Evaluation of energy and cost savings in mobile Cloud RAN

Checko, Aleksandra; Christiansen, Henrik Lehrmann; Berger, Michael Stübert

Published in:
Proceedings of OPNETWORK 2013

Publication date:
2013

Evaluation of energy and cost savings in mobile Cloud RAN

Aleksandra Checko, Henrik Christiansen, Michael S. Berger
DTU Fotonik, Technical University of Denmark, Kgs. Lyngby, Denmark
E-mail: amch@fotonik.dtu.dk

Abstract

The load in mobile networks is subject to variations during the day, due to user mobility and varying network average usage. Therefore, the traditional or Distributed Radio Access Network (D-RAN) architecture, where the BaseBand processing Units (BBUs) are assigned statically to a number of cells, is sub optimal, comparing to a novel, cloud based architecture called Cloud Radio Access Network (C-RAN). In C-RAN a group of cells shares processing resources, and hence benefit from statistical multiplexing gain is expected.

In this paper, the energy and cost savings in C-RAN are evaluated numerically using OPNET Modeler. A real case scenario is built upon the mobile traffic forecast for year 2017, a number of recommendations on traffic models and a proposed C-RAN implementation.

The results achieved show that the maximum statistical multiplexing gain for user plane traffic in C-RAN architecture is 4 compared to a traditional and D-RAN architecture.

1. Introduction

Mobile data traffic is expected to increase 13-fold from 2012 until 2017, according to [1]. Therefore, in order to support such traffic, mobile network operators are forced to increase network resources. This results in significant Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) increase, due to the fact that more or more powerful equipment needs to be deployed, which consequently increases overall network cost and electricity consumption. Meanwhile, the Average Revenue Per User (ARPU) stays flat or even decreases over time, as the typical user gets more and more data-hungry but expects to pay less for data usage. As presented by Juniper (Figure 1), mobile operators are facing cases (2014-1015) where network cost may exceed revenues if no actions will be taken. Thus, a solution which can maintain the network Total Cost of Ownership (TCO) at a reasonable level becomes of utter importance.

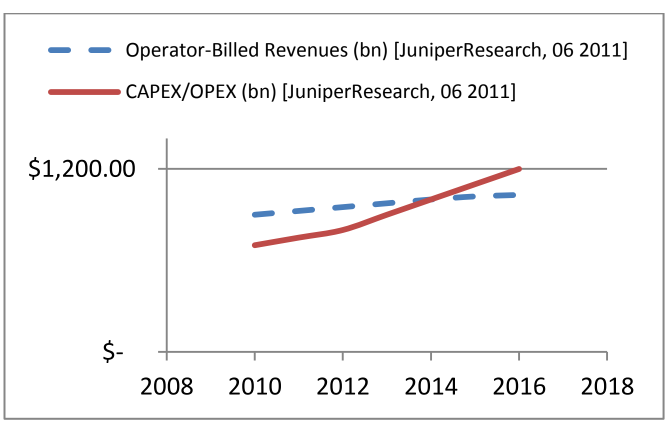

Figure 1 Network costs exceed revenue

Mobile C-RAN has the potential to lower the network cost and energy consumption. The main goal of this paper is to present our approach of modeling Statistical Multiplexing Gain in C-RAN, which is coming from user mobility. We model traffic flows in a future mobile network and assess how many fewer BBUs will be needed in a C-RAN architecture compared to D-RAN to meet future user requirements.

The paper is organized as follows. In Section II we introduce the concept of C-RAN. In Section III we present the approach used to model data traffic in a future mobile network. In Section IV we describe our OPNET model of C-RAN. In Section V we evaluate the cost and energy consumption reduction in C-RAN and compare it to a traditional/D-RAN. In Section VI we present our conclusions and recommendations derived from this work.

2. C-RAN Architecture Overview

Traditionally, in cellular network, users communicate with a Base Station (BS) that is statically assigned to them. Radio and baseband processing units are located close to each other (few meters) as lossy RF cables are used to connect modules. In Distributed architecture, known as D-RAN, the BS is separated into radio unit and signal processing unit, as shown in Figure 2. Radio unit is called Remote Radio Head (RRH) and performs digital signal processing, digital to analog conversion, power amplification, filtering and optical conversion. The signal processing part is called BaseBand Unit (BBU). Common Public Radio Interface (CPRI) [8] is the radio interface protocol widely used for data transmission between RRH and BBU. Distance between RHH and BBU can span up to 40 km, thanks to the use of lossless optical cables. In distributed architecture RRHs are statically assigned to BBUs in the same way as in the traditional one.

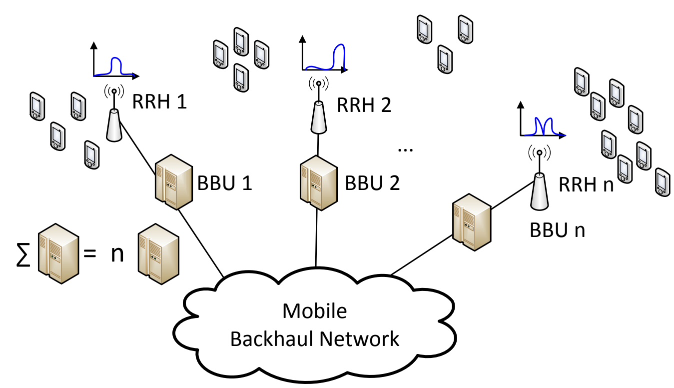

Figure 2 D-RAN architecture

C-RAN was first proposed in [2] and widely described in [3]. Figure 3 shows this novel base station architecture, where baseband processing is shared among sites in a virtualized BaseBand Unit Pool (BBU Pool). C-RAN aims at reducing power and inter-cell interference, as well as improving performance, energy efficiency, scalability, upgradeability and utilization of base stations compared to D-RAN architecture [3]. Moreover, this novel architecture has also the potential to decrease cost of network operation.

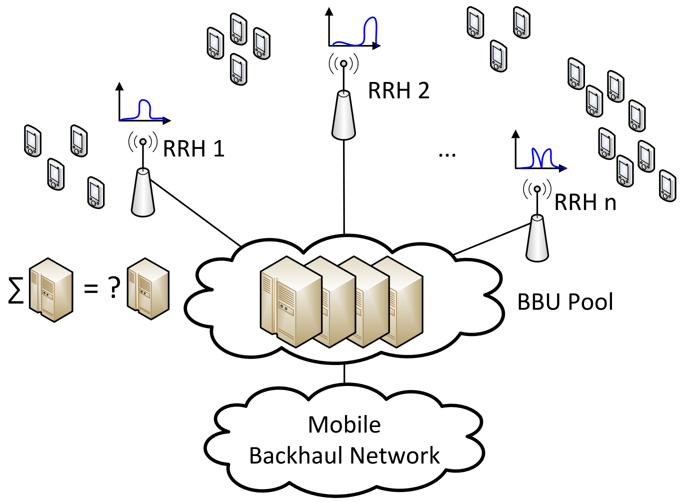

Figure 3 C-RAN architecture

In C-RAN several units of BBU hardware, collocated in the same pool, are perceived as a one entity from the outside, and internally can be flexibly adjusted to adapt for varying load in many cells. This is more optimal compared to traditional or distributed architecture, where the assignment is static. Later on in the paper we will refer to the comparison between C-RAN and D-RAN, however it also applies to comparison of C-RAN and traditional architecture.

C-RAN efficiently addresses the so called "tidal effect", meaning that the load of base stations depends on which type of area they are situated and fluctuates throughout the day. For example, base stations located in office areas will experience the highest load in the morning and after lunchtime, while residential ones in late evening. Commercial areas will experience the highest traffic load during lunch break and in late afternoon. Sharing BB resources for those different zones will introduce a Statistical Multiplexing Gain, as fewer BBUs will be needed to serve a particular area compared to the traditional case. This paper numerically evaluates this gain, which will have an impact on both cost and energy savings. However, the need in C-RAN for higher transmission bit rates between the BBU and RRH may, on the contrary, increase the expenses. The overhead of L1 data to application data might reach 270 times.

3. Mobile Traffic Modeling

In order to model traffic in C-RAN architecture, the following inputs are essential:

- Traffic forecast allowing evaluation of traffic volume processed by each Base Station.
- Traffic models that will reflect the nature of different applications (e.g. data, video, file sharing, etc).
- Data on daily variation of network load for different types of areas.

Cisco has published a traffic forecast [1] for 2017, where both data volume per world region and per application are stated, as presented in Table 1. Next Generation Mobile Networks alliance in [4] has also given recommendations on how to evaluate network performance, specifying among others traffic mix and defining specific applications. However, for the purpose of this project, Cisco forecast has the advantage of providing the actual traffic volume, therefore this one was chosen for our simulations. Nevertheless, the traffic models developed were inspired by NGMN.

Table 1 Global Mobile Data Traffic in 2017 [1]

Application	TB per month	%
Web/Data	2,778,386	24.91 %
File sharing	395,342	3.54 %
Video	7,418,322	66.50 %
M2M	563,481	5.05 %

We take as an example the city of Cologne (Germany), knowing than there are around 2000 Base Stations [6]. According to [1], in 2017 there will be 10B mobile subscribers in the world. By collecting data on population, we could make a projection for the number of mobile subscribers in Cologne in 2017. Knowing the number of Base Stations in Cologne and assuming that the average traffic is the same in each base station, the daily traffic volume per base station was calculated. Average mobile user will transmit and receive 75.8 MB of mobile data daily, which implies a monthly consumption of 2.3 GB. Those numbers do not include signaling - control messages.

Assuming that 40 % of the traffic is in uplink direction, uplink daily traffic volume per application for 10 BSs has been calculated. Ratio is an exemplary value and was inspired by different structures of LTE Time Division Duplex (TDD) frames-one out of seven possible configurations assigns 4 out of 10 sub-frames for UL. The results obtained in the paper are relative, therefore this number doesn't influence the results obtained in Section 5.

Web/Data, File sharing and Video applications are modeled, as their traffic characteristics will considerably influence the aggregated traffic characteristics. Table 2 summarizes the models used in simulation.

Table 2 Traffic Models

Traffic Parameters	Statistical Characterization
Web/Data [5]	
Web Page Size	Lognormal Distribution Mean = 321979 b, Variance = 10.91 Gb, corresponding to Standard Deviation = 413 kB
Aggregated Interpage Request Time	Exponential Distribution Mean = 1.1471s
File sharing – FTP [4]	
File Size	Lognormal Distribution Mean = 2MB, Standard Deviation = 0.722 MB
Reading Time	Exponential Distribution Mean = 180 s
Mobile Video [4]	
Inter-Arrival time between the beginning of each frame	Deterministic 100 ms (based on 10 frames per second)

Number of packets (slices) in a frame	Deterministic, 8 packets per frame
Packet (slice) size	Truncated Pareto Distribution Mean = 100 B, Maximum = 250 B, Minimum = 20 B, Shape α = 1.2, and a location x_m = 133 calculated based on above
Inter-arrival time between packets (slices) in a frame	Truncated Pareto Distribution Mean = 6 s, Maximum = 12.5 s, Minimum = 2.5 ms, Shape α = 1.2, and a location x_m = 1 calculated based on above

No recommendation on Machine2Machine (M2M) traffic modeling from 3GPP or NGMN has been found so far. This type of traffic is usually generated by networks of measuring or sensor devices. Since they usually send short updates or alarms, data are expected to be uniformly distributed and small in size. This means that they will not influence significantly the overall traffic profile.

In [3] a daily load of Base Stations in Residential and Office areas is presented, as shown on Figure 4. Such a distribution is proposed for the third area type, the Commercial area, as presented in Figure 4. For the BS located in the Commercial area, the relative load is smaller than for the former ones, since small cells are expected to be deployed. The highest load of Base Stations in Commercial areas is during lunch time and in the evening when people spend time after work for example/doing shopping on their way home. In the beginning of the night, and up to 04 o'clock the load is nonzero, considering small number of people staying in public places.

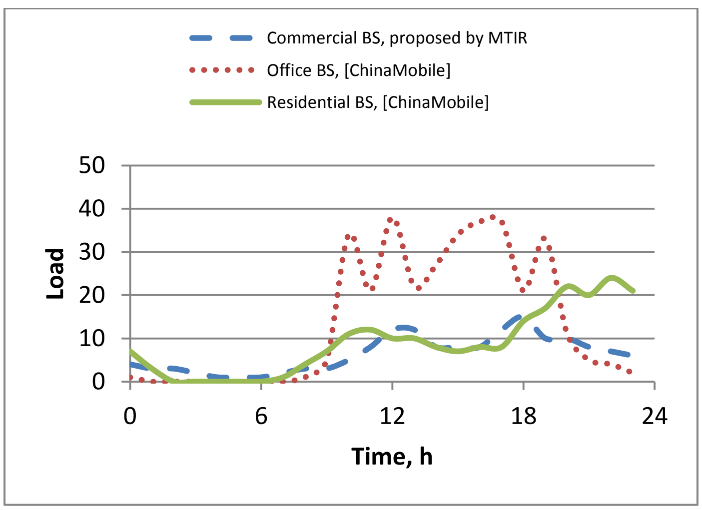

Figure 4 Daily Load on different types of BSs

The simulations have been performed under the assumption that in the city there are 30% of Office BS, 50% of Residential BS, and 20% of Commercial BS. Those are exemplary values and should be studied separately for each real-case scenario. The daily load in bits has been calculated per hour. The load is then mapped into the traffic volume for each application, using as a reference the Cisco forecast and following the daily traffic load distribution presented in Figure 4.

4. An OPNET D-RAN and C-RAN Model

To model user data processing gain energy and cost savings in C-RAN, two scenarios examining both C-RAN and D-RAN architectures have been created, as presented in Figure 5 and Figure 6, respectively.

For the D-RAN scenario, throughput has been measured for each BBU. Peak values of the throughput for each BBU were added and considered as a reference value against which peak value of throughput in C-RAN is compared to.

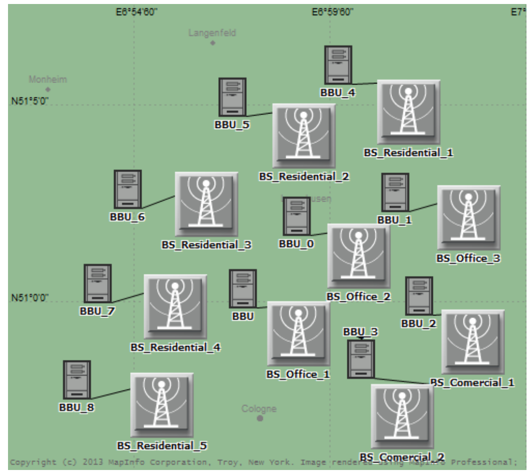

Figure 5 D-RAN Simulation Scenario

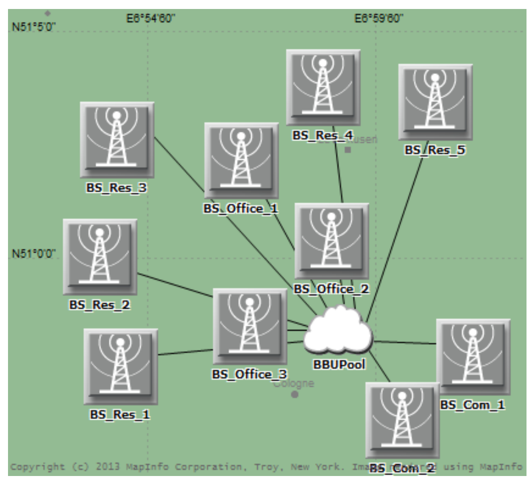

Figure 6 C-RAN Simulation Scenario

For the C-RAN scenario, peak of aggregated throughput at BBU Pool is measured. Its value is then compared with the result obtained from the D-RAN scenario. The comparison between the two results allows us to assess the relative savings in terms of the number of BBUs and their power consumption necessary to meet user requirements.

In both scenarios, two types of nodes are used. First is a Base Station constituting of three types of traffic generators, with specific models described in the section below, a common transmitter with four channels and a receiver as shown in Figure 7. The transmitter has a separate channel defined for each application type. The node model is suited for adding other traffic types as well as duplex transmission. It should be noted that the model in the current shape can be used for modeling both UL and DL transmission, however only in simplex mode. The node attribute "Cell type" is also defined, which can be of a value "Office", "Residential" or "Commercial", in order to define which type of area the BS is situated in. This parameter can be set for each Base Station from the list in node attributes at scenario level. Based on this parameter a proper file defining traffic load distribution through the day is used in the simulation. Separate files have been created for each Base Station type and application types providing the number of single traffic generators which should be running in each base station to meet Cisco forecast and desired daily load distribution.

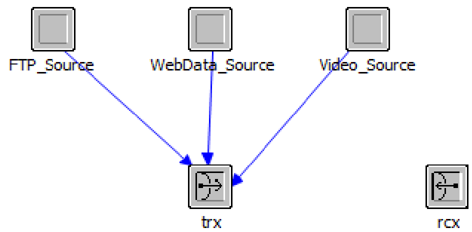

Figure 7 Base Station Node

Second node simulates BBU and BBU Pool. Both instances are modeled as a traffic sink with a separate receiver for each BS, as shown in Figure 8, and transmitters as a part of possible extension towards duplex transmission. Each receiver has a separate channel for different applications, with the 4th channel added to make the node extensible for adding 4th type of traffic. In the heart of the node lies the traffic sink model described below.

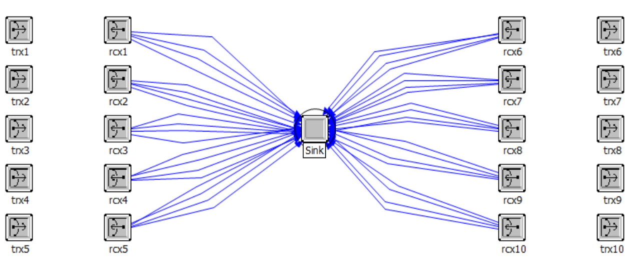

Figure 8 BBU/BBU Pool Node

A Point to Point (P2P) link of a 100 Gbps throughput is chosen to connect Base Stations to BBUs and not limit the aggregated throughput.

4.1 Traffic sources

All three traffic sources are modeled using the same design: a child process generates packets according to the recommendations outlined in Table 2 and a parent process controls the number of child processes running, in order to comply with traffic generation throughout the day according to load presented in Figure 4.

4.1.1. Parent process

The parent process is designed in the same way for all application types and it is presented in Figure 9.

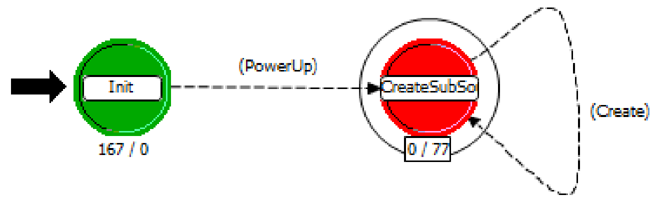

Figure 9 Parent process for traffic generators

In the **Init** state variables, statistics and a shared memory gathering statistics from the child processes have been initialized. Data on the number of child processes to be generated for specific BS type and application type is accessed from an assigned file, based on BS configuration on the scenario level.

In **CreateSubSources** state the number of running child processes is adjusted at each specified time interval – here one hour. Statistics are updated.

4.1.2. Child processes

The design of child processes is the same for Web/Data and File sharing sources. It is shown in Figure 10.

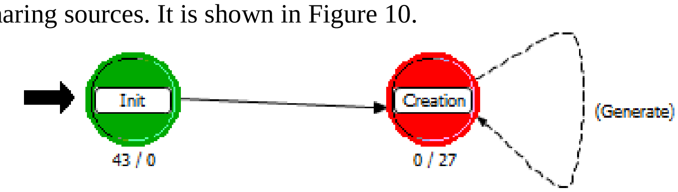

Figure 10 Web/Data and File Sharing Child process

In the **Init** state variables, statistics and a shared memory for sending statistics to the parent processes have been initialized. Inputs on traffic generators parameters are being read. In the **Creation** state packets of the given traffic characteristics have been created.

The child process for video data is designed as presented in Figure 11. On top of what is described for Web/Data and file sharing child processes, it takes care of preparing a frame that consists of 8 packets with specific sizes and inter-arrival times.

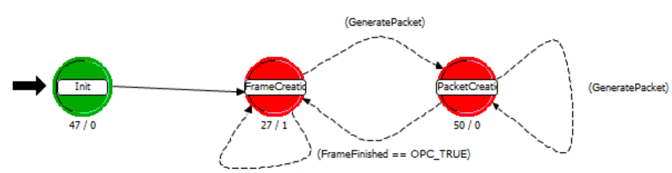

Figure 11 Mobile video child process

4.1.3. Traffic characteristics

Fig. 12 illustrates characteristics of FTP, Video and Web/Data traffic source. As presented in the first subfigure, For FTP traffic, packets size vary between 7 Mb and 35 Mb, packets are send in few minutes interval, as defined in Table 2. Video traffic illustrated in second and presented in zoom for 0.2 s on third subfigure constitutes of 8 frames with distributed sizes, released each 0.1 s. Web/data traffic, shown in fourth subfigure constitutes of packets of varying size generated every 1,1471 s.

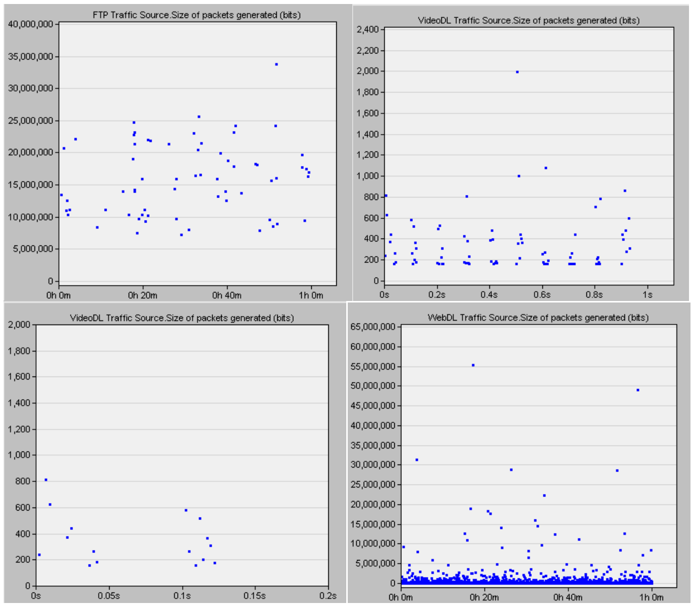

Figure 12 Size of FTP files (bits), characteristics of Video frames and size of Web/Data files.

4.2 Traffic sink

The design of the sink process that lies in the heart of BBU/BBU Poll node is presented in Figure 13. In the **Init** state variables and statistics have been initialized. The **Discard** state destroys the packets and writes the statistics each time a packet arrives (Arrival interrupt) or updates a crucial statistic on throughput each second (Second interrupt).

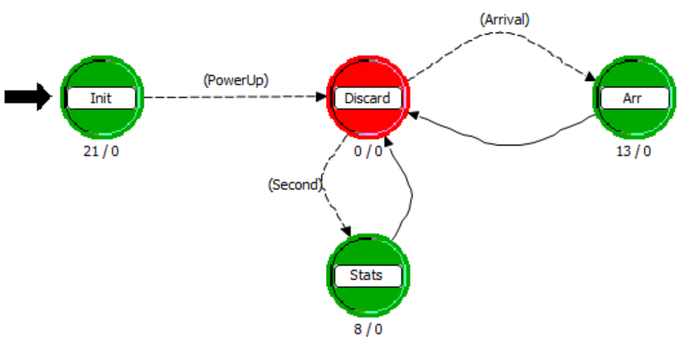

Figure 13 Process model for traffic sink

Table 3 Simulation parameters

Simulation Parameters	Values
Simulation for year	2017
City	Cologne, Western Europe
Application data	As in Table 1
Daily traffic per BS	448 Gb, uniformly distributed per BS
User data/Control data	User data
UL/DL data volume	40 %/60 %
Length of Simulation Run	24 hours
Number of runs	24
Time interval for updating data flows	1 hour
Link Speed	100 Gbps
% of total traffic coming from different BS types:	Office 30 %
	Residential 50 %
	Commercial 20%
Traffic models	As in Table 2
Traffic processing by the mobile network	Based only on user requests

5. Simulation Results

Simulations are carried on for 24 hours to study the impact on varying traffic load throughout the day. 24 runs with different seeds values were conducted to obtain results as less as possible influenced by a random number generator of OPNET Modeler. The traffic observed at Commercial, Residential and Office Base Station has characteristics as shown in Figure 14, for the exemplary representative of each type. The picture is not clear due to traffic peaks. The sample sum of throughput, presented in Figure 15 shows the trend, where the office Base Station experience highest accumulated throughput in office hours – 10 to 18, while residential ones in the evening. It follows input trend, as in Figure 4. An interesting aspect is represented by the traffic peaks, due to the way the specific applications (especially file sharing) are modeled.

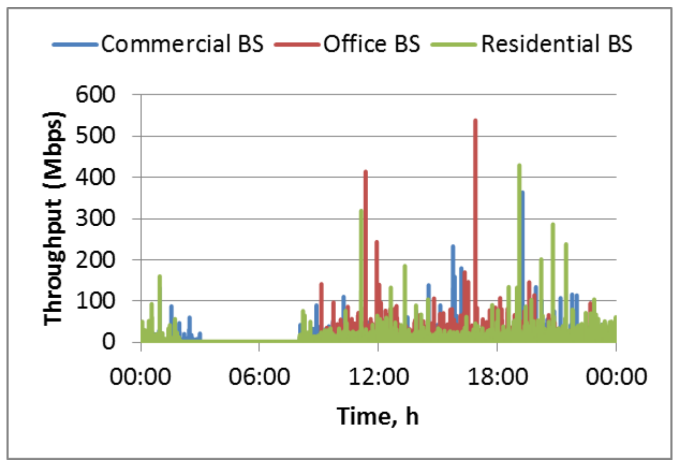

Figure 14 Daily throughput for different types of BSs

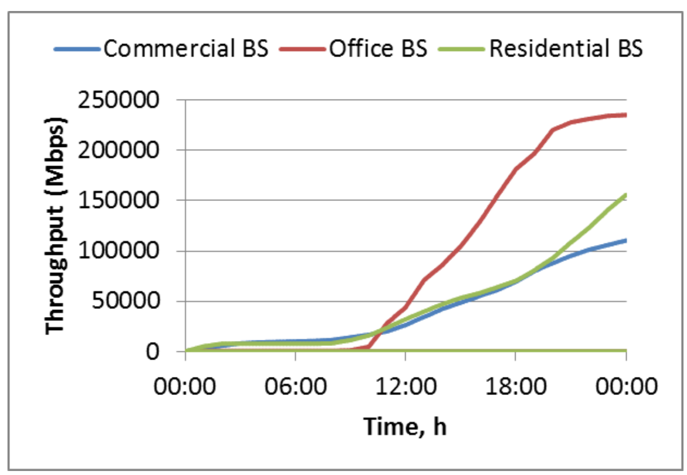

Figure 15 Sample sum of daily throughput for different types of BSs

For the C-RAN scenario, the aggregated throughput observed in the BBU Pool is presented in Figure 16. The general trend is as expected, a very low activity in the night hours, and quite stable throughput starting from 10 o'clock, as users are attached to either Office or Residential BS. The particular peaks depend on the variance of packet sizes. The overall peak is observed at noon and in the evening, which follows other daily load distributions, like the one presented in [7]. The latter source was not used as an input to the OPNET model; therefore it represents an external proof for validating simulation results.

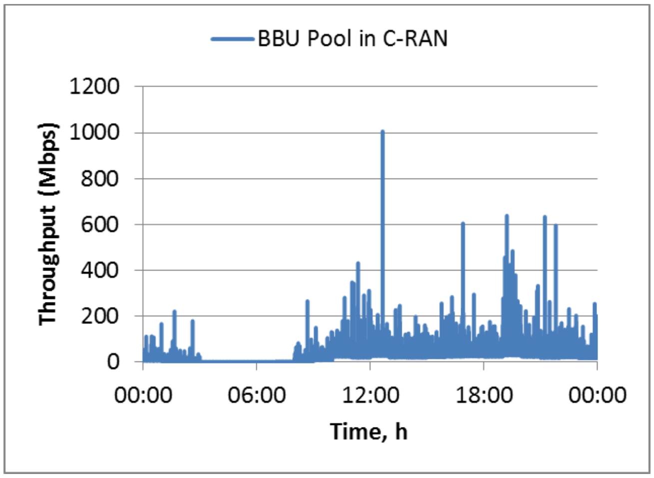

Figure 16 Aggregated traffic in BBU Pool in C-RAN for different types of BSs

The most relevant results of the performed simulations are the statistics on maximum throughput for each Base Station in the D-RAN architecture versus maximum throughput for BBU Pool in the C-RAN case. Table 4 summarizes the results for Seed = 128.

Table 4 Evaluation of Statistical Multiplexing Gain in C-RAN

D-RAN			C-RAN	
BS Type	BS	Peak throughput, Mbps	BS Type	Peak throughput, Mbps
Office	1	538.56	BBU Pool	**1005.46**
Office	2	253.96		
Office	3	350.25		
Commercial	4	363.69		
Commercial	5	267.57		
Residential	6	429.64		
Residential	7	157.74		
Residential	8	977.72		
Residential	9	608.97		
Residential	10	604.36		
Sum		4552.44		**1005.46**
Ratio		**4.53**	**Mean = 4.34, σ = 1.42**	

Analyzing the results for 24 different seeds, the mean value is 4.34 and standard deviation σ is 1.42, therefore, statistically for the given simulations the result falls into the interval 1.50-7.17 with 95 % of probability.

To sum up, the peak of aggregated user data traffic in BBU Pool is 4 times lower than the sum of peak user data throughputs in BBUs assigned to each of the cells. However, signaling needed to carry on the connections will stay the same in both cases, as the number of sessions won't change. Therefore if the operator is dimensioning network either for peak or average traffic, relatively it is a reasonable assumption to set up 4 times less BBU for user data processing in pool compared to the distributed set up. This result in lower CAPEX cost for network deployment and lower OPEX due to lower energy consumption. If the operator wishes to aggregate existing BBU into the pool it is possible to operate them on lower processing power (in terms of processor speed in MHz), saving on energy consumption and, at the same time, being ready for significant traffic growth.

The reduction of BBU resources by 75% is quite high. The expected result should be above 50%, assuming that people at least move from work to home using base stations in two places for different time of the day. The particular value depends on the daily traffic load distribution and application definition as well as traffic scheduling to be processed by the BBU. The simulation results are very sensitive to the traffic peaks, as seen from the variance of the result, which come from particular traffic modeling, having significant variance of packet sizes. In the model we assume that the user gets the maximum number of resources he requested in each session and that the signal processing doesn't introduce any delay and is handled in real time basis, which may not be the case in a real, congested network. Any mobile standard will introduce traffic processing of protocols of different layers, which will shape the traffic. Moreover, the resources needed for signaling traffic need to be taken in the account. Therefore the real value might be lower.

6. Conclusion

This paper compares the amount of resources needed for BaseBand user data processing in mobile C-RAN versus D-RAN network. It has been calculated that C-RAN enables reduction of user data signal processing resources 4 times. This has an impact on CAPEX and OPEX reduction as well as power savings. The results are based on traffic forecasts, data on daily load distribution of base stations located in office, commercial and residential areas and nature of different application. Traffic forecast for 2017 is used in order to provide useful conclusions for future networks in which user's needs grow significantly compared to nowadays.

In this paper it has been shown that looking at user data processing, C-RAN addresses the problem of growing network maintenance cost associated with growing traffic consumption when ARPU is flat. However, additional evaluation is required for signaling processing resources needed in BBU. As in C-RAN the overhead of L1 data to application data is very high, the cost of renting the fibers is needed to be taken in the account.

The simulation results are of great interest not only for mobile network operators, but also for equipment vendors who would like to assess the benefits of C-RAN.

7. Future work

OPNET Modeler provided a framework to create a network model with precise application definition and flexible traffic assignment. We believe that it will be relevant to study how sensitive the results are to changes in the traffic distributions and share of commercial, office and residential traffic.

Basing on the current model, we can observe how the link rate needed to support user demands is changing over the day. It can be calculated how to adjust bit rate on CPRI links over the time to lower energy consumption on the links.

The developed model can be further used to model a scenario with duplex transmission, with more base stations, each having a dedicated traffic profile, with real time updates on traffic load. M2M application could be also added, and the parameters for FTP, Web/Data and Video applications could be made configurable. Extensions could be added to assign different ratio of UL and DL traffic for different applications. Also the traffic going through BSs of different generation could be differentiated.

Furthermore, the model can be expanded to cover also:

1. Impact of signaling traffic on the load of BBU. In that case a more elaborated model of BBU that is subdivided into signaling (including scheduling) and simple packet forwarding will be needed.
2. Power consumption - this could be used to estimate the AC requirements (and potential savings) using C-RAN.
3. LTE (or other mobile standard) protocol stack processing.

Acknowledgement

Authors would like to thank MTI Radiocomp for kindly providing inputs on C-RAN solution and company view on deployment scenarios.

This work was partially sponsored by the 7th Framework Programme for Research of the European Commission HARP project, under grant number HARP-318489.

References

[1] "Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017," tech. rep., Cisco, February 2013.

[2] 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, January-February 2010.

[3] "C-RAN. The Road Towards Green RAN," tech. rep., China Mobile Research Institute, October 2011.

[4] R. Irmer et. al. "NGMN Radio Access Performance Evaluation Methodology", NGMN Alliance, January 31, 2008

[5] M. Molina, P. Castelli, and G. Foddis, CSELT "Web Traffic Modeling Exploiting TCP Connections'. Temporal Clustering through HTML-REDUCE", IEEE Network May/June 2000.

[6] http://www.fst-gsm.de, cited June 2013

[7] Vodafone, found in: A. Aijaz et.al, "Energy Savings for Mobile Communication Networks through Dynamic Spectrum and Traffic Load Management ", 2013

[8] "Common Public Radio Interface (CPRI); Interface Specification V5.0," September 2011.