A Heterogeneous Load Balancing Approach in Centralized BBU-Pool of C-RAN Architecture

Byomakesh Mahapatra*, Rahul Kumar†, Shailesh Kumar ‡ and Ashok Ku. Turuk§
Dept. of Computer Science & Engg.
National Institute of Technology, Rourkela, Odisha, India
*Email: [*byomakesh22, †rahulchauhan533, ‡shaileshkumar478, §akturuk] @gmail.com*

***Abstract*—The Cloud Radio Access Network (C-RAN) is a new paradigm for the next generation wireless and mobile communication for handling a large volume of heterogeneous data-centric devices and communication devices on the same platform. The introduction of cloud technology along with virtualization on the mobile communication platform added more scalability and flexibility to the cellular base station (BS). The centralization of Base Band Unit (C-BBU) also leads to the minimization of the capital cost (CAPEX), and operational cost (OPEX) which incurred during construction and maintenance of the BS. As in C-RAN Architecture, all the signal processing and controlling functionalities are carried out at the C-BBU, which sometimes leads to some overload and under load conditions at C-BBU. This situation can be avoided by proper balancing the load at C-BBU.**

This paper investigated the different load balancing issues in C-RAN and proposed a Heterogeneous Load Balancing (HLB) algorithm for handling the various type of radio technology on the same BBU-Pool. HLB assign the incoming load to a particular VM based on the different factors like packet length, packet headers and VM capacity. The proposed algorithm is compared with a different existing algorithm, and the result shows that HLB is more energy efficient and scalable.

***Index Terms*—C-BBU, Cloud-RAN, HLB, Load Balancing, Virtualization.**

I. Introduction

The need for higher data speeds, fast and efficient network services are pushing towards the next generation of mobile network like 5G . It is envisioned as the next big step in mobile technology which would offer a giant leap in terms of data speeds and quality of service. Many mobile network architectures and technologies are in the proposal to use with the 5G. C-RAN architecture which offers many solutions to the various number of problems which are face by the cellular operators, while at the same time ensuring the quality of services provided to the end users [1]. The C-RAN architecture is based on the idea of pooling the traditionally distributed Baseband Units (BBUs) into a centralized BBU Pool for reducing the wastage of resources, reducing power consumption and to increase the quality and fastness of the services. It also performs the load balancing of the signals originating from different base stations which increase the network capacity [2]. The C-RAN environment is similar to the traditional cloud computing environment in which the tasks(signals) are migrated from the over-load VMs (BBUs) to underutilized VMs using the scheduling and load balancing concepts for effectively sharing their resources. In this paper, we discuss the various load balancing algorithms that can be applied to the C-RAN architecture. It was founded that load balancing ensured effective sharing of the load across VMs and minimized the execution time and the active time of the physical resources [3]. Cloud radio access network (C-RAN) is a type of mobile network architecture, which addresses above-mentioned problems. The alphabet C can be interpreted as: Cloud, Collaborative, Clean or Centralized processing or Cooperative radio. It was first proposed in 2009 and went through many developments over the period of time [4]. The China Mobile Research Institute proposed an updated version of C-RAN in which the centralization and sharing of the baseband processing done between the sites in a virtualized BBU Pool. It decreases the CAPEX and OPEX of the core network by energy saving methods in networks RAN[6].

In this paper we proposed a heterogeneous load balancing algorithm which can be used for the C-RAN platform which handle different type of packet or frame coming from a heterogeneous radio access technologies(RAT). The rest of the paper is organized as follows: the Section II, described about related prior work mentioned by different authors on load balancing concept in cloud environment. In Section III and IV, a detailed architecture of C-RAN is given and also described the advantages of C-RAN architecture. Section V, described about different load balancing algorithms which are closely related to the proposed algorithm. Section VI and VII, described about mathematical model and proposed algorithm. Finally, the simulated results are given in Section VIII. A concluding remarks of the paper is given in Section IX.

II. Literature Survey and Prior Work

The exponentially increment of wireless and mobile connectivity tends to a data explosion in near future. From [4], it is assuming that the volume of data transmission will increase more than 12 times by 2020. To manage this explosion of data traffic, C-RAN can be considered as the efficient technology for next generation cellular network. Additionally, the C-RAN tends to increase the computational resources by sharing its baseband resources in a shared BBU pool for multiple RRH sites [5]. The main challenge at the C-BBU of C-RAN architecture is to allocate the packets, coming from RRH, to the Virtual Machines (VMs) of a C-BBU pool in an efficient way and manage the excessive data if at all generated during some peak session. The process of managing the load and

compare with the capacity of the VM is known as load balancing. In another way, load balancing can be expressed as the process of migration of the tasks from overloaded VM to under loaded VM to reduce the wastage of the resources in a C-BBU. In [7], the author describes different load balancing algorithms for the data center scenario, they used different load balancing techniques, some of them are static load balancing and some are dynamic load balancing techniques. In [9], the algorithm is focused on the reduction of power usage in the data centers. It takes a group of heterogeneous multi-core servers with different sizes and speed and forms a queue to addresses the problem of load distribution and optimal power allocation and for multiple heterogeneous multi-core server processors across data centers in C-BBU. In [11], an updated weighted round robin algorithm for the cloud environment was proposed, which considers resource capabilities and job length. The scheduling technique in [12] works for both the cloud computing and grid computing. It is a dynamic technique and performs well in reducing the workflow execution time and the scheduling overhead.

III. Architecture of C-RAN

The Cloud RAN architecture includes all the components of a traditional architecture as Base Band Units (BBU) and Remote Radio Heads (RRHs) along with antennas mounted on the cell tower. But the BBUs are centralized into a virtualized BBU pool and shared among the sites. The RRHs are coupled with the processor in BBU pool using high bandwidth and low latency transport links. C-RAN optimizes the BBU utilization between excessively and moderately loaded base stations. This is where the cloud concepts of task allocation and load balancing come in to act [19]. The network manages the mobile users by forming a virtual pool of resources by potentially applying the cloud computing to the RANs. The delivery of the service applications to the end user through the cloud reduces the communication latency to support delay-sensitive real-time applications which is one of the main goals of 5G.

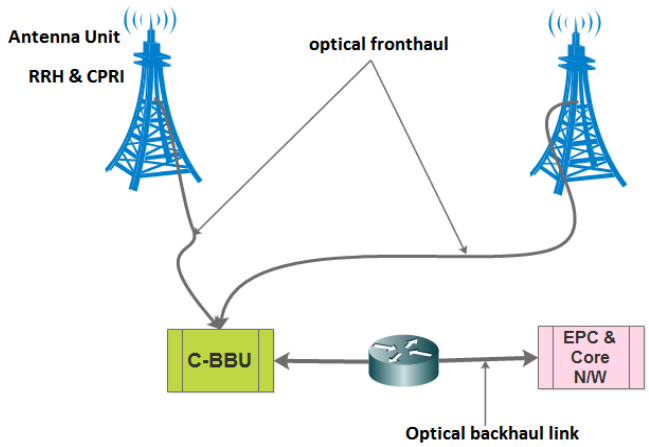

Fig. 1. Cloud-RAN Architecture.

IV. Advantages of C-RAN

C-RAN architecture is useful for both macro and small cell. In case of the deployment of a macro base station, the BBU Pool, which is centralized, enables a well-organized utilization of BBUs . The C-RAN Architecture help to reduce the power consumption associated with the BS. The C-RAN also increase BS scalability by adopting multiple non-uniform traffic [8]. Furthermore, in the LTE networks, the co-location of BBUs simplifies the X2 interface and also makes the associated latency and synchronization issues insignificant. The X2 interface provisions information exchange between BBUs for smooth handover and coordination [17] [18].

Here, the motivation and advantages of C-RAN are described as follows:

- The centralization and virtualization in C-RAN leads to reduce the number of BBUs and its associated cost like capital and maintenance cost.
- The reduction in number of BBUs directly reduces the power and energy consumption associated with the BS.
- Ease of adding and upgrading the BBUs to the pool improves scalability and ease network maintenance.
- Sharing and renting of RAN as a cloud service between different network operators.
- Lower delay in interactions between BBUs which increase throughput, spectral efficiency and reduce delays during handover.
- Adaptability to non-uniform traffic - Facilitation of the environment and methodology for implementing load balancing between the cells.

V. Different Load Balancing Approach in C-BBU of C-RAN

Load balancing is the process of migration of the load from over-utilized nodes to under-utilized nodes to reduce the wastage of the resources in a cloud environment. It can be done in two ways :

- *VM Selection and Migration:* This is used to migrate the VMs to another host to implement server consolidation. Appropriate VMs are selected either from the overloaded or under-loaded host for migration. VM migration aims at achieving the least no. of active physical machines(hosts) to increase energy efficiency by mapping the VMs with the active PMs as compactly as possible. The non-active PMs switched to a power saving mode [10].
- *Dynamic Task scheduling and Migration:* This is used to migrate the tasks from the overloaded VMs to the underloaded VMs for sharing the load effectively in order to achieve the optimal resource utilization. This helps in reducing the overall completion time of tasks and also the decrease of the total execution time in the cloud environment thus helping in the reduced the usage time of the physical resources.

Different type of scheduling and load balancing algorithms are proposed earlier for packets or tasks allocation at C-BBU, which are coming from RRH through a fronthaul network in C-RAN. Some of them are as follows:

1) First Come First Serve (FCFS) : The first come first serve(FCFS) algorithm is most frequently used to schedule the tasks. The FCFS does not take any parameters into account as it just assigns the tasks in a particular order of the arrival. This is required in case the first occurred task is needed to be processed first but most of the times it is un-desirable.

2) Improved Weighted Round Robin (IWRR) : IWRR assigns the tasks to the appropriate VMs based on the VMs information like the length of incoming tasks, the load on the VMs, processing capacity, and priority of the tasks. It employs a static scheduler, a dynamic scheduler, and the load balancer. The static scheduler of this algorithm takes the parameters like the length of each task, the no. of current and arriving tasks, and processing capacity of the VM into account and decides on the allocation to the most suitable VM. The dynamic scheduler of this algorithm, which is triggered at run time, uses the load on each of the VMs in addition to the above-mentioned parameters to select the most suitable VM for assigning the task. The load balancer rearranges the jobs according to the computed values in the queue of the underloaded VMs by migrating a job in the service queue of an overly loaded VM and passes this information to the scheduling controller to help the dynamic allocation. This algorithm is of interest because of its combination of the static and dynamic schedulers with the load balancer as the optimal scheduling is very important to reduce the migration overhead during the load balancing.

3) Opportunistic Load Balancing (OLB): Opportunistic Load Balancing (OLB) is to attempt each node keep busy, therefore does not consider the present workload of each computer. OLB assigns each task in free order to present node of useful.The advantage is quite simple and reach load balance but its shortcoming is that it is not considered the expectation execution time of the task, therefore the whole completion time (Makespan) is very poor. In other words, OLB dispatches unexecuted tasks to currently available nodes at random order, regardless of the node's current workload [13].

4) Task based load balancing (TLB): In this algorithm, the allocation of resources is done based on task to meet dynamic requirements of users. This algorithm achieves load balancing by first mapping tasks to VM and then all VM to host resources. It is improving the task response time and also provide better resource utilization. [14].

VI. System Model And Problem Formulation

To add more flexibility and scalability to the C-RAN, it can use the functionality which are used in the traditional datacenter, which is consist of a number of virtual machine control by some virtual machine manager(VMM) or a hypervisor. The C-RAN VM are real-time VM which can support different radio functionality and growth over a multi-radio supported hardware like software define radio(SDR) and general purpose processor (GPP). In the C-RAN number of service providers can use the same BBU-Pool by mutual authentication and service level agreement(SLA) [15], [16]. Each service provider either holding a single VM or multiple VMs based on the mutual agreement between service providers and the infrastructure providers.

In this paper, we have taken a number of assumption like :

- All service providers are having one or more VMs on a C-BBU pool.
- All VMs are co-operative to share their load with other VM.
- The user entities (UE) are connected to only a single RRH at a particular time.
- Packets data coming from the RRH through fronthaul to BBU are consider as task, which are further assigned to a particular VM for processing.

A. Mathematical Model

Let BBU-pool consists of a set of VMs for processing real-time data. Considering a set of RRHs, Rh = [R1, R2,...Rn] connected to a set of N number of VMs such as VM = [VM1, VM2,...VMn], process M packets represented by the set Pt = [P1, P2,...Pm]. Load of all VMs in a BBU-pool is depended upon the load of RRH unit connected to a particular VM. The RRH load it self-depends upon the number of users connected per unit time. So a total load of a VMs can define as:

$$L_{vm} = \sum\nolimits_{i \in |N|} L_i \tag{1}$$

where,
L_i = load of one VM
The load per unit capacity (L_c) is defined as:

$$L_c = \frac{L_{vm}}{\sum_{i \in |M|} Cp_i} \tag{2}$$

Then, threshold of each VM can be calculated as:

$$TH_k = L_c \times Cp_k \tag{3}$$

where,
Cp = capacity of VM.
The load imbalance factor of a particular VM can be defined as:

$$TH_k > \sum\nolimits_{i \in |Z|} l_i(UnderloadingCondition) \tag{4}$$

$$TH_k = \sum\nolimits_{i \in |Z|} l_i(BalanceCondition) \tag{5}$$

$$TH_k < \sum\nolimits_{i \in |Z|} l_i(OverloadingCondition) \tag{6}$$

where,
$\sum_{i \in |Z|} l_i$ = load generated by users coming through RRH.

VII. Proposed Heterogeneous Load Balancing Algorithm (HLB)

In this work, we proposed a dynamic load balancing algorithm for C-BBU which is more suitable for the real-time load balancing application. The HLB is more efficient than the earlier proposed algorithm such as IWRR, OLB, and TLB in term of processing and execution time. In the next generation data-centric network, like LTE and LTE-A system, the UE are assigned to the RRH based on the channel bandwidth and

resource block (RB) availability. As in the C-RAN, the core network (Fig. 2) are more busy to handle different real time signals such as voices and streaming video, it is necessary to give more preferences to real-time (R_t) packet than non-real times packets (NR_t). The BBU-Pool used a classifier to classify the incoming packet based on the following tuples of the packet like, sources index (SI), destination index (DI), and priority bit (PI). Unlike other algorithms, the HLB more focus on the priority bit instead focus on length of the packet such as in TLB. The HLB based load balancing is carried out based on VM capacity (C_p), load (L_i) on a VM_i and packet headers. When the load (L_i) on a particular VM is more than its capacity (C_p) then the next incoming load to the VM_i search for another VM known as VM_j based on the availability of physical and computational resources like bandwidth (BW), RAM, and task execution time as shown in Fig. 3 and described in Algorithm 1.

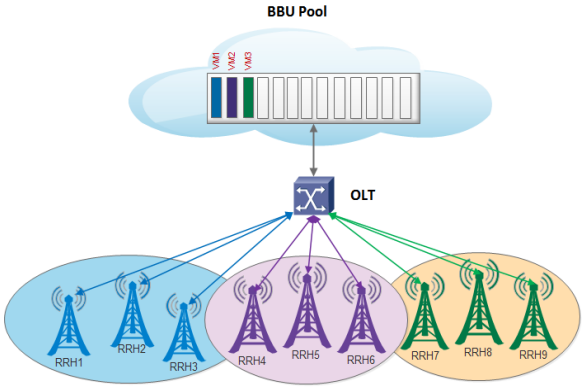

Fig. 2. C-RAN system model with Virtualized BBU-pool and Corresponding RRH assignment

A. Proposed Algorithm for Packet Classification and allocation to at C-BBU

Algorithm 1 HLB Packet Classification and Allocation Algorithm for C-BBU of C-RAN

procedure PROCEDURE
  Accept packets from RRHs
  classify incoming packets as R_t *and* NR_t
  repeat
  for each packet coming from RRHs
  $p \leftarrow$ *prefix of destination IP address of packet*
  if *p match with the prefix of any VM IP address* **then**
    if *PI* $==$ *defined priorities and* $Cp \geq L_i$ **then**
      classify that packet as R_t *and* NR_t
      forward the packets towards corresponding VM
    else
      search for new VM
  else
    Discard the packet.
End Procedure

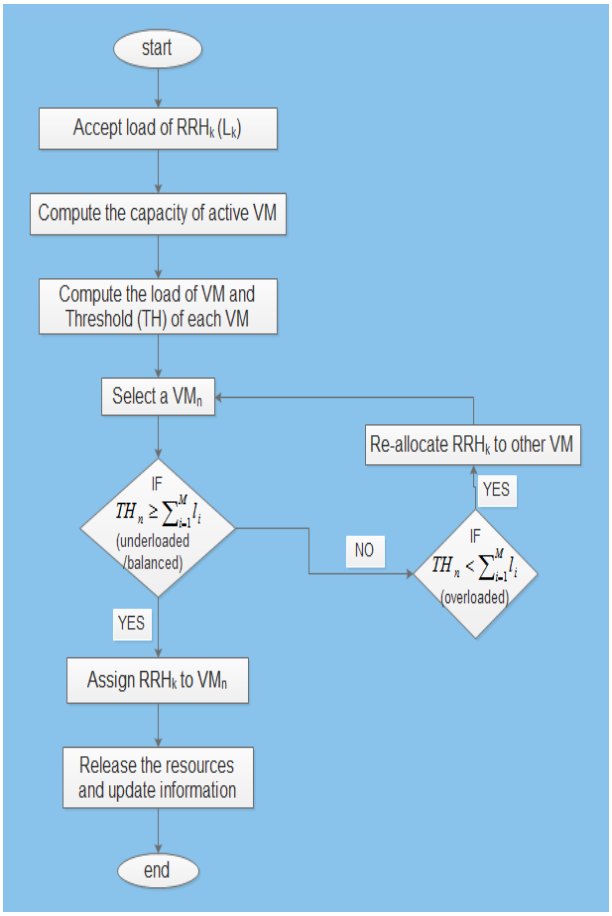

Fig. 3. Flow chart of proposed algorithm

VIII. Simulation and Performances Evaluation

The simulation is carried out with the help of Cloud-sim and Matlab simulator on a Linux environment. The main simulation parameters are the speed of VM and size of the incoming packets. We have created a host(BBU-Pool) and run a number of VMs over it. The incoming packet is assigned to the corresponding VM based on its header. The linear increase in packet arrival rate (R) leads to increase in the completion time or execution time (Te) in the VM as shown in Fig. 4. Taking the packet arrival constant and increase the number of VM as shown in Fig. 5, leads to decrease in the waiting time (Tw) and increase in the Te.

The results are given in Fig. 4 and Fig. 5 shows that the HLB is more effective to handle the real-time traffic. This proposed algorithm required less (Tw) to process the number of the incoming packet. The (Te) required for this proposed algorithm is very less for same packet arrival rate as compared to other mentioned algorithms. As the HLB required less Te and Tw, it is more energy efficient than the mention TLB and OLB algorithms.

IX. Conclusion

The load balancing approach in C-RAN is used to balance the traffic load and made the C-BBU more energy efficient

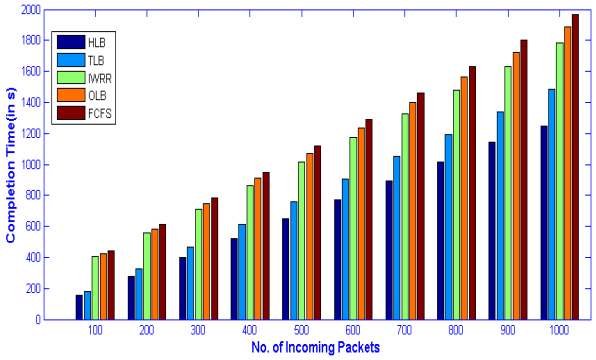

Fig. 4. Execution Time (Te) Vs Packet arrival (R) for linear and hierarchical searching technique

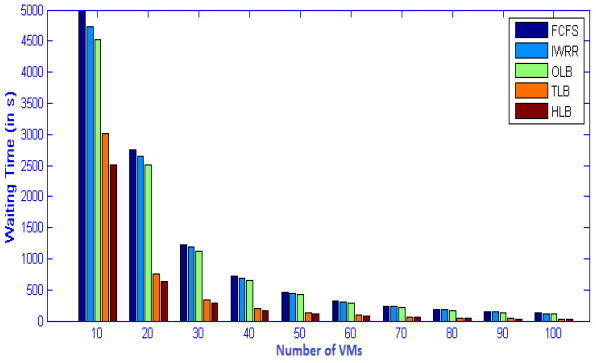

Fig. 5. Waiting Time (Tw) Vs Packet arrival (R) by using linear and hierarchical searching technique at C-BBU

and cost-effective. The performance analysis and experiment results of the HLB algorithm proved that it is suitable to handle the homogeneous/heterogeneous jobs with help of scalable resources (VMs) compared to the remaining algorithms. This makes them feasible to use in the cloud environment and in turn in C-RAN architecture. The above algorithms reduced the usage of the resources which are valuable in mobile networks. The deployment of these algorithms in the C-RAN environment can be helpful in balancing the load imposed by the number of signals originating from different base stations.

REFERENCES

[1] I. Chih-Lin,C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, Toward green and soft: a 5G perspective, IEEE Communications Magazine, vol. 52(2), pp.66-73, 2014.

[2] A. Checko, L. Henrik, Christiansen, Y. Yan, L. Scolari, G. Kardaras, S. Michael, and L. Dittmann, Cloud RAN for Mobile Networks-A Technology Overview, IEEE Communication surveys tutorials, vol.17, no.1, 2015.

[3] A. Feldman and S. Muthukrishnan. Tradeoffs for packet classification, Proceedings of Infocom, vol. 3, pages 1193-202, March 2000.

[4] Cisco visual networking index: Global mobile data traffic forecast update 2016.

[5] M. Artuso and H. Christiansen, "Fronthaul dimensioning in C-RAN with web traffic for coordinated multipoint joint transmission," Communication Workshop (ICCW), 2015 IEEE International Conference, pp. 50-55, vol.2, sept. 2015,.

[6] A. Revar, M. Andhariya, D. Sutariya, M. Bhavsar, Load balancing in grid environment using machine learning-innovative approach, International Journal of Computer Applications, vol.2, no.3, pp.975-980, 2010.

[7] Wang, Pi-Chung. "Scalable packet classification for data center networks." IEEE Journal on Selected Areas in Communications, vol. no. 1, pp.124-137, 2014.

[8] P. Gupta, and N. McKeown, Algorithms for packet classification, IEEE Network, 15(2), pp.24-32, 2010.

[9] J. Cao, K. Li, and I. Stojmenovic, Optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and data centres, IEEE Transactions on Computers, vol.63, no.1, pp.45-58, 2014.

[10] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, Wireless network cloud: Architecture and system requirements, IBM Journal of Research and Development, 2010.

[11] R. Basker, V. Rhymend Uthariaraj, and D. Chitra Devi, An enhanced scheduling in weighted round robin for the cloud infrastructure services, International Journal of Recent Advance in Engineering & Technology, vol.2, no.3, pp.81-86, 2014.

[12] M. Rahman, R. Hassan, R. Ranjan, and R. Buyya, Adaptive workflow scheduling for dynamic grid and cloud computing environment, Concurrency and Computation: Practice and Experience, vol.25, no.13, pp.1816-1842, 2013.

[13] S Wang, K. Q. Yan, W. Liao and S. Wang, "Towards a Load Balancing in a three-level cloud computing network," International Conference on Computer Science and Information Technology, Chengdu, pp. 108-113, 2010.

[14] Y. Fang, F. Wang, J. Ge," A task scheduling algorithm based on load balancing in cloud computing," Web Information Systems and Mining, PP.271-277,2010.

[15] R.Chen, S.Wang, etc. Optimal load balancing in Cloud Radio Access Networks, IEEE Wireless Communication and Network Conferences (WCNC), pp 1006-1011, Mar 2015.

[16] D. Mishra, P.C Amogh. A. Ramamurthy, A. Franklin, A. and B.R Tamma, 2016, Load-aware dynamic RRH assignment in Cloud Radio Access Networks, IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-6.vol.2, 2016.

[17] H. Yang, J.Zhang, etc, Multi-stratum resources optimization for cloud-based radio over optical fiber networks, IEEE International Conference on Communication, pp. 1-5, May 2016.

[18] C Parada Mobile Cloud Networking, Algorithms and mechanisms for the mobile network cloud, A Project Report, 2014.

[19] S. Namba, T. Warabino,and S. Kaneko," BBU-RRH switching schemes for centralized RAN", Communications and Networking in China (CHINACOM), pp. 762-766, 2012.

.