for small cell resources management, by using fog load balancing technique. In [13], An advances network architecture by integrating Cloud and Fog technology is proposed in [13]. This paper analyzed different network architecture based on radio resources control and mobility management. The performances analysis given in this paper shows that due to the harmonization of both RAN technology increase the throughput and scalability of C-RAN architecture. In work [14], backhaul and fronthaul solution for Fog-RAN based small cell network was carried out. This work proposed an optical wavelength division multiplexing (WDM), for integrating some operator or services over a shared optical network backbone. This paper also used a Cloud-RoF based centralized architecture for resources sharing in the C-RAN network. In [15] [16], authors described about a load balancing approach to reduce the heavy burden on the backhaul and fronthaul system of C-RAN. They proposed a joint resources allocation technique and a cached enable RRH for increasing QoS. Although these paper are described the edge computing for the different C-RAN scenario, but all this paper is lack of a concrete model for load balancing approach. To fill the void, in this work we have proposed a detail mathematical for load balancing by using the heterogeneous architecture.

III. System Model and Problem Formulation

The load balancing phenomenon of any BS depends upon two factor i.e associated traffic load and the base station capacity. The traffic load give an idea about the number of UE and channel occupancy, whereas the capacity represents the efficiency of the system [9]. In this section, we have given system model for both C-RAN and H-RAN, under the same packet arrival rate λ. Suppose an H-RAN is having 'N' number of eRRH unit connected to a centralized BBU, which contains 'M' number of VB for processing 'P' number of packets at a time unit 't'. Considering packet arrival as birth and death process (M/M/1) at eRRH, the packet arrival and departure time Fig.2 can be express as:

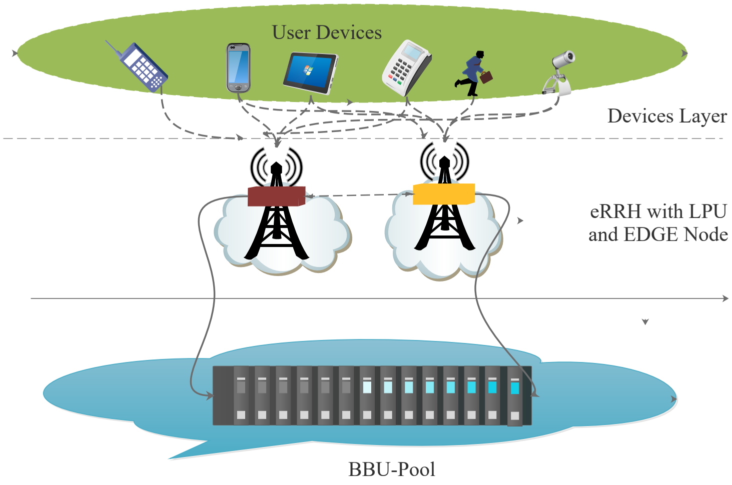

Fig. 1. Different layer of H-RAN and its associated components

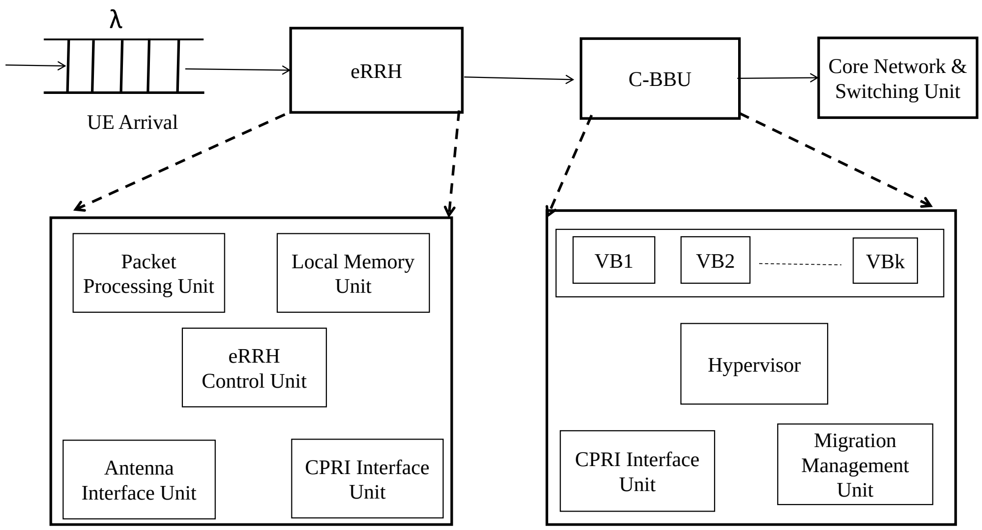

Fig. 2. H-RAN block diagram internal Architecture

Departure Time = Arrival Time + Workload Found in the System + Transmission time

The rate of change of capacity of an eRRH is depend on UE arrival or the total traffic load at the eRRH at a particular interval time which can be express as:

$$\frac{\partial C_{eRRH}}{\partial t} = \gamma C_{eRRH} - \alpha \sum_{i \in |K|} L_{UE} \tag{1}$$

Where C_{eRRH} = Total Capacity of eRRH
L_{UE} = load generated by each user entity which are linked with the eRRH L_{eRRH} = Total load on the eRRH Based on the above equation the threshold value for eRRH switching decision can be express as:

$$L_{th} < L_{eRRH} < U_{th} \tag{2}$$

Similarly, the capacity of the BBU depend upon the load on the BBU i.e number of eRRH presently associated with the BBU. This load is generally the processing load which is forward by the eRRH after a local processing.

$$\frac{\partial C_{BBU}}{\partial t} = \kappa C_{BBU} - \beta \sum_{i \in |N|} L_{eRRH} \tag{3}$$

which can be express as

$$\frac{\partial C_{BBU}}{\partial t} = \kappa C_{VB} - \beta L_{BBU} \tag{4}$$

TABLE I
PARAMETER AND ABBREVIATION

Parameter	Abbreviation
α	The efficiency of RRH load link with the VB
β	The efficiency of VB load link with the BBU
κ	BBU capacity constant

Algorithm 1 Load-aware dynamic BS switching algorithm (LDBSS) [19] [20]

$N \leftarrow$ Initialize the number of eRRHs in a cluster
Calculate total traffic load generated by UE

Calculate the present load at eRRH
Calculate the threshold value of each eRRH
if $L_{RRH} < TL_{eRRH}$ **then**
  Switch off the RRH and migrate the load to other RRH
  N = N - 1
  Update the load of other eRRH
  Update the power consumption
end
else
  if $L_{RRH} > TH_{eRRH}$ **then**
    Migrate some current load to associated C-BBU
    N = N + 1
    Update the load and power consumption
  end
end
Release the resources and update the Information

A. Algorithm for eRRH Switching

The H-RAN entertains two basic types of incoming traffic the real-time traffic (voice, streaming video, D2D communication) are need to be more priority than non-realtime traffic such as mail services, web-services, etc. To provide priority to the real-time task packet classification and allocation are carried out at the eRRH. Initially, the classifier of eRRH classified the whole incoming packet into two separate entity group (queue)such as:
$e_1 = \{R1, R2, R3....Rn\}$ = The set of real-time task focused to be processed on the Edge node (eRRH)

$e2 = \{T1, T2, T3....Tn\}$ = The set of non-real time task (delay aware) which are to be processed in the centralized cloud (C-BBU).

Power consumption can be optimized by switching *On* and *Off* the eRRH based on the capacity and load condition of the eRRH as given in eq.(1) and (2). The switching is based on the eRRH and C-BBU threshold load condition. This threshold condition is lower threshold condition TL_{eRRH} and upper threshold TU_{eRRH}. The incoming packets are accepted or rejected for further processing was decided by the present load and capacity of the eRRH which must be lies between the upper and lower threshold value. When the incoming traffic load exceeds these boundary threshold value (TU_{eRRH}), the eRRH migrate some of the running tasks to either the assigned C-BBU or the eRRH which are lying in his cluster area. This upper threshold load balancing condition is more concentrated on maintaining the constant load condition at eRRH and BBU. To optimize the power consumption BS used the lower threshold (TL_{eRRH}) based BS switching technique. This switch *Off* the eRRH which having a minimum load by transferring the running load to other eRRH in his cluster area. This selected switching technique of eRRH or BBU based on the load and capacity reduces the overall power consumption of the H-RAN system.

IV. Simulation and Result

The simulation was carried out with the help of simulator CloudSim simulator 3.0.2 on a core i7, 20GB RAM, and Windows7 OS platform. For the execution of the task, we have used NetBeans IDE 8.2 tools. Initially, we have taken a centralized BBU consist of a number of virtual box (VB), growth over a data centeric environment in CloudSim. The incoming UE generated packets, which are coming to the C-BBU through the RRH, are allocated to the different VB based on the allocation policy of hypervisor[17],[18]. By using First Come First Serve (FCFS) allocation policy, we have increased the number of the packet from 100 to 1000 for 10 VB and noted down the total waiting time required for a packet to serve. In the next phase, we have taken three small eRRH nodes each having some local processing unit over a virtualization platform. The LPU consist of some small capacity VB to handle a real-time task. This small VB has mostly handled the delay aware real-time task. Taking the same number of packet arrival, the blocking probability and waiting time are calculated. The results are shown in Fig.3 and Fig.4 shows that with the introduction of the Fog layer at the RRH reduce the services waiting time by balancing the load across the distributed edge cloud. Again Fig 5 shows that due to the heterogeneous layer packet execution time or processing time is in balanced condition after a certain load condition. As both eRRH and C-BBU have a processing unit to handle different type of task this new architecture reduces the latency by increasing the quality of services.

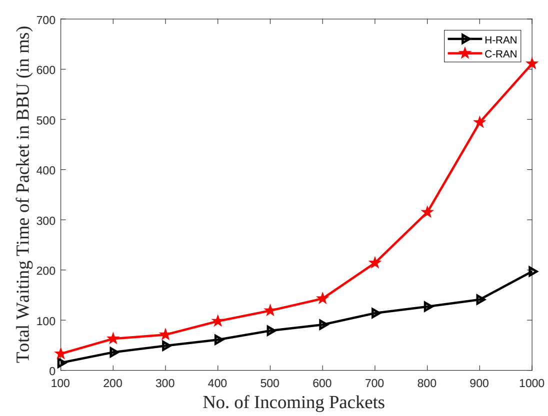

Fig. 3. Total Waiting Time for UEs in C-RAN and H-RAN architecture

V. Conclusion

The increase in the number of the wireless and cellular device needs a high capacity and reliable base station architecture for end to end communication. In this paper, we have focused on two different C-RAN architecture for the next-generation cellular network. The C-RAN is a centralized collaborative RAN architecture use for processing different radio services

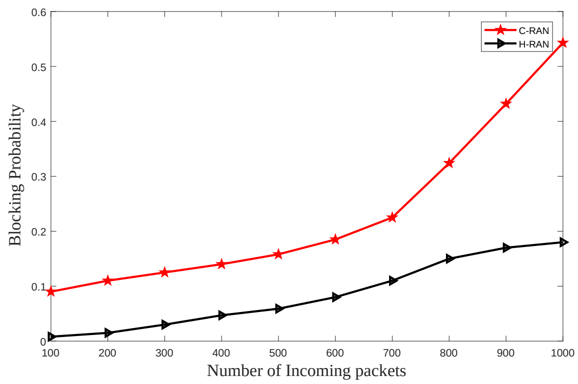

Fig. 4. Blocking probability with respect to the packet arrival for C-RAN and H-RAN architecture

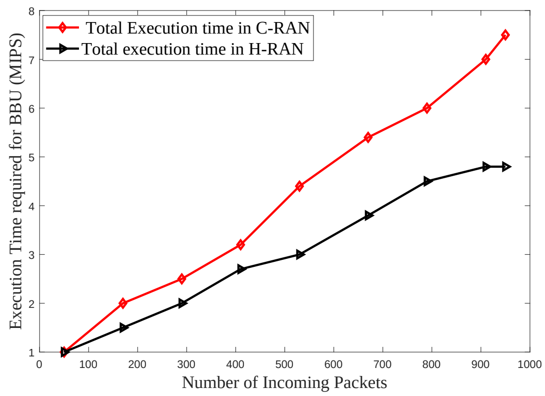

Fig. 5. Execution time with respect to the packet arrival for C-RAN and H-RAN architecture

on a single platform. On the other hand, the proposed H-RAN is a new architecture based on the edge node and cloud computing technology for accessing the local devices and to handle the real-time task. We have used CloudSim simulator for performing packet-level simulation, and results show that the implementation of the edge layer at the RRH decrease the service waiting time with a decrease in queue length. Again the reduction in blocking probability at the C-BBU decreases the network latency which is most significant for the next generation network architecture.

REFERENCES

[1] C. I, J. Huang, R. Duan, C. Cui, J. Jiang and L. Li, "Recent Progress on C-RAN Centralization and Cloudification," in IEEE Access, vol. 2, pp. 1030-1039, 2014.

[2] N. Saxena, A. Roy, and H.-S. Kim, "Trafc-Aware Cloud RAN: A Key for Green 5G Networks," IEEE Journal on Selected Areas in Communications, vol. 8, no. c, pp. 111, 2016.

[3] M. Hadzialic, B. Dosenovic, M. Dzaferagic and J. Musovic, "Cloud-RAN: Innovative radio access network architecture," Proceedings ELMAR-2013, Zadar, 2013, pp. 115-120.

[4] A. V. Dastjerdi and R. Buyya, "Fog Computing: Helping the Internet of Things Realize Its Potential," in Computer, vol. 49, no. 8, pp. 112-116, Aug. 2016.

[5] P. Varshney and Y. Simmhan, "Demystifying Fog Computing: Characterizing Architectures, Applications and Abstractions," 2017 IEEE 1st International Conference on Fog and Edge Computing (ICFEC), Madrid, 2017, pp. 115-124.

[6] B Mahapatra, R Kumar, S Kumar, AK Turuk,"A Heterogeneous Load Balancing Approach in Centralized BBU-Pool of C-RAN Architecture," IEEE International Conference for Convergence in Technology (I2CT),2018.

[7] M. Peng, S. Yan, K. Zhang and C. Wang, "Fog-computing-based radio access networks: issues and challenges," IEEE Network, vol. 30, no. 4, pp. 46-53, July-August 2016.

[8] Y. Ku, D. Lin and H. Wei, "Fog RAN over General Purpose Processor Platform," IEEE 84th Vehicular Technology Conference (VTC-Fall), Montreal, QC, pp. 1-2,2016.

[9] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow and P. A. Polakos, "A Comprehensive Survey on Fog Computing: State-of-the-Art and Research Challenges," in IEEE Communications Surveys Tutorials, vol. 20, no. 1, pp. 416-464, Firstquarter 2018.

[10] J. Oueis, E. C. Strinati and S. Barbarossa, " The Fog Balancing: Load Distribution for Small Cell Cloud Computing," 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), Glasgow, pp. 1-6,2015.

[11] J. Clarke, M.E Luna, and N. Wood," Method and apparatus for controlling network traffic," U.S. Patent 7,237,034.

[12] Z. Zhao, H. Guan, J. Wigard, S. Lee, D. H. Kim and K. M. Hwang, " Network Upgrade with LTE-Advanced Small Cells," 2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall), Boston, MA, pp. 1-6, 2015.

[13] S. Hung, H. Hsu, S. Lien and K. Chen, "Architecture Harmonization Between Cloud Radio Access Networks and Fog Networks," in IEEE Access, vol. 3, pp. 3019-3034, 2015.

[14] D. Chen and V. Kuehn, "Adaptive Radio Unit Selection and Load Balancing in the Downlink of Fog Radio Access Network," IEEE Global Communications Conference (GLOBECOM), Washington, DC, pp. 1-7,2016.

[15] C. Liu, L. Zhang, M. Zhu, J. Wang, L. Cheng and G. K. Chang, "A Novel Multi-Service Small-Cell Cloud Radio Access Network for Mobile Backhaul and Computing Based on Radio-Over-Fiber Technologies," Journal of Lightwave Technology, vol. 31, no. 17, pp. 2869-2875, Sept, 2013.

[16] S. H. Park, O. Simeone and S. Shamai Shitz, " Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks," IEEE Transactions on Wireless Communications, vol. 15, no. 11, pp. 7621-7632, Nov. 2016.

[17] Y. N. Shnaiwer, S. Sorour, P. Sadeghi and T. Y. Al-Naffouri, "Online Cloud Offloading Using Heterogeneous Enhanced Remote Radio Heads," IEEE 86th Vehicular Technology Conference (VTC-Fall), pp. 1-5, 2017.

[18] SK Mishra, B Sahoo and PP Parida," Load Balancing in Cloud Computing: A big Picture," Journal of King Saud University-Computer and Information Sciences, 2018.

[19] E. Balevi and R. D. Gitlin, "Optimizing the Number of Fog Nodes for Cloud-Fog-Thing Networks," IEEE Access, vol. 6, pp. 11173-11183, 2018.

[20] E. Oh, K. Son and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," in IEEE Transactions on Wireless Communications, vol. 12, no. 5, pp. 2126-2136, May 2013.