Networked Wireless Systems Lab - IIT Hyderabad

Load-Aware Dynamic RRH Assignment in Cloud Radio Access Networks

Debashisha Mishra, Amogh PC, Arun Ramamurthy, Antony Franklin A, Bheemarjuna Reddy Tamma

Dept. of CSE, Indian Institute Of Technology, Hyderabad, INDIA

IEEE WCNC
April, 2016

Introduction

Challenges in traditional cellular system

- Traffic inhomogeneity → Diurnal Base station (BS) utilization.
- To meet data requirement by mobile users → Deploy more BSs
- Bigger landscape for Macro BS → High CAPEX and OPEX
- Unattractive Average Revenue per User (ARPU) → BS evolution

Cloud Radio Access Network / Centralized RAN / C-RAN

- Functional Split of base stations into Remote Radio Head (RRH) and Baseband Unit (BBU)
- RRH at cell site with a much smaller footprint than traditional BS
- BBU at a centralized data center few tens of KMs away

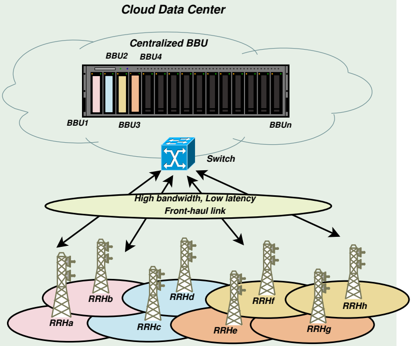
Cloud RAN

An architectural Overview

Key Terminology

- RRH - With some part of RF circuitry, Geographically distributed
- BBU - Digital signal processing for compute intensive tasks hosted on standard IT server
- Fronthaul - High bandwidth and low latency communication medium between BBU and RRH
- IQ Sample - Inphase Quadrature sample transmission on fronthaul (*i.e.*,CPRI,OBSAI,ORI)

Fig 1 : Cloud RAN Architecture



Motivation

- Many-to-One Dynamic Mapping from RRH to BBU
- Reduced Network CAPEX and OPEX → Operator's preferred choice

Fig 2 : RRH Assignment - Scene 1

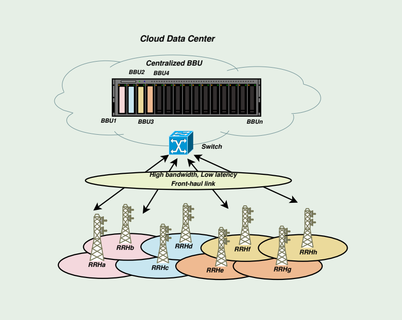
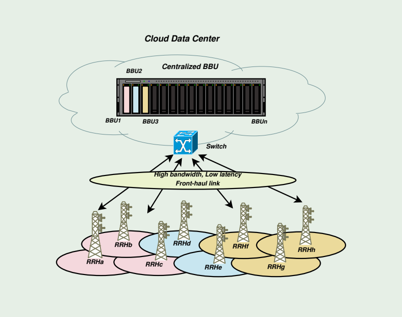


Fig 3 : RRH Assignment - Scene 2



RRH to BBU Assignment

A generalization of Bin Packing Problem

What is Bin Packing Problem (BPP) ?

- Given, bins of a fixed size, insert items into the bins to maximize the individual bin space utilization with goal of minimizing number of bins
- Exact optimal assignment solution is NP-hard, Use various heuristics and approximation procedures to find a near optimal solution

First Fit Decreasing (FFD) Scheme for BPP

- Provides near optimal clustering/assignment of items into the bins
- Number of used bins are bounded by $\left(\frac{11}{9} \times OPT + \frac{6}{9}\right)$ with OPT being the optimal number of bins
- Very efficient in terms of computation time complexity

Contribution of the Work

Disadvantages of FFD

- FFD is not suitable for practical RRH to BBU mapping in Cloud RAN
- FFD algorithm has to be run frequently in order to capture traffic inhomogeneity of mobile subscribers with fine granularity of time

Major Contributions

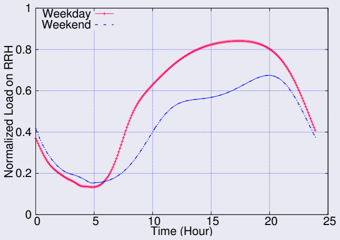
- Efficient, light-weight, and load-aware dynamic RRH assignment (DRA) algorithm for many-to-one mapping of RRHs to BBU
- Performance comparison of proposed DRA algorithm with FFD
- Extensive simulation experiments for an urban network area (200 RRHs to 1000 RRHs) including both weekday and weekend traffic profiles

Load Characterization at RRH

Spatio-temporal Pattern of Traffic Load

- **Temporal Load Pattern**
  Load on individual base stations do not follow any periodicity, however the trend is consistent with diurnal activity patterns of human beings.

Time Varying Nature



Load Characterization at RRH

Spatial Load Pattern
The residential zones tend to be active in off-hours (nights, weekends and holidays) while business or office areas are active during daytime in weekdays.

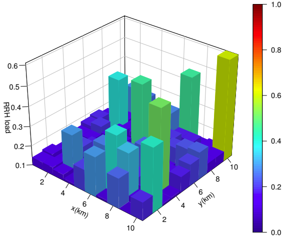


Figure : Weekend Spatial Plot

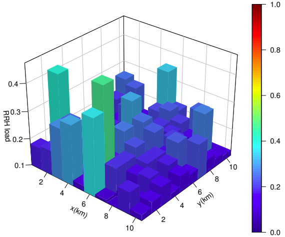


Figure : Weekday Spatial Plot

System Model

Probabilistic Distribution of RRH Loads

- Let l_i be the load on i^{th} RRH, then, value of l_i is an exponential random variable with mean value of $\frac{1}{\lambda}$.
- The probability density function of an exponential distribution is expressed by

$$f(t) = \lambda e^{-\lambda t}, t \geq 0 \tag{1}$$

  where λ is the rate parameter and mean value is $\frac{1}{\lambda}$.
- The GMM used for modeling of time-varying rate parameter is given by

$$\lambda = \sum_{i=1}^{n} a_i e^{-\left(\frac{t-b_i}{c_i}\right)^2} \tag{2}$$

  where a_i is the amplitude, b_i is centroid location, c_i is the peak width, n represents number of peaks in data series and t is any time instant in 24 hours of the day.
- Using these two models, we generated snapshots of spatial loads on each of RRHs of whole network under study for a given time instant.

System Model
BBU Model

- Processing load incurred by a UE at a BBU in the BBU pool in a given Transmission Time Interval (TTI) depends on the number of Resource Blocks (RBs) allocated to that UE and their corresponding Modulation and Coding Scheme (MCS) values
- Overall processing load of an RRH on BBU in a given TTI is the summation of processing loads of all the UEs connected to that RRH in that TTI
- A single BBU may be capable of processing peak processing loads of one or more RRHs

System Model
Energy Model

- The total power consumed at any active BBU at time instant t is

$$P_{BBU_t} = P_{BB} + \sum_{i=1}^{n} P_{RRH_i} \tag{3}$$

where P_{BB} is power consumed by that particular BBU and $\sum_{i=1}^{n} P_{RRH_i}$ is sum of the power consumed by all the RRHs associated with that particular BBU at time t.

- The power consumed by a specified RRH is given by

$$P_{RRH} = \left(\frac{P_r}{E_{pa}} \right) + (P_{rf} \times N_{tx}) \tag{4}$$

where P_r is the radiated power, E_{pa} is the power amplifier efficiency, P_{rf} is the power used by RF circuits and N_{tx} is the number of transceiver antennas.

System Model
Cluster Optimization as an Integer Linear Programming model

Notation	Definition
N	Set of all RRHs
M	Set of all BBUs in BBU Pool
l_i	Processing load incurred by RRH i
z_m	1 if BBU m is active; otherwise 0
y_{im}	1 if RRH i is associated with BBU m; otherwise 0
l_{max}	Peak BBU capacity

Objective Function: Minimize $\sum_{m=1}^{|M|} z_m$.

Constraints:

$$\sum_{m=1}^{|M|} y_{im} = 1, \qquad \forall i \in N \tag{5}$$

$$\sum_{i=1}^{|N|} y_{im} \times l_i \leq l_{max} \times z_m, \qquad \forall m \in M \tag{6}$$

- Eqn 5 - Each RRH is associated with exactly one BBU.
- Eqn 6 - Sum of loads from RRHs associated to a BBU does not exceed the BBU peak capacity.

Dynamic RRH Assignment

Concepts and Terminology

- Basic principle of this algorithm is to offload one or more RRHs (known as *Candidate_RRH(s)*) from an overloaded BBU to a less loaded BBU with enough available computation capacity to accommodate the incoming RRH(s)
- Assume δt is the periodicity of cluster formation; $\delta t \rightarrow 0$
- Partition P_{mod} is the partition containing clusters which need reassignment of RRHs
- l_{max} is the maximum load a BBU can handle.
- r is the total number of RRHs associated with a given cluster
- l_i is the load on i^{th} associated RRH
- $spill\_load$ for a given cluster is the excess spill amount. Mathematically, $spill\_load = ((\sum_{i=1}^{r} l_i) - l_{max})$
- $$spill\_load = \begin{cases} > 0, & \text{spill\_cluster} \\ \leq 0, & \text{non\_spill\_cluster} \end{cases} \quad (7)$$

Dynamic RRH Assignment

Algorithm Design & Complexity Measures

1: Based on $spill\_load$, classify the clusters in P_{mod} as spill_cluster and non_spill_cluster

2: In case, no spill_cluster is found, just return P_{mod}

3: In case, all clusters are spill_cluster, perform FFD on processing loads contained in P_{mod} into BBU. This situation is too unrealistic to occur in practice when $\delta t \to 0$

4: For each spill_cluster, find *Candidate_RRH(s)* to offload using *Offload_Selection* subroutine

5: Find a non_spill_cluster with enough resources to accommodate *Candidate_RRH*, perform RRH assignment to it

6: If no such non_spill_cluster, get a new BBU from BBU pool and perform the RRH assignment to the new BBU

7: Apply merge procedure on non_spill_clusters using FFD and return new partition P_{new}

Test Input	FFD	DRA	Scenario
Best Case	$O(\|N\|log\|N\|)$	$O(\|N\|)$	No spill_cluster
Average Case	$O(\|N\|^2)$	$O(\|N\| + \|k\|^2)$	Both spill_clusters and non_spill_clusters
Worst Case	$O(\|N\|^2)$	$O(\|N\|^2)$	All are spill_clusters

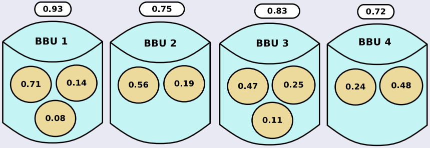
Dynamic RRH Assignment

Numerical illustration

Partition P at time t
$C_1 = \{0.71, 0.14, 0.08\}, \sum C_1 = 0.93$
$C_2 = \{0.56, 0.19\}, \sum C_2 = 0.75$
$C_3 = \{0.47, 0.25, 0.11\}, \sum C_3 = 0.83$
$C_4 = \{0.24, 0.48\}, \sum C_4 = 0.72$
Partition P_{mod} at time $(t + \delta t)$
$C_{1mod} = \{0.56, 0.14, 0.40\}, \sum C_{1mod} = 1.10$ (spill_cluster)
$C_{2mod} = \{0.31, 0.48\}, \sum C_{2mod} = 0.79$ (non_spill_cluster)
$C_{3mod} = \{0.21, 0.39, 0.80\}, \sum C_{3mod} = 1.40$ (spill_cluster)
$C_{4mod} = \{0.11, 0.24\}, \sum C_{4mod} = 0.35$ (non_spill_cluster)
Partition P_{new} at time $(t + \delta t)$ after DRA algorithm
$C_{1new} = \{0.56, 0.40\}, \sum C_{1new} = 0.96$
$C_{2new} = \{0.31, 0.48, 0.14\}, \sum C_{2new} = 0.93$
$C_{3new} = \{0.21, 0.39\}, \sum C_{3new} = 0.60$
$C_{4new} = \{0.11, 0.24\}, \sum C_{4new} = 0.35$
$C_{5new} = \{0.80\}, \sum C_{5new} = 0.80$
After FFD, $C_{3,4new} = \{0.21, 0.39, 0.11, 0.24\}, \sum C_{3,4new} = 0.95$

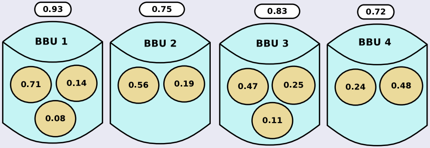
Algorithm illustration

An optimal assignment at time t

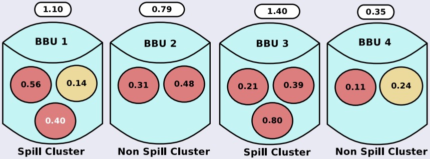


Algorithm illustration

An optimal assignment at time t

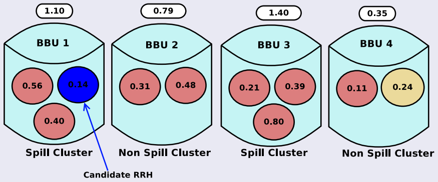


At time $(t + \delta t)$ some RRH loads are modified



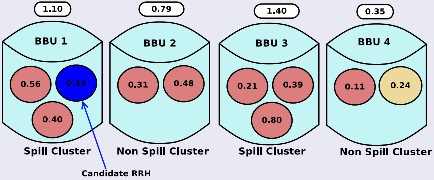
Algorithm illustration

Choose a Candidate RRH

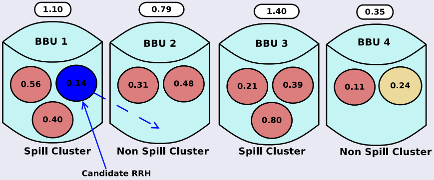


Algorithm illustration

Choose a Candidate RRH as offloading candidate



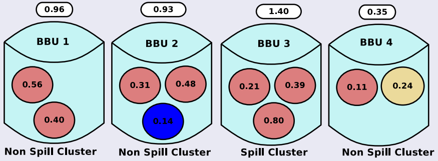
Find a Place



Algorithm illustration

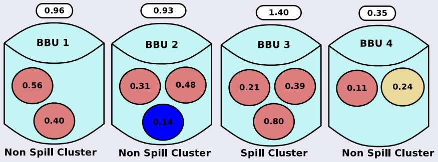
Load-Aware Dynamic RRH Assignment in Cloud RAN

Assign the RRH to Destined BBU

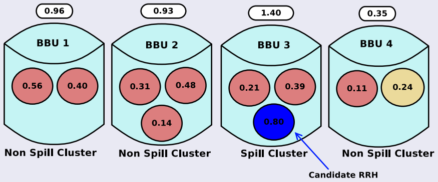


Algorithm illustration

Assign the RRH to Destined BBU



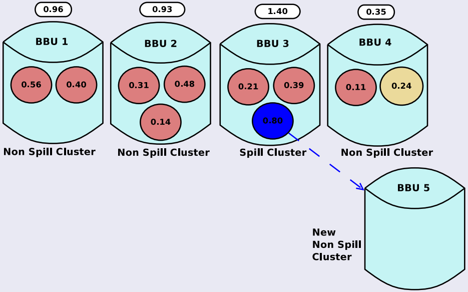
Proceed ...



Algorithm illustration

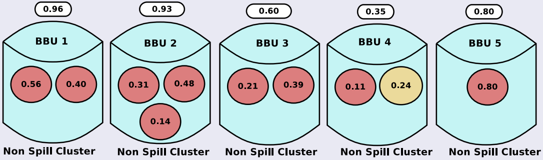
Load-Aware Dynamic RRH Assignment in Cloud RAN

Switch ON a new BBU



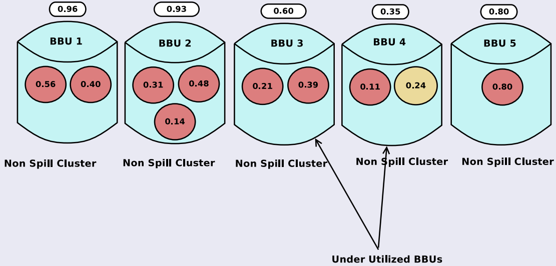
Algorithm illustration

After RRH Assignment



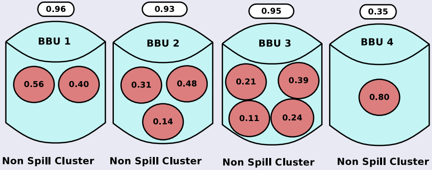
Algorithm illustration

Not Optimal Cluster - Apply FFD Merge on Clusters



Algorithm illustration

End of Procedure



- Total Number of Clusters $= 4$.
- Each BBU is optimally utilized.
- Simple re-assignment schemes over costly FFD.

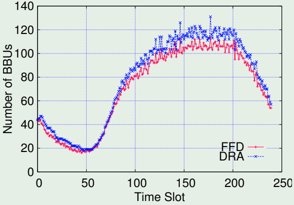
Simulation Setup

- Simulations on a commodity hardware having 64-bit Ubuntu Linux on x86, Intel 4 core, 1.7 GHz processor
- Study of System performance in terms of the computational resource gain (number of active BBUs used)
- Individual RRH loads are generated using probabilistic distribution model which vary over space and time dimensions

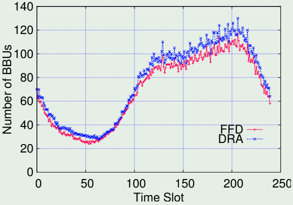
Parameter	**Value**
Number of RRHs	200 to 1000
Sampling periodicity	6 minutes
Traffic duration	24 Hours
Total number of samples	240
Traffic profile	Weekday, Weekend
Geographical region	Urban
RRH load range	[0,1] (0 to 100%)
Maximum Load on BBU	1 (100%)
Spatial load distribution	Exponential
Time-varying rate parameter	Gaussian Mixture Model

Performance Evaluation

Variation in BBUs usage for a weekday with 1000 RRHs



Variation in BBUs usage for a weekend with 1000 RRHs.



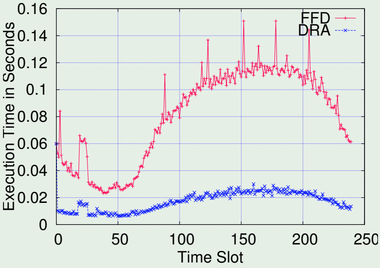
- DRA reduces the required number of BBUs by 87% compared with 1:1 RRH to BBU mapping scheme
- DRA over estimates FFD by 1.7% and 1.4% on weekday and weekend, respectively

Performance Evaluation (Contd ...)

Observation...

- Total running time of the algorithm for 1000 RRHs is order of milliseconds
- With increase in number of RRHs, C-RAN needs to deploy more number of BBUs for serving them
- More than 90% energy savings for C-RAN
- Larger pool size offers more energy saving opportunities

CPU time taken in each δt for 1000 RRHs

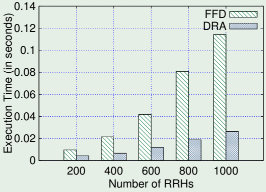


Performance Evaluation (Contd ...)

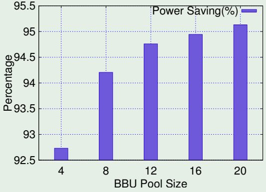
Load-Aware Dynamic RRH Assignment in Cloud RAN

- Introduction
- System Model and Problem Formulation
- Proposed Work
- Performance Evaluation
- Conclusions and Future Work

CPU time taken in each interval for 1000 RRHs



Exec times of FFD and DRA at 16.00 hour on weekday w.r.t varying RRHs



Conclusions and Future Work

- Analyzed and quantified the BBU resource savings and time complexity measures of DRA in contrast to FFD considering spatio-temporal traffic variations from base stations.
- The savings trend follows a diurnal human traffic pattern.
- As part of ongoing work, we aim to define various dependent factors such as UE position, cell edge constraints, BS cooperation (e.g., CoMP) in the processing load characterization of RRH for quantifying savings.

References I

Cisco , *"Visual Networking Index, Global mobile data traffic forecast update", White Paper, 2014*.

"C-RAN - The road towards green RAN", China Mobile White Paper, Vol 2, 2011.

Bhaumik, Sourjya and Chandrabose, Shoban Preeth and Jataprolu, Manjunath Kashyap and Kumar, Gautam and Muralidhar, Anand and Polakos, Paul and Srinivasan, Vikram and Woo, Thomas, *"CloudIQ: a framework for processing base stations in a data center", Proceedings of MOBICOM,125–136*, ACM,2012.

Checko, Aleksandra and Christiansen, Henrik L and Yan, Ying and Scolari, Lara and Kardaras, Georgios and Berger, Michael S and Dittmann, Lars, *"Cloud RAN for mobile networks - a technology overview", IEEE Communications Surveys & Tutorials, Vol 17, 405–426*, 2014.

Paul, Utpal and Subramanian, Anand Prabhu and Buddhikot, Milind Madhav and Das, Samir R, *"Understanding traffic dynamics in cellular data networks", Proceedings IEEE INFOCOM,882–890*, 2011.

References II

Liu, Jingchu and Zhou, Sheng and Gong, Jie and Niu, Zhisheng and Xu, Shugong, *"On the statistical multiplexing gain of virtual base station pools"*, *IEEE GLOBECOM, 2283–2288*, 2014

Zhu, Dalin and Lei, Ming, *"Traffic and interference-aware dynamic BBU-RRU mapping in C-RAN TDD with cross-subframe coordinated scheduling/beamforming,"*, *IEEE ICC Workshops, 884–889*, 2013.

Chen, Xi and Li, Na and Wang, Jing and Xing, Chengwen and Sun, Liang and Lei, Ming, *"A Dynamic Clustering Algorithm Design for C-RAN Based on Multi-Objective Optimization Theory"*, *IEEE VTC Spring,1–5*, 2014.

Checko, Aleksandra and Christiansen, Henrik Lehrmann and Berger, Michael Stubert, *"Evaluation of energy and cost savings in mobile Cloud RAN"*, *OPNETWORK*, 2013.

Wang, Huandong and Ding, Jingtao and Li, Yong and Hui, Pan and Yuan, Jian and Jin, Depeng, *"Characterizing the Spatio-Temporal Inhomogeneity of Mobile Traffic in Large-scale Cellular Data Networks"*, *HOTPOST '15, 10.1145/2757513.2757518, 19–24*, ACM 2014.

References III

Nan, E and Chu, Xiaoli and Guo, Weisi and Zhang, Jie, *"User data traffic analysis for 3G cellular networks"*, *IEEE CHINACOM,468–472*, 2013.

Khan, M and Alhumaima, RS and Al-Raweshidy, HS, *"Reducing energy consumption by dynamic resource allocation in C-RAN"*, *EuCNC,169–174*, IEEE 2015.

Falkenauer, Emanuel and Delchambre, Alain, *"A genetic algorithm for bin packing and line balancing"*, *International Conference on Robotics and Automation,1186–1192*, IEEE 1992.

Dosa, Gyorgy, *"The tight bound of first fit decreasing bin-packing algorithm is $FFD(I) \leq 11/9 \times OPT(I) + 6/9$"*, *Combinatorics, Algorithms, Probabilistic and Experimental Methodologies*, Springer 2007.

Thank You

QUERIES ?