Dynamic Resource Prediction and Allocation in C-RAN with Edge Artificial Intelligence

Wei-Che Chien, Chin-Feng Lai, *Senior Member, IEEE*, Han-Chieh Chao *, *Senior Member, IEEE*

***Abstract*— Artificial Intelligence (AI) is one of the important technologies for industrial applications but it needs a lot of computing resources and sensing data to support. Therefore, big data transmission is a challenge for current network architectures. In order to have high-performance computing requirements, this study proposes an emerging network architecture that combines edge computing and cloud computing to reduce the transmission of useless data and solve bottleneck problems. Moreover, we define the resource allocation problem about multiple Remote Radio Heads (RRHs) and multiple Base Band Units pools (BBU pools) in the Cloud Radio Access Network (C-RAN) for 5G. The Long Short-Term Memory (LSTM) is used to predict dynamic throughput and GA-based Resource Allocation Algorithm (GARAA) is used to optimize resource allocation. The simulation results represented that the proposed mechanism can achieve high resource utilization and reduce power consumption.**

***Index Terms*—Internet of things, Edge computing, Long-short term memory, Metaheuristic algorithms, C-RAN**

I. INTRODUCTION

With the rapid development of mobile phones, Internet of Things (IoT) [1], and mobile terminal devices in recent years, network resources are gradually insufficient and the burden on existing cellular networks is increasing. In addition, a large number of Artificial Intelligence (AI) applications will cause more throughput and computing resources. To maintain the Quality of Service (QoS), telecom operators continue to expand base stations to meet the requirement of users. However, since the UEs are mobile, they don't require a lot of network resources all day. For example, in the technology park or the tourist attraction, when a large number of tourists or employees enter here, the traffic of the network will be increased suddenly. Constantly, tourists or employees return home during the off-peak hours. Therefore, the demand for network traffic will be decreased. If the base station always provides the same resource at all time, it will lead to a lot of unnecessary resource waste.

To improve the efficiency of network resources, the Cloud of Radio Access Network (C-RAN) architecture is considered in the 5th generation mobile networks (5G) standard by 3rd Generation Partnership Project (3GPP) [2]. It splits the traditional base station into two parts, including the remote radio head (RRH) and the baseband unit (BBU). One BBU can be connected to one or more RRHs for signal processing, and the integrated BBUs form a baseband unit pool (BBU pool). The BBU pool can easily manage and dynamically allocate computing resources, and can flexibly provide resource allocation according to the environment. C-RAN not only can improve resource utilization but also reduces power consumption and reduces interference. Therefore, the dynamic adjustment of C-RAN resource allocation can greatly reduce the operator's cost.

In order to maintain the QoS, there are two main solutions. The first is to use Extremely High Frequency (EHF) to enhance throughput. The second is to use Multi-input Multi-output (MIMO) [3] to use spatial multiplexing technology to improve the efficiency of the used bandwidth. These two solutions inevitably require the expansion of the base station. Therefore, future network architecture will have architectures of multiple RRHs and multiple BBU pools.

The base station is deploying continuously. When the network requirement in a certain place is increased, more RRHs are needed to improve the coverage. If the BBU pool cannot support network requirements, it is necessary to increase the capacity of the BBU pool or add more the BBU pool. Therefore, In the future, the network architecture will inevitably have a situation in which the multiple RRHs connect to multiple BBU pools, and the resource allocation problem will become more and more important for the QoS.

For resource allocation in C-RAN, the first step is to predict traffic. Accurately predicting traffic can reduce the number of migrations of C-RAN resources and reduces transmission costs and energy costs. Most of the methods for predicting traffic in the early stage use statistical analysis. However, statistical methods cannot comprehensively consider global factors. Although the neural network was proposed several years ago, it was ignored because of the problem of gradient disappearing in the BackPropagation (BP). In recent years, the nonlinear activation function has been used to solve the gradient disappearance problem as well as causes the AI to be re-proposed.

Although artificial intelligence has been applied to many fields such as natural language processing [4] and speech recognition [5], most of the machine learning methods require

This work was supported in part by the Ministry of Science and Technology of Taiwan, R.O.C., under Contracts MOST 107-2221-E259-005-MY3.

W.-C. Chien and C.-F. Lai are with the Department of Engineering Science, National Cheng Kung University, 70101 Tainan, Taiwan (e-mail: cinfon@ieee.com; b9944006@gmail.com).

H.-C. Chao is with Department of Electrical Engineering, National Dong Hwa University, 974 Hualien, Taiwan, and Department of Computer Science and Information Engineering, National Ilan University, 260 Yilan, Taiwan. (e-mail: hcc@niu.edu.tw).

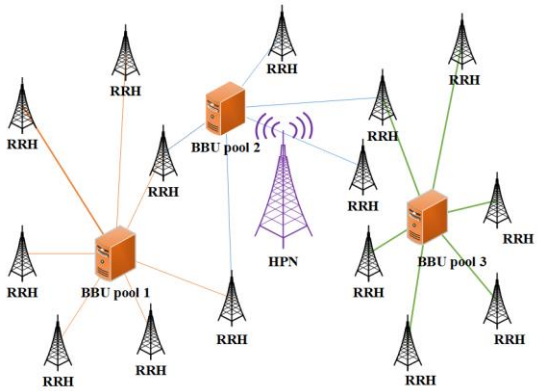

Fig. 1. The C-RAN architecture of multiple RRHs and multiple BBU pools

a certain amount of data and training time to achieve accurate models. Therefore, there are many limitations to the application. The combination of edge computing and AI can improve the weakness of AI. Its applications can be simply divided into two parts. The first part is the improvement of traffic, which uses the simple operation to filter out the useless data in advance so as to reduce the transmission of backhaul. The second part is the improvement of computing resources. It can aggregate terminal devices for distributed computing so that the computing performance can be improved. However, there are some challenges in edge computing in AI applications.

Although the computational efficiency can be improved by edge computing, when data size is large, it still needs to spend a lot of training time. Therefore, it is hard to apply to environments with low latency requirements, such as internet of vehicles (IoV). Combining other mechanisms to achieve AI applications for 5G is a trend. Metaheuristic algorithms can find good solutions in a limited time. For network applications with real-time requirements, they are potential solutions such as Simulated Annealing (SA) [6], Tabu Search (TS) [7], Genetic Algorithm (GA)[8], Ant Colony Optimization (ACO)[9], Particle Swarm Optimization (PSO)[10], etc. In this work, a novel network architecture for energy minimization is proposed for large-scale C-RAN. The contributions in this article can be summarized as follows.

- According to current network trends, we proposed a novel architecture about C-RAN with edge AI and defined the problem for resource allocation in multiple RRHs to multiple BBU pools.
- Considering low transmission latency and low power consumption for the real-time problem, we using Long Short-Term Memory (LSTM) to predict network traffic and using GA to allocated BBU resources.
- Based on the dynamic network environment, a mechanism combining edge AI is proposed, which can configure BBU resources at any time in response to environmental changes.

The remainder of this article is organized as follows. Section II introduces the related works of resource allocation in C-RAN and traffic prediction. The network model and problem definition are shown in section III. The proposed mechanism for resource allocation and traffic prediction are presented in Section IV. The simulation result and discussion are represented in Section V and Section VI. Finally, we conclude this work in Section VII.

II. RELATED WORK

C-RAN is regarded as one of the important architecture for the mobile network. In this section, we will introduce the current development of resource allocation and traffic prediction for C-RAN.

A. Resource allocation in C-RAN

C-RAN is an emerging and developmental network architecture in mobile networks. Many researches have explored the issue of energy loss in C-RAN and the number of BBU minimizations. In [11], author proposes that a BBU can link multiple RRHs to reduce network loss. They formulate the resource allocation problem as a bin packing problem. In the case of limited resources, a heuristic algorithm is proposed to link the BBU to the RRH so as to minimize the number of used BBUs and improve energy efficiency. [12] considered different hardware capabilities, processor computing power, memory size, and disk access capabilities in the BBU. The problem is defined by a multi-dimensional bin packing problem with the goal of minimizing the number of BBU launches within a limited hardware resource. [13] proposed that the tasks in the RRH can be calculated and processed by different BBUs to reduce the impact of BBU failures on the overall system. However, it needs to consider the additional cost of different BBU calculations. They propose HeuGA (heuristic genetic) which mixed with First-Fit and genes to reduce the number of BBU. [14] formulated the BBU-RRH allocation problem as a linear integer-based programming problem. The proposed Particle Swarm Optimization (PSO) based algorithm to maximize the network QoS and load fairness. Moreover, the concept can be extended to develop Network Function Virtualization (NFV) [15] and Software Defined Network (SDN) [16] solutions for C-RAN. However, the traditional C-RAN architecture is unable to support the huge network requirements. The large-scale resource allocation problem about multiple RRHs and multiple BBU pools needs to be considered.

B. Traffic prediction

In order to meet the rapid growth of network traffic, 5G networks aimed to optimize transmission efficiency and provide a higher QoS. Small cell is referred to a promising solution. However, the off-peak period is easy to cause the waste of many unnecessary resources when small cell is extensive deployment. Therefore, in the C-RAN structure, how to make effective spectrum prediction is very important for resource allocation.

[17] used support vector machine and regression model to estimated and predicted traffic. The proposed approach depends on a multistep sleep mode. It can control the Base Station (BS) on/off operation (sleep mode) so as to mitigate the excessive power consumption. In [18], Poisson distribution is used to predict network traffic, authors used the prediction data to design cache mechanism. However, these methods cannot fully consider environmental factors and have certain challenges in application. [19] proposed a video multicast orchestration scheme based on SDN to reduce the bandwidth consumption of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2913169, IEEE Transactions on Industrial Informatics

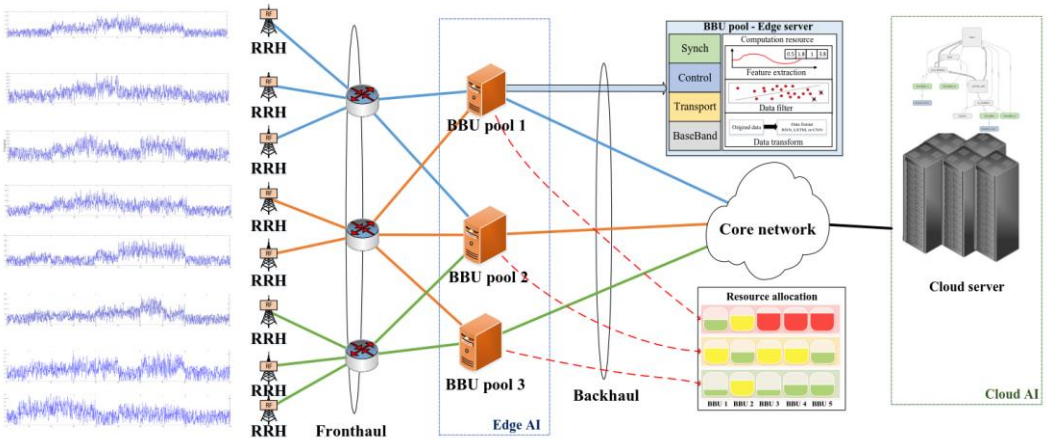

Fig. 2. System architecture with edge AI in 5G

a 5G network. Moreover, they proposed a Markov-based prediction method to predict network traffic load. [20] proposed a prediction algorithm based on Hidden Finite State Markov Chains (HFSMC) to guarantees service continuity in QoS networks. They proposed two integrated schemes. The first one is to predict mobile host movements, the second one is to minimize the wasted bandwidth for passive reservations. However, the Hidden Markov Model (HMM) is memoryless and cannot use contextual information. Since it is only related to its previous state, it is not used for medium and long-term predictions.

III. SYSTEM MODEL

A. Network model

According to the current network development, the network architecture for C-RAN will form as multiple RRHs to multiple BBU pools [25][26]. Each BBU pool is composed by a lot of BBUs. We define that $r_n\ (t)$is a Boolean value that represents the used statues of n^{th} RRH within period time *t*. The $bp_i(t)$is a Boolean value that represents whether the i^{th} BBU pool is used or not within period time *t*. The *m* is the total number of BBU pools. The $b_j^i(t)$ is a Boolean value that means the used status of j^{th} BBU of i^{th} BBU pool within period time *t*. The *k* is the total number of BBUs. Since the distance limitation of fronthaul, not all RRHs can connect to all BBU pools. Therefore, we define a Boolean value $T_{n,i,j}$. $T_{n,i,j}$=1 means that the n^{th} RRH can connect to the j^{th} BBU of the i^{th} BBU pool and vice versa. The task is the basic transmission unit. The *h* is the total number of tasks. Fig. 1 shows the C-RAN architectures of multiple RRHs and multiple BBU pools. In order to calculate the power consumption of the C-RAN. The $P\_bp_i\ (t)$ is the power consumption of the i^{th} BBU Pool within period time *t*. Please refer to (1)

$$P\_bp_i\ (t)= bp_i(t)\ \cdot\ (\sum_{j=1}^{k} P\_b_j^i(t)+P\_bp_{start}^i(t)), \quad (1)$$

When the BBU pool receives the task, BBU pool will allocate resources to the BBU for calculation. Where the task is the minimum unit of resource transmission. $P\_b_j^i(t)$ is power consumption for the j^{th} BBU of the i^{th}BBU Pool within period time *t*. $P\_bp_{start}^i(t)$ is the basic power consumption for starting the i^{th} BBU Pool within period time *t*.

$$P\_b_j^i\ (t)=b_j^i(t)\ \cdot\ (\gamma\ \cdot\ \sum_{z=1}^{h} G_{i,j,n}^z(t)\ \cdot S_z(t)+ P_{b}{}_{start}^{\ j}(t)), \quad (2)$$

Where γ is the weight value which is positively correlated with the size of the task. $G_{i,j,n}^z(t)$ is a Boolean value, $G_{i,j,n}^z(t)\ =1$ means that the z^{th} task is allocated from n^{th} RRH to j^{th} BBU of i^{th} BBU pool within period time *t*. $S_z(t)$ is number of z^{th} task. $P\_b_{start}^j(t)$ is the basic power consumption for starting the j^{th} BBU within period time *t*.

B. Problem definition

In the environment of multiple RRHs and multiple BBU pools, resource allocation becomes quite complicated. This problem can be divided into two parts. The first part is how to allocate RRH resources to the BBU pools. Since allocated order of RRH is a continuous problem, each solution will affect the next allocation result. How to effectively map RRH to the appropriate BBU pools to reduce power consumption is a key issue. The second part is the resource allocation of the BBU pool. This problem can be formulated into bin bagging problem. Since each BBU pool is composed of a lot of BBUs, how many BBUs are needed will affect the total power consumption. In addition, the above two problems will affect each other and need to be considered at the same time. In order to clearly illustrate the objectives of this study, we have refined this problem into a linear programming model:

- Objective function

$$\text{Min } \tau(t) = \alpha(t) +\beta(t)\ \ , \quad (3)$$

- Cost functions

$$\alpha(t) =\sum_{i=1}^{m} P\_bp_i(t)\ \ , \quad (4)$$

$$\beta(t) = \delta \cdot \sum_{z=1}^{t} G'^{z}_{i,j,n}(t) \cdot S'_z(t), \quad (5)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2913169, IEEE Transactions on Industrial Informatics

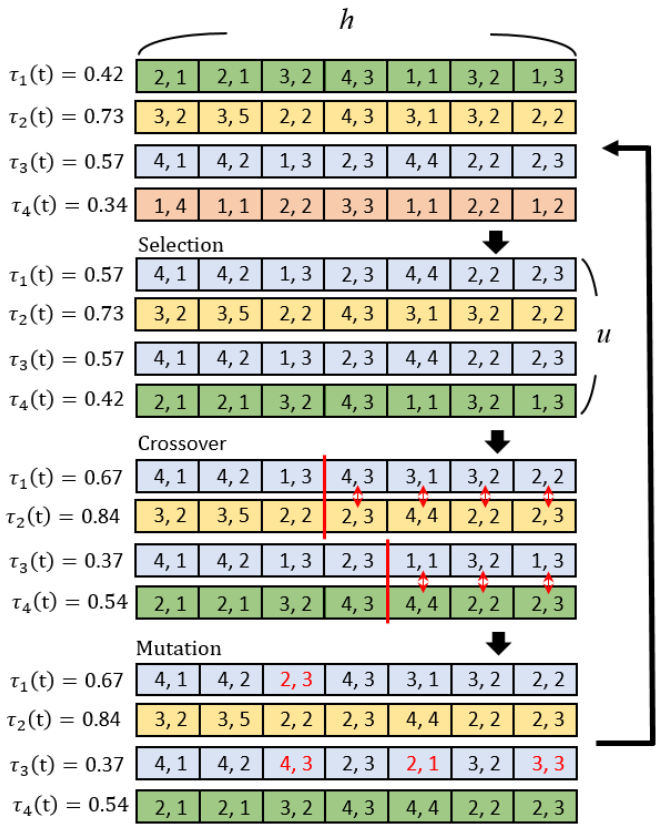

Fig.3. The schematic diagram of selection, crossover, and mutation in GARAA

- Subject to

$$P\_bp_i(t) > \sum_{j=1}^{k} P\_b_j^i(t) \quad , \tag{6}$$

$$P\_bp_{start}^i(t) > P\_b_{start}^j(t) \quad , \tag{7}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{k} T_{n,i,j} > 1 \quad , \tag{8}$$

$$\gamma, \delta > 0 \quad , \tag{9}$$

$$\sum_{j=1}^{k} BP\_tp_j > \sum_{z=1}^{h} tp_z \quad . \tag{10}$$

Where $\tau(t)$ is the total power consumption of C-RAN within period time *t*, including sum of the power consumption of the BBU pools $\alpha(t)$ and sum of the migrated power consumption $\beta(t)$. According to the application type of the terminal device, it can be easily divided into two types. The first one is a long-term and stable data transmission, such as IoTs. The second one is a mobile terminal device, such as smartphones, IoVs. When we want to shut down the less used BBU pool during the off-peak period, the data will be migrated. Therefore, it will cause additional power consumption. The β is the power usage of resource migration, and δ is the weight of power consumption when task migrates. $G'^{z}_{i,j,n}(t)$ represents the migrated status of z^{th} task. $G'^{z}_{i,j,n}(t) = 1$ means that the z^{th} task is migrated from n^{th} RRH to j^{th} BBU of i^{th} BBU pool. $S'_z(t)$ means the number of migrated task. The main goal of this research is to minimize total power consumption of the BBU pools $\tau(t)$.

We define some restrictions based on environmental constraints as follows. The (6) indicates that the power consumption of the i^{th} BBU pool is larger than the power consumption of total BBUs of i^{th} BBU pool. Moreover, the basic power consumption for starting the BBU Pool is also larger than the basic power consumption for starting the BBU that is shown in (7). The (8) indicates that the number of BBU pools and BBUs that can be mapped by n^{th} RRH is at least 1. The (9) represents γ and δ belongs to a positive number. BP_tp_j is the throughput of j^{th} BBU pool. tp_z is the throughput of z^{th} task. The (10) restrict that the total throughput of the task must be less than the total throughput that the BBU Pools can load.

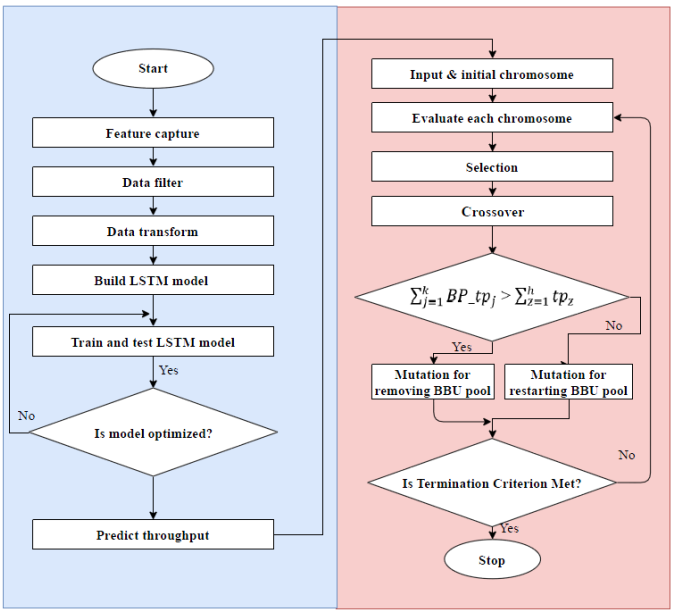

Fig. 4. The flow chat of LSTM+GARAA

C. Edge AI

For resource allocation, the first step is to accurately predict the traffic of each RRH. Therefore, in this study, LSTM is used to predict the traffic of each BBU pool. Since deep learning requires a lot of training time, in order to meet the low latency requirements of 5G, this study proposes an edge AI architecture for C-RAN. Fig. 2 is system architecture with edge AI in 5G. Each BBU pool can support some computing resources, which can run some simple operations such as data filtering, data converting, etc. When the BBU pool receives the request, the connection between the BBU pool and the RRH is coordinated by the HPN. Noticeably, the BBU pool that each RRH can be connected to is limited by the restricted environment. Moreover, the low computation request operations can be performed by the edge of the BBU. Conversely, the complex computation can offload to cloud. Therefore, integrating edge computing and cloud computing can enhance AI applications that require a lot of computing time.

IV. PROPOSED MECHANISM

A. Flow chart for resource allocation

We use LSTM to predict network traffic and treat the predicted data as GARAA input. Then, using GARAA to allocate resource for BBU pool. We aim to minimize the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2913169, IEEE Transactions on Industrial Informatics

TABLE I
FEATURES SELECTION

Features	Features selected as basis of classification
Throughput	Throughput is regarded as the main feature, we perform resource allocation of the BBU pool by predicting the throughput of each RRH.
Number of UEs	The number of UEs will affect the amount of throughput. We simply divide the number of UEs into high throughput UEs (HTUE) and low throughput UEs (LTUE). Number of HTUEs may be few but they will cost a lot of multimedia audio and video transmissions, such as mobile phones, tablets, and laptops. Conversely, there may be a lot of LTUEs but the throughput is not high, such as NB-IoT, eMTC devices.
Number of APPs	Although the number of APPs is not necessarily proportional to throughput, the number of APPs will indirectly affect production. Therefore, the number of APPs can be considered a feature.
Number of device types	Although the number of types of devices may not be proportional to throughput because some HTUEs will affect the overall result, the number of device types has an indirect effect on throughput.

number of used BBU and migrations so as to reduce power consumption. First, the RRH will randomly connect to the BBU bool and send the requests. The BBU pool will pre-process the collected traffic through the edge server for a period of time, including feature extraction, data filtering, and data conversion. Then build the LSTM model and start learning. If the loss value meets the termination condition, the model after learning can predict the traffic. Otherwise, it should continue training. Through the prediction of throughput, we can let the BBU which has less throughput shut down or let the overloaded BBU pool offload the traffic to other BBU pools. Therefore, the resource allocation will be more accurate. Finally, the GARAA is used for resource allocation, and the steps of selection, crossover, and mutation are repeated until the termination condition is met. Through the above mechanism, we can dynamically allocate resources and minimize the power consumption of the BBU pool. The detailed flowchart is shown in Fig. 3.

B. LSTM for traffic prediction

We use the hour as the basic unit and make predictions every hour. For feature capture, the throughput, the number of UEs, the number of apps, the number of device types are used as feature values. The detailed reason for using features is shown in Table I. In order to avoid affecting the forecast results because some of the data is too large, we filter out the flow of the standard deviation greater than 2, please refer to (10). After filtering out the data, we convert the data into a data pattern that conforms to the Tensor flow. In this study, the model was established using LSTM, which is divided into three layers, including input layer, hidden layer and output layer. The loss function (loss function) using Huber; the optimizer uses Adam. The activation function uses Relu to reduce the gradient disappearance problem so that the predicted loss value after training is minimized and then the prediction ability is improved.

$$\bar{x} - 2\sigma < x < \bar{x} + 2\sigma, \;\; x > 0. \tag{10}$$

C. GARAA for resource allocation

In our scenario, each RRH can connect to multiple and different BBU pools but it is limited by the geographical environment. The BBU resource allocation of each BBU pool can be formulated as a bin packing problem. Therefore, it belongs to the NP-hard problem. In order to dynamically allocate resources. We proposed a GA- Based Resource Allocation Algorithm (GARAA) in C-RAN architecture. The concept of GARAA is based on the theory of evolution. There are three main steps, including selection, crossover, and mutation. Where "chromosome" represents the solutions. The "gene" represents which BBU of BBU pool can process this task. For example: (1,2) means that 2^{th} BBU of 1^{th} BBU pool can process this task. Firstly, GARAA will generate u chromosomes and each chromosome has h gene. All of the genes will be randomly generated. Then the following steps will be taken.

Selection: We can choose a better solution according to fitness function ($\tau(t)$). This is an important mechanism for convergence. The better solution will be selected, the worse will be eliminated. Our selection strategy is that two chromosomes will be chosen randomly within all chromosomes. Then we can compare two chromosomes and find the best one to save it. Finally, repeat above steps u times. The detailed schematic of the selection step is shown in Fig 4.

Crossover: In this step, chromosomes can exchange its gene with each other. Since the worse chromosomes have been removed in the selection step, the crossover has a chance to make the solution better. Our crossover step is that two chromosomes will be chosen randomly within all chromosomes. Then, we randomly swap a segment of two chromosomes. The detailed schematic of the crossover step is shown in Fig 4.

Mutation: In this step, it has the chance to change the status of a gene in each chromosome so that solution can avoid falling into local optimization quickly. If total throughput of the task is less than the total throughput that the BBU Pools can load, the mutation step has the opportunity to randomly remove BBU Pools to reduce power consumption. Otherwise, the mutation step has the opportunity to randomly restart BBU Pools to guarantee that the restrictions are met. The benefit of this strategy is that we can reduce the number of BBUs and avoid slow convergence. Therefore, the mutation step not only made the number of BBU pools to converge to meet the needs of the environment but also avoids falling into the best solution in the region. The detailed mutation step is shown in fig 4.

V. RESULT ANALYSIS

D. Environment setup

1) System environment

Our environment is a Linux operating system, using the Python programming language, Anaconda as the Python execution environment and suite management. The installed TensorFlow suite is a 1.0.1 GPU version. Since TensorFlow supports distributed computing architecture, it can be deployed on multiple computing devices. For computing resource allocation, the edge AI is used to pre-process data that not only

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2913169, IEEE Transactions on Industrial Informatics

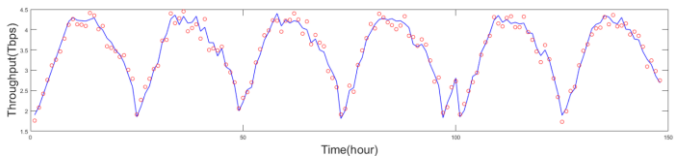

Fig. 5. Throughput for scenario (a)

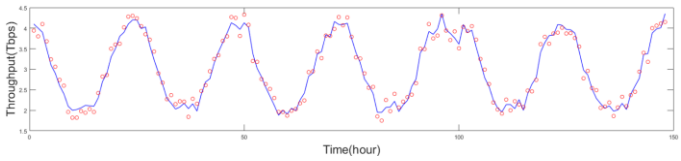

Fig. 6. Throughput for scenario (b)

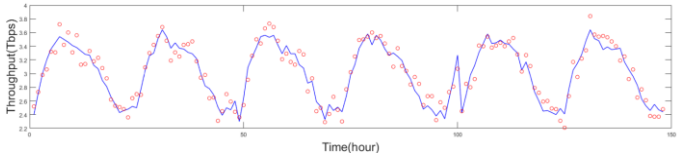

Fig. 7. Throughput for scenario (c)

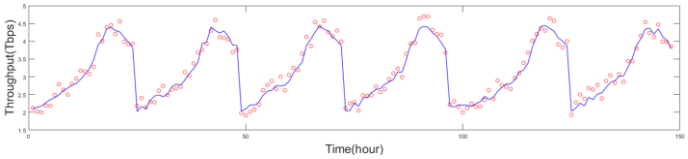

Fig. 8. Throughput for scenario (d)

can reduce the useless data transmission in the backhaul but also can filter out unnecessary data. The cloud server is used to process more complex computations such as neural networks and GARAA that makes deep learning to have better performance. If the data is large and the model complexity is high, multiple machines can be used to achieve the acceleration in the cloud servers, and the real-time requirement can be solved.

2) Data acquisition:

In our scenario, the number of UEs per RRH is between 1 to 1000, the number of HTUEs is between 10 to 20, the number of LTUEs is between 50 to 100, the number of APPs is between 1to 20, and there are four UE types, including mobile phones, tablets, IoT, laptops, car wifi. For each RRH, the number of UE, type of UE, and Apps are all different per hour. Therefore, the throughput of each RRH will continue to change. Each RRH can connect to BBU pools is different because of environmental constraints but it can be connected to at least one BBU pool.

E. Simulation result

Fig. 5 to Fig. 8 are throughput predictions by using LSTM. The x-axis is the time axis and the y-axis is throughput. Fig. 5 shows that the throughput is concentrated from 10:00 am to 3:00 pm, and the scenario is similar to the general working place. The throughput in Fig. 6 is concentrated from 1 am to 9 am and from 5 pm to 12 pm, and the scenario is similar to a residential area. The throughput in Fig. 7 is concentrated in the morning that the scenario is similar to the market. The throughput in Fig. 8 is mostly concentrated after 5 pm, and the scenario is similar to the night market. We use LSTM to predict the throughput and the accuracy can achieve 90%. In scenario of Fig. 7 and Fig. 8, the types of UEs, the traffic, the number of users, and the throughput are always changed so that the accuracy in prediction is not high. Conversely, the scenario in Fig. 5 and Fig. 6, due to there are more UEs that are stably transmitted, the accuracy in prediction is higher.

Fig. 9 shows the comparison of power consumption. The x-axis is the number of users, and the y-axis is the value of power consumption. The power consumption using GARAA is significantly higher than LSTM+GARAN. The reason is that after throughput prediction by LSTM+GARAA can allocate resources based on these data so as to reduce the power consumption of BBU pool. Although it is not 100% fully predictable, LSTM+GARAN can achieve good resource allocation and reduce power consumption of the BBU pool when most of the traffic can be predicted in advance. The detailed discussion about comparison of migrations, number of BBU pools, BBUs are as follows.

Fig. 10 shows the comparison of number of BBUs. As the number of UE increases, so does the number of BBU, and the problem of resource allocation will become more complex. Comparison of the number of BBU pool, the number of BBUs of LSTM+GARAA is slightly larger than the GARAA. The reason is that LSTM+GARAA can shut down unnecessary BBU pools by predicting throughput. Therefore, all the throughput will be concentrated in part of the BBU pools, causing the resource allocation of some BBUs to be uneven. Conversely, for GARAA, since there is no predicted data of throughput, resource allocation can only be performed according to the current traffic. Therefore, it needs to use a large number of BBU pools to maintain QoS and avoid BBU Pool service interruption due to overload. In other words, GARAA can allocate more resources at the same time, and the BBU can be effectively allocated.

The comparison of the number of BBU pools is shown in Fig. 11. GARAA can only allocate resources according to the current network throughput so that it is easy to cause excessive BBU pool waste during the off-peak period. The number of GARAA BBU pools will be increased while the number of UE is increasing. Conversely, LSTM+GARAA can avoid this problem because it can predict the throughput and pre-allocated resource so as to reduce the number of BBU pools.

Fig. 12 and Fig. 13. shows the comparison of number of migrations and its power consumption. The power consumption of LSTM+GARAA data migration is higher because the throughput of the BBU pool can be reduced during the off-peak period so as to reduce the total power consumption and keep the overall BBU pool in a high-load state. In other words, it can reduce excessive useless BBU pool usage

Fig. 14 shows the comparison of total power consumption. When the number of people is small, the overall power consumption is not obvious. Because the location of the UE is very average, the number of RRHs used is also relatively average. Therefore, in order to meet the service needs of all users, the BBU has a small combination of resource allocation. As the number of people increases, the combination of allocable resources increases so that the LSTM+GARAA method can effectively reduce power consumption. Although the power

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2913169, IEEE Transactions on Industrial Informatics

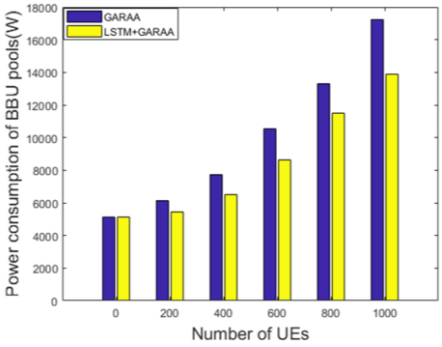

Fig. 9. The comparison of power consumption.

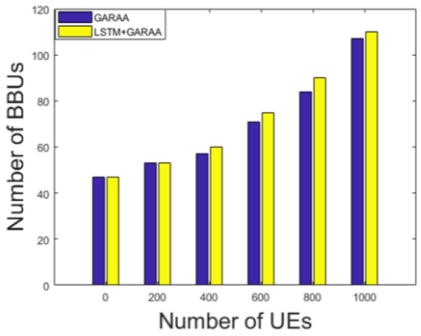

Fig. 10. The comparison of number of BBUs

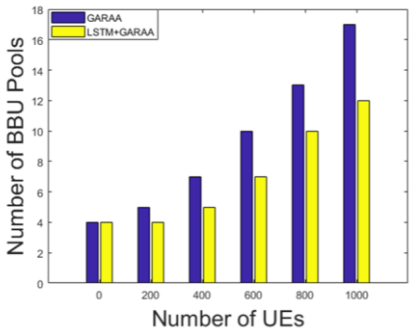

Fig. 11. The comparison of number of BBU pools

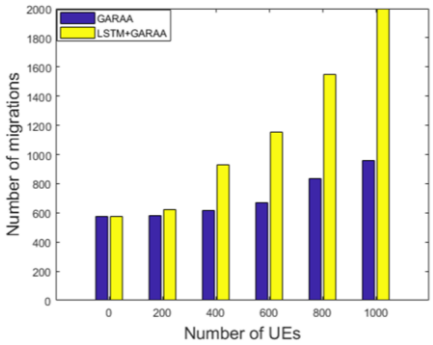

Fig. 12. The comparison of number of migrations

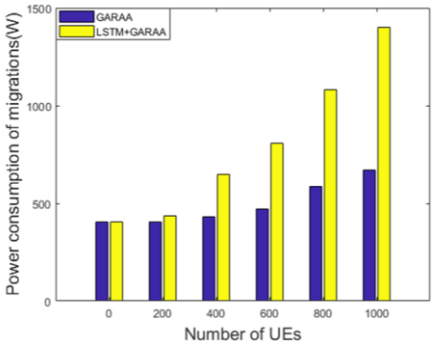

Fig. 13. The comparison of migrating power consumption

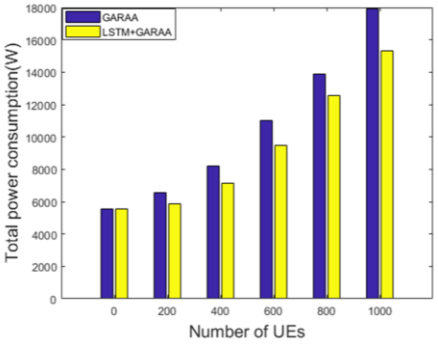

Fig. 14. The comparison of total power consumption

consumption of LSTM+GARAA data migration is high, the amount of use of the BBU pool can be greatly reduced as well as the total power cost can be reduced.

VI. Discussion

The simulation results show that LSTM + GARAA can reduce power consumption in the network architecture of multiple BBU pools and multiple RRHs. However, there are some limitations to this research. For concerts, Olympics, and so on, the throughput for these scenarios changes rapidly. When the types of UEs, the traffic, the number of users, and the throughput are always changed, it will affect the performance of power saving. The reason is that the proposed strategy can allocate the resource of BBU pools and BBUs according to prediction result. For scenarios with large changes in throughput, it is necessary to reserve enough buffer in advance to avoid overloading BBU pools, thus increasing unnecessary energy consumption. Moreover, we consider the energy consumption and migrated power consumption of BBU pools and BBUs. Although the migrated power consumption of LSTM+GARAA is higher than GARAA, the overall energy consumption is less than GARAA. The reason is that GARAA needs to spend a lot of BBU pool resources in order to meet network requests in different time periods. Conversely, LSTM+GARAA can concentrate resources on part of BBU pools to reduce the number of BBU pools so as to reduce energy consumption.

VII. CONCLUSION AND FUTURE STUDIES

AI has led to the rapid development of many industries but many applications are based on high-speed transmission and powerful computation resources. In the future, a large number of base stations will be built to meet the network requirements of AI but this also extends some challenges, including the waste of resources during the off-peak period, unnecessary and huge data led to transmission bottleneck. Therefore, according to the current development trend, this study proposes an emerging 5G network architecture which combing edge computing and cloud computing. It can reduce the computing cost and useless data transmission in backhaul through the good division of edge servers and cloud servers. In addition, in order to allocate resources efficiently in C-RAN, including multiple RRH and multiple BBU pool, we use LSTM to predict the throughput of each RRH. Then the GARAA is used to allocate resources to minimize the number of BBU pools and task migrations so as to minimize power consumption. The results show that this method can effectively allocate resources according to predictive data, reduce the number of BBU migrations and reduce power consumption. In the future works, we will focus on improving the accuracy of throughput predictions and reducing the computational time for deep learning.

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347-2376, June 2015

[2] H. Chungwoo, 3GPP RAN#72- RP-160750 2016. [online]. Available from: https://portal.3gpp.org/ngppapp/CreateTdoc.aspx?mode=view&contributionId=709524

[3] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186-195, February 2014.

[4] T. Young, D. Hazarika, S. Poria, and E. Cambria, " Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55-75, July 2018.

[5] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," *in proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5934-5938, April 2018.

[6] L. Wei, Z. Zhang, D. Zhang, and S. C. Leung, "A simulated annealing algorithm for the capacitated vehicle routing problem with two-dimensional loading constraints," *European Journal of Operational Research*, vol. 265, no. 3, pp 843-859, March 2018.

[7] G. L. A. Méndez, P. E. J. Gómez, and V. A. Terré, "Application of Tabu search based algorithms for symbol detection in L-MIMO systems," *in proceedings of 2017 IEEE Colombian Conference on Communications and Computing (COLCOM)*, pp. 1-6, Aug. 2017.

[8] G. Oliveri, G. Gottardi, F. Robol, A. Polo, L. Poli, M. Salucci, and R. Lombardi, "Codesign of Unconventional Array Architectures and Antenna Elements for 5G Base Stations," *IEEE Transactions on Antennas and Propagation*, Vol. 65, no. 12, pp. 6752-6767, Dec. 2017.

[9] B. Chen, J. Zhang, Q. Zhu, X. Wang, and M. Gao, "Energy-Efficient Traffic Grooming in 5G C-RAN Enabled Flexible Bandwidth Optical Networks," *in Asia Communications and Photonics Conference*, pp. M3B-2, Nov. 2017.

[10] T. Shahjabi, and K. V. Babu, "Cooperative scheme for wireless energy harvesting and spectrum sharing in cognitive 5G networks," *2017 International Conference on Innovations in Electrical, Electronics, Instrumentation and Media Technology (ICEEIMT)*, pp. 57-61, February 2017.

[11] K. Boulos, M. El Helou, and S. Lahoud, "RRH clustering in cloud radio access networks," *2015 International Conference on Applied Research in Computer Science and Engineering (ICAR)*, pp. 1-6, October 2015.

[12] N. Yu, Z. Song, H. Du, H. Huang, and X. Jia, "Multi-resource allocation in cloud radio access networks," 2017 IEEE International Conference on Communications (ICC), pp. 1-6, May 2017.

[13] F. Zhang, J. Zheng, Y. Zhang, and L. Chu, "An Efficient and Balanced BBU Computing Resource Allocation Algorithm for Cloud Radio Access Networks," *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1-5, June 2017.

[14] M. Khan, A. Sabir, and H. S. Al-Raweshidy, "Load balancing by dynamic BBU-RRH mapping in a self-optimised Cloud Radio Access Network," *In 2017 IEEE 24th International Conference on Telecommunications (ICT)*, pp. 1-5, May 2017.

[15] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network*, Vol. 28, no. 6, pp. 18-26, November 2014.

[16] S. Sezer, S. Scott-Hayward, P. K. Chouhan, B. Fraser, D. Lake, J. Finnegan, and N. Rao, "Are we ready for SDN? Implementation challenges for software-defined networks," *IEEE Communications Magazine*, Vol. 51, no. 7, pp. 36-43, July 2013.

[17] H. Pervaiz, O. Onireti, A. Mohamed, M. A. Imran, R. Tafazolli, and Q. Ni, "Energy-Efficient and Load-Proportional eNodeB for 5G User-Centric Networks: A Multilevel Sleep Strategy Mechanism," *IEEE Vehicular Technology Magazine*, Vol. 13, no. 4, pp.51-59, October 2018.

[18] W. C. Chien, H. H. Cho, H. C. Chao, and T. K. Shih, "Predictable NDN Cache over Mobile Communication," *2015 IEEE 39th Annual Computer Software and Applications Conference*, Vol. 1, pp. 23-28, July 2015.

[19] X. Huang, S. Tang, Q. Zheng, D. Zhang, and Q. Chen, "Dynamic Femtocell gNB On/Off Strategies and Seamless Dual Connectivity in 5G Heterogeneous Cellular Networks," *IEEE Access*, Vol. 6, pp. 21359-21368, January 2018.

[20] P. Fazio, M. Tropea, S. Marano, "A distributed hand-over management and pattern prediction algorithm for wireless networks With mobile hosts," *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, July 2013.

[21] W. Mo, C. Gutterman, Y. Li, G. Zussman, and D. Kilper, "Deep Neural Network Based Dynamic Resource Reallocation of BBU Pools in 5G C-RAN ROADM Networks," *Optical Fiber Communication Conference*, pp. Th1B-4, March 2018.

[22] A. Pelekanou, M. Anastasopoulos, A. Tzanakaki, and D. Simeonidou, "Provisioning of 5G services employing machine learning techniques," *2018 IEEE International Conference on Optical Network Design and Modeling (ONDM)*, pp. 200-205, May 2018.

[23] H. Pang, J. Liu, X. Fan, and L. Sun, "Toward Smart and Cooperative Edge Caching for 5G Networks: A Deep Learning Based Approach," *In Proc. of IEEE/ACM International Symposium on Quality of Service (IWQoS)*, 2018.

[24] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel state information prediction for 5g wireless communications: A deep learning approach," *IEEE Transactions on Network Science and Engineering, 2018*.

[25] L. Velasco, A. Castro, A. Asensio, M. Ruiz, G. Liu, C. Qin, R. Proietti, and S. J. B. Yoo" Meeting the Requirements to Deploy Cloud RAN Over Optical Networks"IEEE/OSA Journal of Optical Communications and Networking. Vol. 9, No. 3, 2017.

[26] B. R. Chen, C. Y. Huang, C. H. Chien, C. N. Lai, and C. W. Yuan, "Network infrastructure and software defined remote radio head controller." U.S. Patent Application No. 15/367,538.

Wei-Che Chien received the B.S. and M.S degree in Computer Science and Information Engineering from National I-Lan University, Taiwan in 2013 and 2016. He is currently pursuing the Ph.D. degree at the Engineering Science, National Cheng Kung University. His research interests include wireless rechargeable sensor networks, next generation network, IoT application, and cloud computing

Chin-Feng Lai (SM'14) is an associate professor at Department of Engineering Science, National Cheng Kung University since 2016. He received the Ph.D. degree in department of engineering science from the National Cheng Kung University, Taiwan, in 2008. He received Best Paper Award from IEEE 17th CCSE, 2014 International Conference on Cloud Computing, IEEE 10th EUC, IEEE 12th CIT. He has more than 100 paper publications. He is an associate editor-in-chief for Journal of Internet Technology. His research focuses on Internet of Things, Body Sensor Networks, E-healthcare, Mobile Cloud Computing, Cloud-Assisted Multimedia Network, Embedded Systems, etc. He is an IEEE Senior Member since 2014.

Han-Chieh Chao (SM'04) received his MS and Ph.D. degrees in Electrical Engineering from Purdue University in 1989 and 1993 respectively. He is currently a Professor with the Department of Electrical Engineering, National Dong Hwa University, where he also serves as the President. He is also with the Department of Computer Science and Information Engineering and the Department of Electronic Engineering, National Ilan University, Taiwan. He is a fellow of IET (IEE) and a Chartered Fellow of the British Computer Society. He serves as the Editor-in-Chief for the Institution of Engineering and Technology Networks, the Journal of Internet Technology, the International Journal of Internet Protocol Technology, and the International Journal of Ad Hoc and Ubiquitous Computing.