SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •
Special Focus on Mobile Network Virtualization

April 2017, Vol. 60 040304:1–040304:11
doi: 10.1007/s11432-016-9038-7

Storage and computing resource enabled joint virtual resource allocation with QoS guarantee in mobile networks

Xiaodong XU*, Jiaxiang LIU, Wenwan CHEN, Yanzhao HOU & Xiaofeng TAO

National Engineering Laboratory for Mobile Network Technologies,
Beijing University of Posts and Telecommunications, Beijing 100876*, China*

Received November 8, 2016; accepted February 15, 2017; published online March 17, 2017

Abstract Virtualization is the trend for the future mobile networks. With the advantage of virtualization, we can abstract the physical mobile network into the virtual network function (VNF) and design the network without the details. In this paper, we focus on the virtualization of the physical resources so that the resource allocation scheme considers not only the time-varying characteristic of wireless channels but also the amount of storage and computing resources. Virtual resources are composed of radio, storage and computing resources based on the virtualization technology. Since the cloud radio access network (C-RAN) is a successful paradigm to introduce computing resources into mobile networks, we investigate the virtual resource allocation scheme in the C-RAN architecture. With the content caching technology, we introduce the storage resources into joint resource allocation scheme further. In order to evaluate the performance of proposed scheme, we choose the effective capacity as the metric to include the influence of service latency. The purpose of the optimization problem is maximizing the system effective capacity with constraints of radio, storage and computing resources. It is simplified and converted into a convex problem solved by the subgradient method. Simulation results are provided to demonstrate performance gain of the effective capacity based joint resource allocation scheme.

Keywords virtualization, resource allocation, cache, computing resource, effective capacity, C-RAN

Citation Xu X D, Liu J X, Chen W W, et al. Storage and computing resource enabled joint virtual resource allocation with QoS guarantee in mobile networks. Sci China Inf Sci, 2017, 60(4): 040304, doi: 10.1007/s11432-016-9038-7

1 Introduction

Capacity and ultra-reliable communication are the requirements for the fifth generation (5G) wireless communication system [1]. Due to the mobile data traffic explosion with the rapid emergence of various mobile services, it is obvious that the exponential growth of mobile data will continue in the foreseeable future [2]. To meet the requirements above, 5G wireless communication system should achieve system capacity growth by a factor of at least 1000 [3]. Many novel technologies, such as CoMP [4], content caching and Mobile Edge Computing (MEC), have been proposed to achieve this goal. Network Function Virtualization is a promising technology to combine them into the same architecture by incorporating

* Corresponding author (email: xuxiaodong@bupt.edu.cn)

radio, storage and computing resources together, which extends the concept of resources of radio access network (RAN).

Virtualization derives from the wired networks and develops for decades. The applications of this technology are virtual private networks (VPNs) over wide area networks (WANs) and virtual local area networks (VLAN) in enterprise networks, which aims at overcoming the resistance of the current Internet to fundamental architecture changes [5]. Because of the enormous data traffic demand in mobile networks, it is promising and necessary to introduce virtualization into wireless networks. With the help of virtualization, wireless networks can decouple the services from the infrastructure so that different services can share the same infrastructure, reducing the capital expenses and operation expenses of mobile network operators. Above all, virtualization can abstract and isolated different kinds of physical resources into virtual resource blocks so that radio, storage and computing resources can work together to maximize the utility of system resources. Virtual resource blocks are divided into various quality of service (QoS) supporting abilities, which can correspond the specific service requirement effectively. In other words, the mobile network is abstracted into the virtual network functions to support the network slices and are shared by multiple users through isolating from each other.

Cloud radio access network (C-RAN) is being recognized as a possible scheme of future mobile network architecture. Its main idea is to decompose tradition base stations (BSs) into remote radio heads (RRHs) and baseband unit (BBU). RRHs are distributed geographically and responsible for the transmitting and receiving signals for users while BBU is centralized into a BBU pool which processes the joint decompressing and decoding schemes effectively. The connection between RRHs and BBU is called the fronthaul link so that they can exchange information conveniently. The capacity limits of fronthaul links increase the service delay of users and degrade the performances of C-RAN. To overcome the disadvantages of C-RAN with long delay experienced, heterogeneous C-RAN is proposed to meet this challenge in [6]. The control plane is transferred from BBU pool to high power node (HPN) to alleviate the burden of fronthaul links. HPN can provide the seamless coverage and broadcast the control signals to user equipment without fronthaul links and RRHs can concentrate into high-speed data rate transmission. To make the cloud closer to user and decrease the delay of services further, Ref. [7] proposes a distributed mobile cloud computing architecture where radio and computing resources are orchestrated jointly under the constraints of QoS requirements of services. It also puts forward the scheme of computing offloading to minimize the overall energy consumption at the mobile slides. The advantage of MEC is revealed in this paper which explores the benefit of scheduling radio and computing resources jointly.

Along with the cloud edge computing, it is a natural idea to explore the potential of storage resources in the clouds. Content caching is proposed to bring contents, which can be stored in clouds, much closer to users shrinking the latency of service caused by backhaul links. According to the research of [8], most parts of data traffic are caused by a small number of popular contents (such as popular videos and music). If we have cached the popular contents in advance, numerous duplicate downloads from servers in the mobile networks would be reduced and the service respond rate would be improved significantly.

Therefore, motivated by the capacity of C-RAN in storage and computing, the virtualization of radio, storage and computing resources can take the advantages of content caching and MEC. The main contributions of this paper can be summarized as follows.

1) We packet the radio, storage and computing resources into virtual resources and investigate the joint resource allocation scheme. Compared with the existing work, such as [7, 9], which only introduce storage or computing resources into RAN separately, we take both of them into the resource allocation based on the virtualization technology.

2) With the help of the virtual resources, the user association and scheduling are optimized cooperated with the storage and computing resources. Users prefer to associate with the cloud which has cached the required content already in order to reduce the latency. The user scheduling is not only based on the wireless channel conditions but also the baseband processing ability of computing resources in clouds.

3) Since the effective capacity theory can include the backhaul latency into user data rate, the joint resource allocation problem is formulated as the maximization of overall effective capacity under the constraints of radio, storage and computing resources. With this criterion, the performance gains of

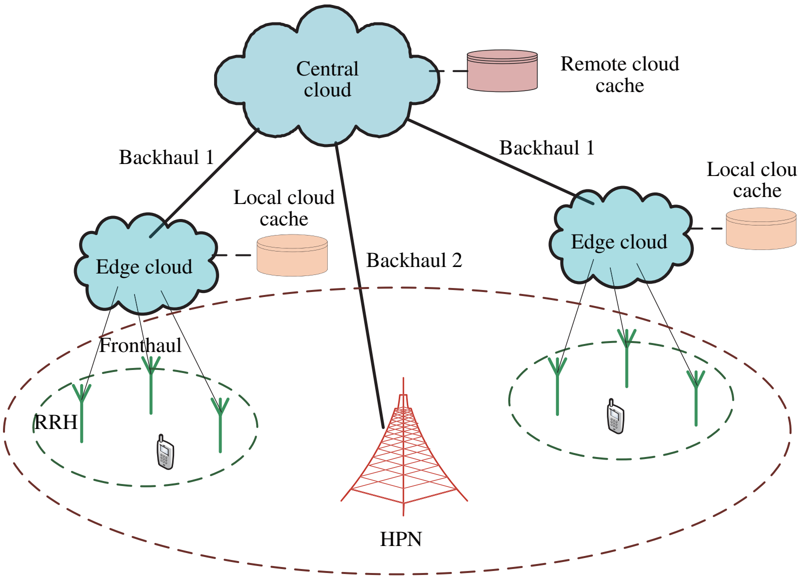

Figure 1 Network model with content caching and edge computing.

content caching and edge computing are explored.

4) As to the impact of content caching for radio resource allocation, we convert the problem into convex and solve it with subgradient method. We formulate the user scheduling scheme which takes computing resources of clouds into the consideration as a knapsack problem. Simulation results show that system performance gains are improved by about 50% with the joint resource allocation scheme.

The remaining of this paper is as follows. System model in the aspects of network, link, computing and cache is introduced in detail in Section 2. In Section 3, the effective capacity based joint resource allocation scheme is formulated and the simplification of the original problem is proposed to find the sub-optimal solutions. Performance evaluation is provided to testify the advantage of content caching and edge computing in Section 4. Finally, conclusions are drawn in Section 5.

2 System model

2.1 Network model

In this paper, we investigate the joint virtual resources allocation and scheduling scheme in the architecture of edge computing with content caching. As Figure 1 shows, RRHs have the access to clouds through the fronthaul links. The wired connections between Edge Clouds and Central Cloud are called backhaul links. RRHs are responsible for transmitting and receiving the radio signals and the baseband processing is executed at the clouds while HPN is mainly in charge of seamless control signals coverage. The storage and computing abilities of Edge Clouds are smaller than Central Cloud's while Edge Clouds bring resources closer to users reducing the latency caused by backhaul links. The radio, storage and computing resources are packaged into virtual resources to facilitate the scheduling of the system. With the purpose of satisfying the QoS requirements, the virtual resources are allocated to users which indicates the association schemes of RRHs and clouds.

2.2 Link model

We consider the downlink scenario of our proposed architecture with K mobile users which associate with N RRHs and Y clouds through M resource blocks (RBs). We assume that RB m can only be allocated to one user at a transmission time interval (TTI) and the allocation scheme is dynamic with the wireless channels changing. Denote $P_{k,m,n}$ as the transmit power allocated to the nth RRH on the mth RB serving the kth user. The channel gain from the kth user to the nth RRH on the mth RB is defined as $|H_{k,m,n}|^2$. The bandwidth of each RB is B and the power spectral density of noise is N_0. Hence, we can use $\mathrm{SNR}_{k,m,n}$ to indicate the channel condition of the kth user on the mth RB associating the nth RRH. The expression is written as follow:

$$\mathrm{SNR}_{k,m,n} = P_{k,m,n} \cdot |H_{k,m,n}|^2/(N_0 B). \tag{1}$$

According to Shannons capacity formula, the data rate achieved for the kth user on the mth RB associating the nth RRH can be formulated as follow:

$$R_{k,m,n} = B\log(1 + P_{k,m,n} \cdot \mathrm{CNR}_{k,m,n}), \tag{2}$$

where the channel-to-noise ratio is formulated as $\mathrm{CNR}_{k,m,n} = |H_{k,m,n}|^2/(N_0B)$. Denote a binary RB allocation matrix $a \in \{0,1\}^{K\times M\times N}$ to indicate the allocation scheme, where $a_{k,m,n} = 1$ represents that the mth RB is allocated to the nth BS to serve the kth user, otherwise $a_{k,m,n} = 0$. Hence, the sum data rate the kth user achieved at the nth RRH can be expressed as follow:

$$R_{k,n} = \sum_m a_{k,m,n} \cdot R_{k,m,n}. \tag{3}$$

2.3 Computing model

The computing capability of the yth cloud is characterized by F_y million operations per time slot (MOPTS). The computing resource allocation scheme is based on the computing complexity of the baseband processing. However, the exact relationship between baseband processing and consumed computing resources is still uncertain. As the paper [10] analyses, the computing in BBU pool is mainly related to the number of RRHs a user associated and the transmission data rate a user achieved. According to [10], the computing resource required for the kth user is defined as follow:

$$w_k = \mu {N_k}^3 + \gamma R_k + w_0, \tag{4}$$

where the number of RRHs the kth user associated is denoted as N_k and the computing complexity of it is as high as $\mathcal{O}(N^3)$. The consumption of computing resource is linear with the transmission data rate R_k. w_0 is the constant which indicates the basic demand of computing resources.

2.4 Cache model

In our senario, clouds are equipped with content cache to alleviate the burden of backhaul links. The storage ability of the yth cloud is defined as A_y. The scope of contents which are requested by mobile users is a library of F files. The numerical results reveal that content popularity distribution follows Zipf distribution. Without loss of generality, the probability of a user call for the jth content z_j is organized as decreasing order $z_1 \geqslant z_2 \geqslant \cdots \geqslant z_F$. The expression of z_j can be written as follow:

$$z_j = \frac{j^{-\gamma}}{\sum_{j=1}^{F} j^{-\gamma}}, \tag{5}$$

where $\gamma > 0$ is the Zipf parameter. Generally, γ indicates the degree of skewness of popularity distribution.

The requirement of a specific content would be checked in the local cache of Edge Clouds first. On the condition that the content is available in the Edge Cloud, the requirement will be responded immediately without the latency caused by backhaul links. Otherwise, the content will be searched in the cache of Central Cloud. For convenience, we assume all contents can be found out in the Central Cloud [9], so the content would be sent back to the Edge Cloud via backhaul links and then transmitted to users later. It is obvious that the application of content caching can reduce the service latency of users and provide a better QoS guarantee at the same time. Let the set $F = \{1, 2, 3, \ldots, F\}$ represents the content library and the size of the fth content is defined as X_f. The cache placement matrix is denoted as $C \in \{0,1\}^{F\times Y}$

to indicate whether the content f is cached in the yth cloud. Hence, the amount of contents the yth cloud can cache is constrained by the following formula:

$$\sum_{f=1}^{F} X_f \cdot C_{f,y} \leqslant A_y. \tag{6}$$

As mentioned in [11], the popularity distribution of the contents, which remains for several days, changes much more slowly than the state of RBs. It is not at the same time-scale of radio resource scheduling. Therefore, we consider the cache placement matrix $\boldsymbol{C}$ as a constant matrix when optimizing the power and RB allocation schemes.

2.5 Effective capacity

The deployment of edge clouds and the application of content caching bring the computing resources and storage resources closer to users so that the system performance is improved significantly. However, it is necessary to find a suitable metric to evaluate the QoS guarantee performance of our schemes. Effective capacity which is proposed to form channel model in link layer is convenient to convert the impact of latency into the user achieved data rate. The criteria indicates the maximum arrival data rate a wireless channel can support under QoS requirements with the queuing theory showing a good performance in [12]. The maximum arrival data rate a wireless channel can support is measured by the log-moment generation function as follow:

$$EC(\theta) = -\lim_{t\to\infty} \frac{1}{\theta t} \log E\left\{\mathrm{e}^{-\theta S(t)}\right\}, \tag{7}$$

where $S(t) = \int_0^t r(t)\mathrm{d}t$ indicates the throughput accumulated on time domain and θ represents the QoS guarantee parameter. A large value of θ implies a tighter QoS guarantee requirement. When the channel coefficients keep constant over the frame duration T and vary independently for each frame, the formula of effective capacity can be rewritten as follow:

$$EC(\theta) = -\frac{1}{\theta T} \log E\left\{\mathrm{e}^{-\theta T R[i]}\right\}, \tag{8}$$

where $R[i]$ indicates the instantaneous channel capacity during the ith frame.

Since the time-varying and randomness of wireless channels, effective capacity denotes QoS guarantee parameter θ to characterize the statistical QoS requirements. According to [13], the delay violation possibility can be formulated as follow:

$$\Pr\{D(\infty) > D_{\max}\} \approx \mathrm{e}^{-\theta \cdot D_{\max}}, \tag{9}$$

where $D(\infty)$ is the steady-state delay experienced by a data flow, and $D_{\max}$ represents the delay bound.

It is obvious that the probability of $D(\infty)$ exceeding a delay bound $D_{\max}$ decreases with the increasing of θ,which indicates a tighter QoS constraints. Note that when θ approaches to zero, the effective capacity converges to the ergodic capability [14].

Based on the theorem 2 in [15], to achieve the same delay experience, the QoS guarantee parameter θ of different users should satisfy the following condition:

$$\theta_{b,k}^{f} = \frac{1}{1 - \frac{2X_f}{r_{\mathrm{BH}} \cdot D_{\max}}} \theta_{a,k}^{f}. \tag{10}$$

If the content requested by the kth user is cached in the nth RRH, in other words $C_{f_k,n} = 1$, $\theta_{k,n} = \theta_{a,k}^{f}$. Otherwise, $\theta_{k,n} = \theta_{b,k}^{f}$.

3 Effective capacity based joint resource allocation scheme

In this section, we firstly formulate the resource allocation problem with the constraints of radio, storage and computing resources. Then we simplify the original problem with the Taylor expansion of the objective function and use Lagrange dual optimization to turn the problem more tractable to solve. And then we optimize the resource allocation scheme of radio and storage resources jointly. Using effective capacity theory, the impact of content caching is calculated with different QoS guarantee parameters. The subgradient method is used to find the sub-optimal solution and a low-complexity algorithm is proposed. Finally, the allocation scheme of cloud computing resources for baseband processing is treated as a knapsack problem.

3.1 Problem formulation

Based on the system model and effective capacity theory previously mentioned, the effective capacity achieved by the kth user in the nth RRH can be expressed by the following formulation:

$$\mathrm{EC}(\theta_{k,n}) = -\frac{1}{\theta_{k,n}T}\log E\left\{\mathrm{e}^{-\theta_{k,n}TR_{k,n}}\right\}. \tag{11}$$

The problem can be formulated as follows with the purpose of maximizing the sum effective capacity of the whole system.

$$\max \sum_k \sum_n \mathrm{EC}(\theta_{k,n}), \tag{12}$$

$$\text{s.t.} \sum_k \sum_m a_{k,m,n} P_{k,m,n} \leqslant P_n, \tag{13}$$

$$\sum_k f_{k,y} \leqslant F_y, \tag{14}$$

$$\sum_f X_f \cdot C_{f,y} \leqslant A_y, \tag{15}$$

$$\sum_k \sum_n a_{k,m,n} = 1. \tag{16}$$

In constrain (13), the whole transmitting power of users associating with RRH n should be no more than P_n. The constraints of computing resources and storage resources are defined as (14) and (15) respectively. Constraint (16) denotes the RB allocation limitation that each RB should only be allocated to one user at the same time. As aforementioned, the allocation of storage and computing resources can be treated as constant when optimizing the power and RB allocation schemes.

Obviously, the optimal problem is a Mixed-Integer Nonlinear Programming (MINLP) and it is highly complicated to find the global optimization solution. In the next section, the original problem is simplified and converted to a convex problem. After that a low-complexity algorithm is proposed to solve the problem.

3.2 Optimization reformulation

Because the original objective is too complex to solve directly, we make the Taylor expansion of it to make the problem more tractable. Taking notice of $\theta_{k,n}T \to 0$ and $\mathrm{e}^{-\theta_{k,n}TR_{k,n}} \to 1$, we make the Taylor expansion around the point 1.

$$\begin{aligned}&\sum_k \sum_n -\frac{1}{\theta_{k,n}T}\log E\left\{\mathrm{e}^{-\theta_{k,n}TR_{k,n}}\right\}\\ &= \sum_k \sum_n -\frac{1}{\theta_{k,n}T}\log E\left\{\prod_m (1+P_{k,m,n}\mathrm{CNR}_{k,m,n})^{-\alpha_{k,m,n}\cdot\beta_{k,n}}\right\}\end{aligned}$$

$$
\begin{aligned}
&= \sum_k \sum_n -\frac{1}{\theta_{k,n}T} \left\{ \left\{ E\left\{ \prod_m (1+P_{k,m,n}\mathrm{CNR}_{k,m,n})^{-\alpha_{k,m,n}\cdot\beta_{k,n}} \right\} - 1 \right\} \right. \\
&\quad \left. -\frac{1}{2} \left\{ E\left\{ \prod_m (1+P_{k,m,n}\mathrm{CNR}_{k,m,n})^{-\alpha_{k,m,n}\cdot\beta_{k,n}} \right\} - 1 \right\}^2 + \cdots \right\}, \qquad (17)
\end{aligned}
$$

where $\beta_{k,n} = \frac{\theta_{k,n}TB}{\ln 2}$. Omitting high order terms, the approximation of the objective function can be rewritten as follow:

$$
\max \left(\sum_k \sum_n -\frac{1}{\theta_{k,n}T} \left\{ E\left\{ \prod_m (1+P_{k,m,n}\mathrm{CNR}_{k,m,n})^{-\alpha_{k,m,n}\cdot\beta_{k,n}} \right\} - 1 \right\} \right). \qquad (18)
$$

Since the popularity distribution of contents changes slowly, ranging from 3 hours to weeks [11], the scheduling of storage resource is much slower than the radio scheduling. So that we can consider the allocation of storage resource is constant when we optimize the RB allocation scheme. Based on (10) we convert the system gain of content caching into the QoS guarantee parameter. The scheduling of computing resources can be regarded as the knapsack problem, in this paper each cloud is treated as a knapsack whoes capacity limit is represented in constraint (14). w_k is regarded as the item packing into the knapsack while the effective capacity achieved by the kth user is treated as the value of items. Above all, the constraints of (14) and (15) can be omitted when we optimize radio resource allocating scheme. The Lagrange function of the primal problem is given by

$$
\begin{aligned}
L(a,P,\lambda) =& \sum_k \sum_n -\frac{1}{\theta_{k,n}T} \left\{ E\left\{ \prod_m (1+P_{k,m,n}\mathrm{CNR}_{k,m,n})^{-\alpha_{k,m,n}\cdot\beta_{k,n}} \right\} - 1 \right\} \\
&+ \sum_n \lambda_n \left(P_n - \sum_k \sum_m P_{k,m,n} \right), \qquad (19)
\end{aligned}
$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_N]$ is the Lagrange vector associated with the power constraints.

The Lagrange dual optimization problem can be expressed as follow:

$$
\begin{aligned}
g(\lambda) =& \max \left(\sum_k \sum_n -\frac{1}{\theta_{k,n}T} \left\{ E\left\{ \prod_m (1+P_{k,m,n}\mathrm{CNR}_{k,m,n})^{-\alpha_{k,m,n}\cdot\beta_{k,n}} \right\} - 1 \right\} \right. \\
&\left. + \sum_n \lambda_n \left(P_n - \sum_k \sum_m P_{k,m,n} \right) \right). \qquad (20)
\end{aligned}
$$

And the dual optimization problem can be written as follow:

$$
\min \{g(\lambda)\} \ \text{ s.t. } \lambda \geqslant 0. \qquad (21)
$$

3.3 Effective capacity based joint resource allocation scheme

The dual problem is convex and more tractable for the reason that we can use convex optimization methods to sovle it. According to [16], a subgradient method is proposed to solve the convex optimization problem by updating λ simultaneously. Similar with the approach [17], the subgradient of (20) can be written as follow:

$$
\Delta\lambda_n = P_n - \sum_k \sum_m P_{k,m,n}. \qquad (22)
$$

The updating of λ follows the following formula:

$$
\lambda_n^{s+1} = \left\{ \lambda_n^{s+1} - \tau^s \left[P_m - \sum_k \sum_m P_{k,m,n} \right] \right\}^+, \qquad (23)
$$

where s denotes the iteration number and τ^s represents a proper step size. There are several step size rules in subgradient method [18] and diminishing step size rule shows a better performance. Then we choose this rule to design the value of τ^s. A typical example is $\tau^s = \frac{a}{\sqrt{s}}$, where $a > 0$.

Suppose $P^*_{k,m,n}$ is the optimal solution of power assignment. Applying the KKT conditions

$$\frac{\partial L(\lambda)}{\partial P^*_{k,m,n}} = 0. \tag{24}$$

We can calculate the expression of $P^*_{k,m,n}$ is

$$P^*_{k,m,n} = \left[\left(\frac{\beta_{k,n}}{\lambda_n \theta_{k,n} T} \prod_{\mathrm{m}} \mathrm{CNR}_{k,m,n}{}^{-\beta_{k,n}}\right)^{\frac{1}{M\beta_{k,n}+1}} - \frac{1}{\mathrm{CNR}_{k,m,n}}\right]^+ . \tag{25}$$

Substituting (25) into (20), the form of Lagrange dual function of is given as

$$\begin{aligned} g(\lambda) = \max & \left(\sum_k -\frac{1}{\theta_{k,n} T} \left\{ E\left\{ \left(\left(\frac{\beta_{k,n}}{\lambda_n \theta_{k,n} T} \prod_{\mathrm{m}} \mathrm{CNR}_{k,m,n}{}^{-\beta_{k,n}} \right)^{\frac{1}{M\beta_{k,n}+1}} \mathrm{CNR}_{k,m,n} \right)^{-a_{k,m,n}\beta_{k,n}} \right\} - 1 \right\} \right. \\ & \left. + \sum_n \lambda_n \left(P_n - \sum_k \sum_m a_{k,m,n} \left[\left(\frac{\beta_{k,n}}{\lambda_n \theta_{k,n} T} \prod_{\mathrm{m,n}} \mathrm{CNR}_{k,m,n}{}^{-\beta_{k,n}} \right)^{\frac{1}{M\beta_{k,n}+1}} - \frac{1}{\mathrm{CNR}_{k,m,n}} \right]^+ \right) \right). \end{aligned} \tag{26}$$

We can use the dual decomposition method and get the effect of allocating the mth RB to the kth user associated with the nth RRH. The expression can be measured as follow:

$$\begin{aligned} g_{k,m,n} = & -\frac{1}{\theta_{k,n} T} \left\{ E\left\{ \left(\left(\frac{\beta_{k,n}}{\lambda_n \theta_{k,n} T} \prod_{\mathrm{m}} \mathrm{CNR}_{k,m,n}{}^{-\beta_{k,n}} \right)^{\frac{1}{M\beta_{k,n}+1}} \mathrm{CNR}_{k,m,n} \right)^{-\beta_{k,n}} \right\} - 1 \right\} \\ & - \lambda_n \left[\left(\frac{\beta_{k,n}}{\lambda_n \theta_{k,n} T} \prod_{\mathrm{m}} \mathrm{CNR}_{k,m,n}{}^{-\beta_{k,n}} \right)^{\frac{1}{M\beta_{k,n}+1}} - \frac{1}{\mathrm{CNR}_{k,m,n}} \right]^+ . \end{aligned} \tag{27}$$

It is obvious that optimal user k and RRH n which RB m should be allocated is the *(k,n)* with the maximum value of $g_{k,m,n}$. So we can get the RB allocation matrix as follows:

$$a_{k,m,n} = \begin{cases} 1, & \mathrm{argmax}\{g_{k,m,n}\}, \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Define the diagonal matrixes a_k to indicate the association relationship between the kth user and RRHs. The expression can be written as follow:

$$a_k = \mathrm{diag}(a_{k,1,1}, a_{k,1,2}, \ldots, a_{k,M,N}). \tag{29}$$

The expression of N_k in (4) can be calculated by $N_k = \mathrm{Tr}(a_k)$. Based on (28), the data rate user k achieved is $R_k = \sum_n R_{k,n} = \sum_n \sum_m a_{k,m,n} \cdot R_{k,m,n}$. The computing resources for supporting baseband processing of the kth user can be obtained according to (4). In this paper, we regard the computing resources in clouds as the volume of knapsack and w_k is regarded as the size of the item with the value of effective capacity. Thus the allocation of computing resources can be converted into a knapsack problem which can be solved by the greedy algorithm.

Table 1 Simulation parameters

System parameters	
Number of BSs	7
Number of subchannels	50
Maximum power of BSs	46 dBm
Carrier frequency	2 GHz
Bandwidth	10 MHz
Cell average radius	500 m
Pathloss model	$\mathrm{PL} = 128.1 + 37.6\log_{10}d, d(km)$
Shadowing standard deviation	8 dB
Fast fading	Rayleigh fading
Noise Density	-174 dBm/Hz
Cache parameters	
Backhaul rate	500 Mbps
Mean of content size	30 Mb
Zipf parameter	0.5
Computing parameters	
Edge cloud ability	1000 MOPTS
Central cloud ability	3000 MOPTS

4 Performance evaluations

4.1 Simulation environment

In this section, we provide simulations to evaluate the performance of joint scheduling scheme of radio, computing and storage resources in the downlink scenario. Our numerical evaluations are based on 19-cell mobile network which contains 3 sectors in each cell. Each sector is deployed with one RRH, so the total number of RRHs is 57. There are 3 edge clouds and 1 central cloud to meet the baseband processing requirements of these RRHs. The edge cloud brings computing and storage resources much closer to users while the central cloud has more resources to schedule. In our scenario, each edge cloud is configured with 1000 MOPTS computing resources and central cloud is configured with 3000 MOPTS [10]. We define the content library with 50 files and the size of each file is a normal random variable with the mean of 30 Mb. The content requests of users follow a Zipf distribution with Zipf parameter $\gamma = 0.5$. More details of simulation environment settings are listed in Table 1.

4.2 Simulation results

We firstly verify the performance gain with content caching. System throughput is calculated by the sum of all users effective capacity. In Figure 2, we investigate the performance of content caching with the cache size of 25 files in each Edge Cloud. The system performance without cache is simulated as the baseline with QoS guarantee parameter $\theta = 0.005$. In the condition of content caching, QoS guarantee parameter θ is calculated by (10). As Figure 2 shows, system throughput is improved about 3 Mbps when we pre-fetch the popular contents in Edge Clouds. That is because contents are much closer to users, which can greatly reduce the service latency caused by backhaul links and the radio resources can be used more effectively. When the number of users is large enough, the radio resources tend to saturation, so the system throughput does not improve any more.

Since content caching can improve the system throughput significantly, it is a natural idea to investigate the relationship between system throughput and the number of cached files. As Figure 3 shows, the system performance with different user number is simulated following the cache size rising. We can find out that system throughput is improved with cache size rising. It illustrates that with cache size increasing the content hit probability rises at the same time, resulting more contents can be found at edge clouds. Thus,

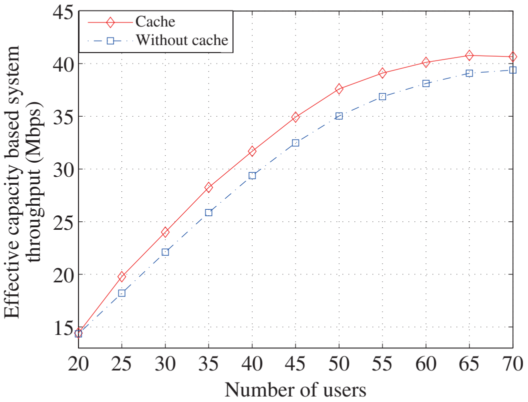

Figure 2 (Color online) System throughput gain with content caching.

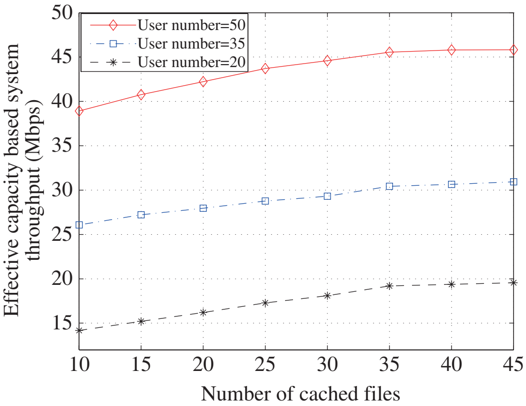

Figure 3 (Color online) System throughput with number of cached files.

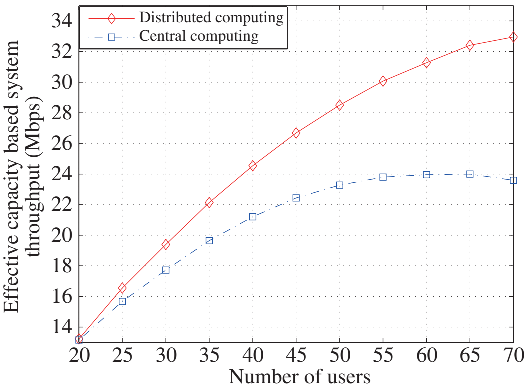

Figure 4 (Color online) System throughput gain with edge computing.

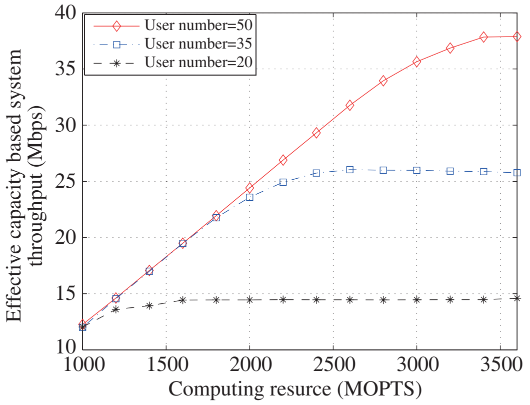

Figure 5 (Color online) System throughput gain with edge computing.

the backhaul latency can be reduced and user data rate is higher.

As aforementioned, the allocation scheme of cloud computing resources supporting baseband process is treated as a knapsack problem. In our scenario, the total computing resources of edge clouds are equivalent with the ones of central cloud. As Figure 4 shows, the edge clouds show a better performance at system throughput. It demonstrates even though the computing resources in edge clouds is smaller the distance reduction between users and computing resources can bring about 50% increase in the performance gain.

Finally, we investigate the influence of computing resources on system throughput to induct the configuration of cloud. Through Figure 5, we find that as computing resource increases, the system throughput improves at first and then tends to steady. As the simulation results reveal, we can find out the proper computing resources to support the associated users in a cloud. For example, a cloud equipped with 2400 MOPTS computing resources is capable to support the baseband processing of 35 users. With the user number rising, the need of computing resources increases at the same time while the system throughput is higher.

As simulation results above show, effective capacity based joint resource allocation scheme has the advantage of releasing the potentials of storage and computing resources. The more storage resources of Clouds can cause more user association, which leads to more demand of computing resources. The system throughput is improved by about 50% with content caching and edge computing. The user association and scheduling scheme are not only based on wireless channel conditions but also storage and computing resources. Furthermore, the requirement of storage and computing resources for communication is investigated, which can instruct the configuration of storage and computing resources in clouds.

5 Conclusion

In this paper, we abstract the physical resources into the virtual ones and proposed the effective capacity based joint resource allocation scheme in the downlink scenario. To evaluate the influence of introducing storage and computing resources into the joint resource allocation scheme, effective capacity is considered as the suitable metric to include the service latency into user data rate. Taking advantage of edge computing and content caching, users are much closer to physical resources which make a difference on user association and scheduling scheme. Users prefer to associate with the cloud which has already cached the required content in order to reduce the latency while scheduling is not only based on the wireless channel conditions but also the baseband processing ability of computing resources in clouds. Simulation results are provided to demonstrate that performance gain of the joint optimal resource allocation scheme could achieve about 50% increase in system throughput. Furthermore, the proper computing and storage capacity configuration for clouds are investigated to achieve a high system throughput.

Acknowledgements This work was supported by International Collaboration Project (Grant No. 2015DFT10-160), National Natural Science Foundation of China (Grant Nos. 61471068, 61421061, 61325006), National High-Tech R&D Program of China (863) (Grant No. 2014AA01A701), National Major Project (Grant No. 2016ZX03001009-003), Beijing Training Project for the Leading Talents in S&T (Grant No. Z141101001514026), and 111 Project of China (Grant No. B16006).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Tesema F, Awada A, Viering I, et al. Evaluation of context-aware mobility robustness optimization and multi-connectivity in intra-frequency 5G ultra dense networks. IEEE Wirel Commun Lett, 2016, 5: 608–611

2 Blasco P, Gündüz D. Learning-based optimization of cache content in a small cell base station. In: Proceedings of 2014 IEEE International Conference on Communications (ICC), Sydney, 2014. 1897–1903

3 Peng M G, Wang C G, Li J, et al. Recent advances in underlay heterogeneous networks: interference control, resource allocation, and self-organization. IEEE Commun Surv Tut, 2015, 17: 700–729

4 Xu D T, Ren P Y, Sun L, et al. Precoder-and-receiver design scheme for multi-user coordinated multi-point in LTE-A and fifth generation systems. IET Commun, 2016, 10: 292–299

5 Liang C C, Yu F R. Wireless virtualization for next generation mobile cellular networks. IEEE Wirel Commun, 2015, 22: 61–69

6 Peng M G, Li Y, Jiang J M, et al. Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies. IEEE Wirel Commun, 2014, 21: 126–135

7 Sardellitti S, Barbarossa S, Scutari G. Distributed mobile cloud computing: joint optimization of radio and computational resources. In: Proceedings of 2014 IEEE Globecom Workshops (GC Wkshps), Austin, 2014. 1505–1510

8 Cha M, Kwak H, Rodriguez P, et al. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, New York, 2007. 1–14

9 Zhao Z Y, Jia S W, Li Y, et al. Performance analysis of cluster content caching in cloud-radio access networks. In: Proceedings of 2015 IEEE Globecom Workshops (GC Wkshps), San Diego, 2015. 1–6

10 Liao Y, Song L Y, Li Y H, et al. Radio resource management for cloud-RAN networks with computing capability constraints. In: Proceedings of 2016 IEEE International Conference on Communications, Kuala Lumpur, 2016. 1–6

11 Shanmugam K, Golrezaei N, Dimakis A G, et al. FemtoCaching: wireless content delivery through distributed caching helpers. IEEE Trans Inform Theory, 2013, 59: 8402–8413

12 Wu D P, Negi R. Effective capacity: a wireless link model for support of quality of service. IEEE Trans Wirel Commun, 2003, 2: 630–643

13 Wu D P, Negi R. Effective capacity-based quality of service measures for wireless networks. In: Proceedings of the 1st International Conference on Broadband Networks, San Jose, 2004. 527–536

14 Liu L J, Chamberland J F. On the effective capacities of multiple-antenna Gaussian channels. In: Proceedings of 2008 IEEE International Symposium on Information Theory, Toronto, 2008. 2583–2587

15 Zhao Z Y, Peng M G, Ding Z G, et al. Cluster content caching: an energy-efficient approach to improve quality of service in cloud radio access networks. IEEE J Sel Areas Commun, 2016, 34: 1207–1221

16 Boyd S, Vandenberghe L. Convex Optimization. Cambridge: Cambridge University Press, 2004. 1–50

17 Han X, Chen H F, Xie L, et al. Effective capacity region in a wireless multiuser OFDMA network. In: Proceedings of Global Communications Conference (GLOBECOM), Anaheim, 2012. 1794–1799

18 Boyd S, Mutapcic A. Subgradient Methods. Stanford: Stanford University Press, 2006. 1–35