Luo *et al. EURASIP Journal on Wireless Communications and Networking* (2016) 2016:23
DOI 10.1186/s13638-015-0503-2

EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

An optimal data service providing framework in cloud radio access network

Yuansheng Luo[1,2], Kun Yang[3*], Qiang Tang[1,2], Jianming Zhang[1,2], Ping Li[1,2] and Shi Qiu[4,5]

Abstract

Much work has been conducted to design effective and efficient algorithms for quality of service (QoS)-aware service computing in the past several years. The wireless mobile computing and cloud computing environments have brought many challenges to QoS-aware service providing. Mobile cloud computing (MCC) and cloud radio accessing networks (C-RANs) are the new paradigms arising in recent years. This work proposes a wireless data service providing framework in C-RAN aiming to provide data service in C-RAN by a more efficient way. The efficiency is measured by cost with time constraint. An abstract formal model is built on the proposed framework, and the corresponding optimal solution is deduced theoretically using queuing theory and convex optimization. The simulation results show that the proposed optimal strategy on the optimal solution works well and has a better performance than compared one.

Keywords: Service computing, Service providing, Cloud radio accessing networks, Queueing theory

1 Introduction

Service computing, as a bridge between modern business services and information technologies, has witnessed the communication infrastructure and environment, which the service-oriented architecture (SOA) build on, varied from wired network to wireless mobile network, from centralized client-server model to distributed peer-to-peer (p2p) model until today's centralized cloud data centers in the past decade. The quality of service (QoS) is always one of the main spots that the users and researchers concern with in both service computing and mobile cloud computing area [1–6]. A service provider should ensure the consistency of the service level agreements (SLA) [3, 7], that is, the QoS in the service advertisement should be consistent with the QoS of the real service delivered to users. However, due to the variable network environments, many efforts should be devoted for service providers to improve the QoS and achieve the QoS level as they promised in the SLA. The QoS-aware service selection has been intensively discussed in many papers [8, 9]. Recently, many research works on QoS-aware service selection in wireless mobile networks and in cloud computing have been proposed [10–13]. The success of these mobile systems lies in their ability to provide users with cost-effective services that have the potential to run anywhere, anytime, and on any device without (or with little) user attention [14]. As far as network-level QoS is concerned, much work has been carried out for both wired networks (i.e., the Internet) and wireless mobile ad hoc networks (MANETs) and mobile cellular networks. For instance, originated for the best-effort wired Internet, Integrated Service (IntServ) provides guaranteed bandwidth for each flow whereas Differentiated Services (DiffServ) offers guarantees on a per service class basis [15]. In the current LTE 4G network, there are 9 bearer types that are used to achieve the different QoS levels for different client application services [6]. While the device-to-device (D2D) communications as a more practical paradigm compared to conventional ad hoc networks, and a more cost-efficient paradigm compared to cellular networks, are advocated by many researchers [16–18]. On the other hand, the cloud data centers can provide computing resources on users' demands and the resources are almost infinite compared to the mobile devices. Thus, tenanting the service in cloud or transmitting the service with data to cloud to execute the computing in data centers is an efficient way to gain the powerful computing capability conveniently. The latter is usually called as offloading [19]. With the prevailing of mobile communication techniques and devices, wireless accessing of cloud data centers, also

*Correspondence: kunyang@essex.ac.uk
[3]School of Computer Science and Electronic Engineering, University of Essex, Colchester Essex, UK
Full list of author information is available at the end of the article

called as mobile cloud computing, has become the next hotspot both in academia and industry. In the context of this paper, the *data service* is specifically referred to a category of services which transmit users' files or data through relay nodes to cloud data center for the purpose of cloud computing and cloud storage.

Not only the software as a service (SaaS) can gain the profit from cloud computing, the cloud platform can be beneficial to radio access networks (RANs) as well, which is an arising communication technology known as cloud radio accessing networks (C-RANs) [3–5, 20–24]. Unlike conventional RANs, the C-RANs decouple the baseband processing unit (BBU) from the remote radio head (RRH), allowing for centralized operation of BBUs and scalable deployment of lightweight RRHs as small cells [20]. BBUs locate in the signal processing cloud with high-speed fronthaul links to the distributed RRHs. The signal processing cloud is connected to backbone network by backhaul links. C-RAN is being advocated both by operators (e.g., China Mobile, SoftBank) as well as service providers (e.g., LightRadio, Liquid Radio). C-RAN is going to be the core technique in the next-generation broadband wireless networks.

In this paper, a QoS-aware data service providing framework is put forward. This framework is based on the C-RAN. The queuing theory is used to theoretically analyze the framework, and optimal service providing strategy is proposed for providers, which aims to minimize the cost for running a candidate data service which is subject to the execution duration constraint. At last, a simulation is conducted to do an empirical study for the proposed framework.

2 Related work

Zeng et al. [9] and Yu et al. [25] presented their models for calculating QoS of composite services. These works are based on the assumption that the service provider can provide a description or agreement on QoS, such as the service level agreement (SLA) [7]. Except for local QoS attributes, network QoS should be considered during the decision-making procedure on selecting services [10, 12, 13]. Some works evaluate network delay according to the geographic locations where the services locate [26]. Service overlay is constructed according to different QoS levels in [27]. The availability or error tendency of services also can be evaluated dynamically according to current states of services with the historical data [28]. Current trend in service selection is selecting services in cloud platform and accessing the cloud services through wireless networks [11, 12, 26]. Offloading services to data centers is also a prevailing research field [19]. The D2D service selection is similar to an ad hoc service selection, but the D2D concept is a newly arising idea in recent years [16–18]. Thus, the work on D2D service selection is few. There are some optimal resource allocation works in C-RAN researching community [20–23]. Cai et al. [20] studied the topology configuration and rate allocation problem in C-RAN with the objective of optimizing the end-to-end performance of mobile cloud computing users. Li et al. [22] studied the queue-aware power and rate allocation optimization for delay-sensitive traffic in fronthaul constrained C-RAN. Liu et al. [23] proposed algorithms to solve the joint optimization problem which combine power control and fronthaul rate allocation for throughput maximization in C-RAN. Sundaresan et al. proposed a scalable, lightweight framework FluidNet for realizing the full potential of C-RAN [21], which can reconfigure its fronthaul and tailor transmission strategies which provide improvement in satisfying traffic demands, while reducing the compute resource usage in the BBU pool compared to baseline transmission schemes. All these works are devoted to improve the link layer QoS in C-RAN, such as throughput, from global viewpoint of optimizing wireless communication service. Dynamically configurable, scalable, sharable, and re-allocable per demand on the communication resources are some of promising benefits of C-RAN. These benefits rely on the virtualized components in BBU pool [24]. Thus, the delay exposed in the BBU pool and the respecting "cloudification" issues for the aggregation of BBUs have to be studied. The work presented in [3, 4] puts forward an integral service framework for optimizing quality of experience (QoE) and efficiency of mobile cloud networking (MCN) architecture based on the LTE C-RAN, which is a package of solutions for communication operators but not for service providers.

The contribution presented in this paper is to propose an optimal framework for application-level data service providers in C-RAN, which minimizes the cost for tenanting a data transferring service when subject to the execution duration constraint.

3 Motivation scenario

In Hunan province, each highway toll station must transfer the collected information on vehicles and roads to the Management Center located in Changsha city from time to time. The Management Center (MC) plays a role as centralized data center for highway information computing, storage, and monitoring. A toll station can transfer the traffic information by private highway fiber network or by VPN rented from China Telecom as a backup line. While in some areas, especially the mountainous areas, the wired channel to MC is very disruption prone due to disrepair and aging of lines. Toll stations would have to choose the wireless broadband service rented from China Mobile when the wired channel failed and toll stations were disconnected from the MC. But in these areas, all stations are not equipped with access points connecting RRHs, and those with access points may have a different wireless channel QoS due to the difference of distance, terrain, and devices. One of possible solutions under this condition is to select a toll station as a relay node, which should have the best wireless QoS compared to other stations nearby. The nearby stations could transfer the traffic information to the relay node by wired channel (providing the wired channel between them is working well) or by D2D communications. After that, the relay node would transfer the traffic information to MC by wireless broadband channel. Thus, it is desirable for the relay node to be dynamically configured to choose an optimal solution for reducing communication cost while subject to time constraints. The service provided by this relay node is the *data service* in this context. The relay node is a service provider who should take responsibilities of configuring and optimizing the data transferring procedure.

4 System model

The system model consists of three parts: users, communication channel, and services. Users come to cloud service broker for selecting optimal services. In conventional case, a service selection broker, which acts as a service registry as well, should be an independent third party locating at a location different with service repository and service providers. In cloud computing context, the broker and services co-locate in a data center in a centralized way. Thus, the service selection broker can be treated as a cloud service as well. When a user selects a service, a communication link is used to connect the user to the cloud service server for data transmitting. The topology of the service providing framework based on cloud radio accessing network is depicted in Fig. 1.

As depicted in Fig. 1, each mobile device connects to the RRHs in a small cell directly or by D2D communications. The service providers have mobile client devices in small cells with other users together. The service users can access cloud services provided by service provider through providers' mobile devices using D2D communications. RRH connects to the signal processing cloud by fronthaul link. In signal processing cloud, re-configurable virtual BBU pools process the signal on the allocated data rate for each user. The backhaul link connects the signal processing cloud to backbone network and establishes connections to SaaS cloud data centers. Thus, a mobile user can establish a connection to a virtual cloud service server by a virtual link. The queue happens at the signal processing cloud because the data rate allocated to a user is limited on the operators' strategy. In SaaS cloud, there would not be a queue event (or the queuing duration is very small) since the cloud resources can be elastically allocated to each user on their demands. The abstract model of service providing framework can be depicted in Fig. 2. We assume the size of each job coming to the queue is much less than the channel transmission data rate in a time unit. For simplicity, we assume each job size is one unit, for example, 1 MB. The service providing framework (SPF) is expressed as following triple tuple:

$$\mathrm{SPF} =< Q, C, S >$$

where Q is the user job queue in service provider, C is communication channel, S is cloud service.

$$Q =< \lambda, L, \mu > \quad C =< \mu, C_1, C_2 > \quad S =< C_s >$$

where λ is arrival rate. L is queue length. μ is service rate, which is also the transmission data rate in channel C. C_1 and C_2 are costs for using the channel. C_s is the cost for hiring the cloud service.

In current mobile communication network, the processing ability of BBUs and the number of accommodated clients are limited. Considering the promising scalability

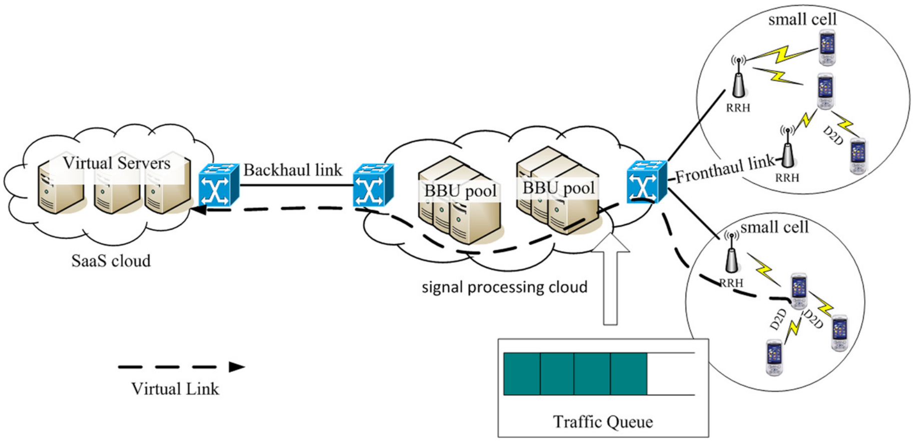

Fig. 1 Topology of service providing framework based on C-RAN

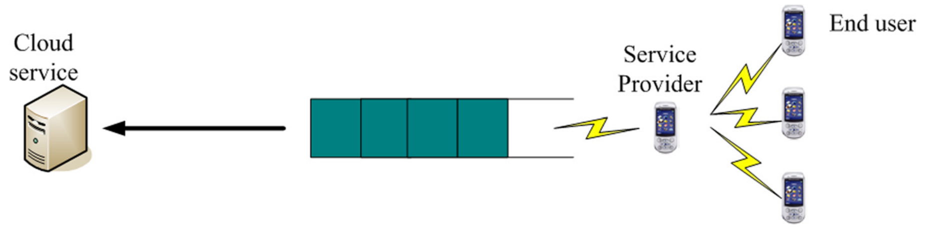

Fig. 2 Abstract model of service providing framework

of C-RAN under heavy load [3–5, 24], we assume that the BBU pool could provide elastic computing capability for meeting the loads of clients on demand. Thus, we have the following dynamical service rate (data rate) reconfiguring procedure used by service providers:

Step 1. Monitoring the job queue, evaluating the arriving rate periodically;
Step 2. Recalculating the optimal service rate on arriving rate;
Step 3. Applying for reallocating data rate to communication operators according to the new service rate;
Step 4. Transmitting data in new data rate and going to step 1.

5 Optimal service providing formulation

As depicted in Fig. 2, a service provider can provide service to multiple users. The service provider needs to transmit the user data to the virtual server in SaaS cloud. User tasks arrive at the provider according to the Poisson distribution, and the arrival rate is λ. The service time of the provider before the user data reaches SaaS cloud obeys the exponential distribution; the service rate is μ. Explicitly, the queuing service and cloud service represent different services. The queuing service consists of forwarding user requests and user data transmitting. The cloud service is the service that processes the user request and executes task on the user data. The cloud service time is correlated to CPU frequency and memories of the virtual server and task type; it is assumed to be a constant t_s for any determined task in this paper. The queuing service time consists of forwarding time and data transmitting time. Forwarding time is correlated to the capability of a provider's device, which can be treated as a constant with small value as well. The data transmitting time is correlated to the data rate allocated to service provider by the signal processing cloud operator, which should be a main factor on the queuing service time. Hence, without loss of generality, the queuing service time is determined by the data transmitting time. The average queuing service time of each task is $1/\mu$. The service selection queuing problem is a typical $M/M/1$ queuing problem. The formulation for the optimal service providing problem can be expressed as following formulas:

$$\min\left\{F(\mu) = C_1\mu + C_2\frac{\lambda}{\mu-\lambda} + C_s\right\} \tag{1}$$

$$\text{subject to}\ \begin{cases} ll\dfrac{1}{\mu-\lambda} + t_s \leq T \\ \dfrac{\lambda}{\mu} \leq 1 \end{cases} \tag{2}$$

where C_1 is the cost of one unit size data transmitted through a virtual link, and C_2 is the cost of one unit size data staying in the system, which represents the cost of the system maintaining the temporary data and states. All costs are in one unit time. C_s represents the cost of cloud service in one unit time. All these costs and data arriving rate can be determined by the historical log, statistics, and operation experiences. For example, the communication operator would give a price per data size of traffic flow. Cloud service provider should give the price for tenanting a service as well. $\lambda/(\mu-\lambda)$ is the average queue length staying in the system. $1/(\mu-\lambda)$ is the average staying time per unit data staying in the system. T is the total execution time constraint for a task. t_s is the execution time per data unit of cloud service. The T can be specified in the SLA file in service selection broker. For example, a typical "less equal than" restrict can be expressed as following WSLA [7]:

```
< Predicate xsi:type="wsla:LessEqual">
<SLAParameter> ResponseTime</SLAParameter>
<Value>1</Value>
</Predicate>
```

which means that the response time should be less equal than one unit time. The formula (1) is the objective function, which means that the optimizing objective is to minimize the total costs. The first part in (2) is a constraint, which means that the total execution time should be less equal than T. If t_s is bigger than T, then the equation has no feasible solution. Thus, the t_s must be less than T. The first part of (2) can be transformed to the following style:

$$1 + \triangle t(\lambda - \mu) \leq 0 \tag{3}$$

where $\triangle t = T - t_s$.

(3) implies the second inequations of (2), hence the original equation can be expressed as:

$$\min \left\{ F(\mu) = C_1\mu + C_2\frac{\lambda}{\mu - \lambda} + C_s \right\} \tag{4}$$

$$\text{subject to } 1 + \triangle t(\lambda - \mu) \leq 0 \tag{5}$$

This equation is a convex optimization problem. Let L be the Lagrangian function:

$$L(\mu,\alpha) = F(\mu) + \alpha(1 + \triangle t(\lambda - \mu)) \tag{6}$$

where α is the Lagrange multiplier. According to the Karush-Kuhn-Tucker (KKT) conditions for optimality [29], we have:

$$C_1 - C_2\frac{\lambda}{(\mu - \lambda)^2} - \triangle\alpha = 0 \tag{7}$$

$$\alpha(1 + \triangle t(\lambda - \mu)) = 0 \tag{8}$$

$$1 + \triangle t(\lambda - \mu) \leq 0 \tag{9}$$

$$\alpha \geq 0 \tag{10}$$

From the above (7), (8), (9), and (10), we can deduce two possible optimal solutions as follows:

$$\mu^* = \lambda + \sqrt{\frac{C_2}{C_1}\lambda} \tag{11}$$

$$\mu^* = \lambda + \frac{1}{\triangle t} \tag{12}$$

Since according to the constraint (9), we can get:

$$\mu^* \geq \lambda + \frac{1}{\triangle t} \tag{13}$$

Hence, we have the following results:

Case 1:
$\triangle t \geq \sqrt{\frac{C_1}{C_2\lambda}}$, then formula (11) should be chosen as the optimal solution, *otherwise*,
Case 2:
Formula (12) should be chosen as the optimal solution.

6 Discussion

An optimal solution which needs the least cost on the task should be chosen, because if the service provider's service rate μ increases, the cost paying on renting service will increase accordingly. On the other hand, if μ decreases, the waiting cost should increase. The optimal solution is a tradeoff which can get the minimum cost. If there is no optimal solution in real scenarios, a suboptimal solution μ' can be chosen as the possible candidate according to the following measurement:

$$\triangle\mu =\mid \mu' - \mu^* \mid \tag{14}$$

If the $\triangle\mu$ is less, the suboptimal solution is closer to the optimal solution. Hence, the greedy selecting strategy could be used to choose the solution which minimizes the (14):

$$\mu_{\text{opt}} = \arg\min \triangle\mu \tag{15}$$

In the current 4G LTE network, the data rate cannot be adjusted smoothly due to the operators' service strategy and technical complexity. There are several bandwidth levels that can be subscribed. For example, the maximum bandwidth is 20 MHz, the second one is 15 MHz, and the minimum bandwidth is 1.4 MHz. In this discrete case, greedy strategy can still be used to select the data rate according to (15).

From a service provider's standpoint, the queuing service rate μ and cloud service time t_s should be adjusted according to the user arriving rate λ, which can be calculated on (11), (12), and (15). The dynamical adjustments include tenanting more cloud resources, applying more data rate from operators. On the other side, operators are expected to provide more elastic mechanisms for users or service providers to configure their communications and cloud resources conveniently when it comes to the promising benefits of Cloud RAN, just as the elastic computing in current SaaS clouds.

7 Simulation study

The simulation study is to evaluate the proposed optimal strategy. Three service rate adjusting strategies are compared. The fix-rate strategy (FR) never changes the service rate since the system initialization. The passive-rate strategy (PR) monitors the job arrival rate periodically and changes the service rate according to formula (8) or (9). The active-rate strategy (AR) also monitors the arrival rate and will increase or decrease the predicated value of arrival rate if the arrival rate keeps increasing or decreasing in successive monitor periods. Let *jobCount* be the variable recording the number of jobs in a monitor period (*MP*), then the estimated arrival rate should be calculated by following formula in PR.

$$\lambda_{\text{PR}} = \frac{\text{jobCount}}{\text{MP}} \tag{16}$$

In AR, the arrival rate is predicated as follows:

$$\lambda_{\mathrm{AR}} = \begin{cases} \dfrac{\text{jobCount}}{\text{MP}} + d & \begin{array}{l}\text{if arrival rate increases}\\ \text{in successive MPs}\end{array} \\ \dfrac{\text{jobCount}}{\text{MP}} - d & \begin{array}{l}\text{if arrival rate decreases}\\ \text{in successive MPs}\end{array} \\ \dfrac{\text{jobCount}}{\text{MP}} & \text{otherwise} \end{cases} \quad (17)$$

where d is an integer, and $d = \frac{\text{the-number-of-successive-MPs}}{2}$. If the arrival rate increases in two successive *MP*s, then d is 1.

7.1 Simulation setup

The simulation was conducted on Omnet++ [30]. The simple queueing model in the Omnet++ library was used and modified for this simulation. The simulation result data was exported and plotted in MATLAB. The simulation setup is as follows:

C_1	0.01
C_2	0.004
t_s	$0.4T$
T	1

Since C_1 is the cost of one unit size data transmitted through a virtual link, it should be the data transmission cost for the usage of mobile wireless networks. The charge per megabyte transmitting of China Mobile is appropriate 0.01 China dollar. C_2 is the cost of one unit size data staying in the system. In this context, it should be the staying cost before the service data is transported. In [31], the per-bit energy consumption of transmission and switching for a private cloud is estimated to be around 0.46 μJ/bit, which is around 3.86 Wh/MB. The industry electric price is around 1.01 China dollars per kilowatt hour in Changsha, hence the cost of per megabyte staying in cloud system is around 0.004 China dollar. The cost of cloud service can be treated as a constant for a particular data service, thus it is omitted in this simulation. What t_s is $0.4T$ means the most time is consumed in data transporting. T is 1 time unit, actually 1 s in this context.

The jobs were generated in a source module and entered into a queue module. The queue module contains the key algorithms for arrival rate evaluation and optimal solution calculation. A sink module is used to count the statistic values. A service broker is composed of the above three modules. There are three service brokers in this simulation which perform FR, PR, and AR, respectively. The arrival rate is configured in two different ways, which will be introduced in following sections. The initial estimated arrival rate of three strategies is 0.5. The service rate of FR is 1. The queue is a first-in-first-out queue and without volume limiting. The time unit is scaled to second for the purpose of accelerating the simulation process. Without loss of generality, we assume that one job is to transmit one unit data.

7.2 Steady arrival rate simulation

In this section, the arrival rate was steadily increased from 0.1 to 10; the step was 0.1. There were 50 jobs initially pushed into each queue. In each step, the source module generated 1000 jobs. The sample interval is 4 s. The service rate is calculated by (12) or (13). The cost is calculated as follows:

$$\text{Cost} = C_1 \times \mu + C_2 \times \text{QueueLength} \quad (18)$$

where the real queue length is used in the cost calculating. The simulation results are depicted in Figs. 3, 4 and 5. The scale of Y axis in Figs. 4 and 5 is adjusted to logarithm for a clear view because the simulation value that came from the FR strategy is much larger than the other two strategies. As shown in these figures, the FR strategy performed badly with the increase of arrival rate. The PR and AR performed well and show a good scalability, which means if the service rate can be adjusted with the change of arrival rate, then the optimal/suboptimal result can be achieved.

7.3 Burst arrival rate simulation

In this section, a burst arrival rate scenario is studied. There have been some researches devoted to characterize the data traffic flow for mobile communication systems [32, 33]. In the most recent work [32] by He et al., The different arrival rates of each hour's Poisson distribution in a day were obtained by fitting. Since the arrival rates are for the base station traffic flow, which should be much larger than a single service provider, we divided them by 10 in this simulation. The arrival rates are listed in Table 1.

Each duration unit is 1 h in a day in real case [32]. Hence, there are 24 arrival rate values in a day. In this simulation, one duration lasted for 100 s. The simulation results are depicted in Figs. 6, 7 and 8. The other setup is the same as the last subsection. The run time of PR and AR is 2400 s. On the other side of the spectrum, the run time of FR was more than 9000 s because the service rate of the FR service broker was too low to finish its jobs in the queue on time. The AR outperformed PR slightly in a few points and they all scaled well with the burst arrival rate. The curve of average cost rose and dropped slower than the queuing time because the predication always falls behind the real changes, that is, the changing of service rates falls behind the real arrival rates. Hence, it will be a better way to setup the service rate on the regress results on long time history data, which will be our future work. The curve trend of queue length is similar as average cost because the cost is directly correlated with queue length.

Table 1 Arrival rates in different durations

Time	λ	Time	λ
0	0.793	12	91.223
1	0.631	13	65.319
2	0.792	14	88.152
3	0.712	15	61.127
4	3.832	16	78.293
5	4.669	17	71.437
6	5.731	18	80.251
7	16.222	19	64.173
8	15.134	20	69.132
9	37.828	21	55.448
10	71.358	22	14.810
11	89.811	23	7.621

8 Discussion on simulation

8.1 Constraint violation

In the simulation, the response time constraint T is set as 1 s (the time resolution in Omnet++ can be adjusted to accelerate the simulation, thus it may not be 1 s in the simulation time). The proposed solution only conforms to this constraint in the statistic concept. The following Table 2 is the average queue time and max queue time in the burst rate simulation.

As presented in Table 2, although the average queue time of PR strategy and AR strategy is less than 1, the maximum queue time is much bigger than 1. It means that whether the QoS is under the constraint or not is uncertain. Thus a further inspection on the queue state should be carried out in running time and multiple-server backup strategy for load balance should be a possible solution. The longtime history data analysis is another possible way to reduce the queuing time by assigning the optimal service rate more accurately.

8.2 Omitted parameters

In the simulation, the price parameters C_1 and C_2, the cost of cloud service C_s are treated as constants. This is not always true in real case. For example, the electricity price may change hourly in a day and be different in different locations [34]. The price for subscribing mobile 3G, 4G, and even the upcoming 5G services also changes with the flow size and service category. The cloud service time is also varied on different service levels. To synthesize all these parameters to choose an optimal or suboptimal solution is a service composition problem [8], which is an NP-hard problem and out of the scope of this paper. Our previous works can be possible solutions [13].

8.3 Other QoS constraints

In service computing, there are always multiple non-functionality attribute constraints, such as power consumption, throughput, availability, and network delay. These constraints will increase the complexity to solve the service providing problem. How to find an optimal solution will be the next step in our work.

8.4 Comparison with related work

Promising virtualization techniques in C-RAN make it possible to jointly allocate the communication resources in the BBU pool to achieve the elastic computing in signal processing cloud. However, there are still several challenges for virtualization of the BBU pool [3–5]. The most important of all are handling delay induced by virtual machines and management of communication procedure in virtualization architecture. The work presented in [3–5] proposes a MCN architecture, which utilizes the software-defined networks (SDNs) and prediction techniques to solve those problems. Bandwidth predication which is used to improve the scalability and response time in [3–5] is put forward with flow-based approach and dimensioning approach. It is a large-scale optimization approach for virtualized BBU pool. Assumptions that the traffic flow obeys normal distribution and traffic model would not change for long periods are coarse granular. The coarse grain approach can hardly optimize a specific service procedure. The strategy proposed in this paper is a fine grain approach, which aims to optimize the data rate of data service by monitoring the job queue. Since service providers have enough knowledge about their own services, the optimization could be more delicate on the basis of the proposed approach if the providers could dynamically adjust their service parameters with supports of communication operators.

In mobile cloud computing environment, the cost can be split into two main parts, that is, the cost in data center and the cost in mobile communications. The research on cost saving for data centers is emerging in recent years [34–36]. The optimal strategies can be classified to location-based strategies and time-based strategies, which are on the basis of the observation that the electricity price varied from time to time, from location to location. On the other hand, character analysis of mobile network traffic has been carried out for more than a decade, which is beneficial to the network design, traffic modeling, resource planning, and network control [32, 33, 37]. The C-RAN can be treated as a kind of mobile resource planning technology. Nevertheless, all these contributions are devoted to improve the operators' efficiency, no matter who are cloud operators or mobile communications operators. Thus, these achievements cannot be directly applied in utility optimization for service providers. Some new methods or strategies should be put forward to optimize the service providers' efficiency in mobile cloud computing, as what is presented in this paper.

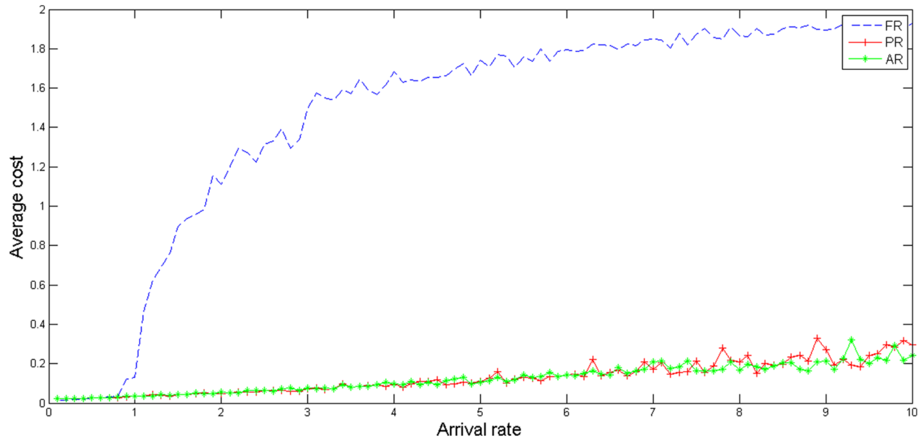

Fig. 3 Average cost in different arrival rates

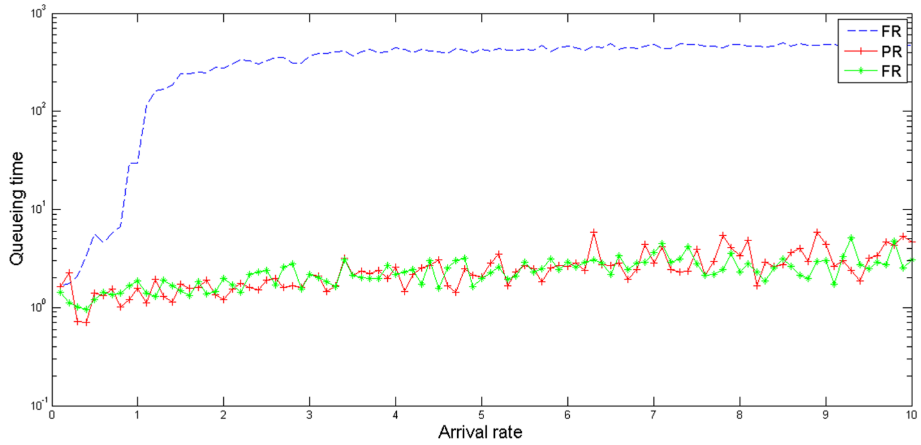

Fig. 4 Average queueing time in different arrival rates

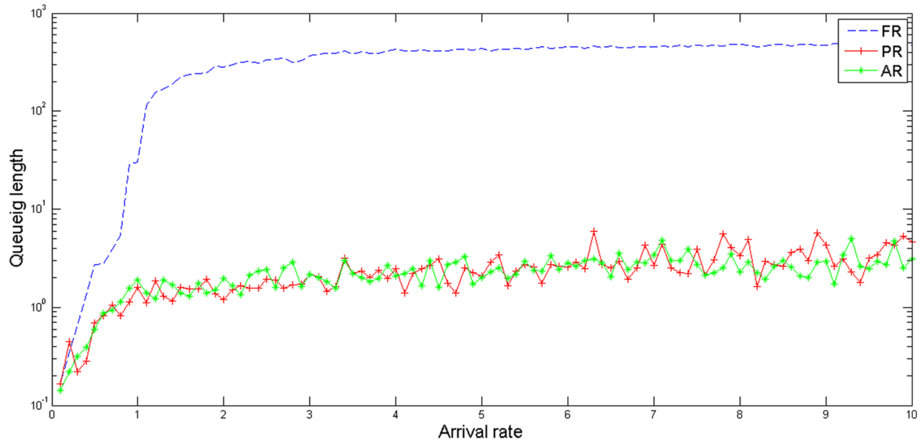

Fig. 5 Average queueing length in different arrival rates

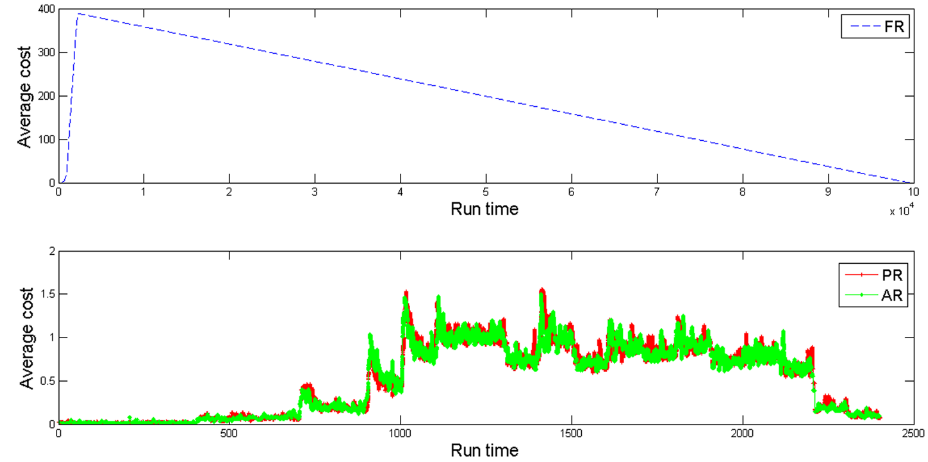

Fig. 6 Average cost in burst arrival rate

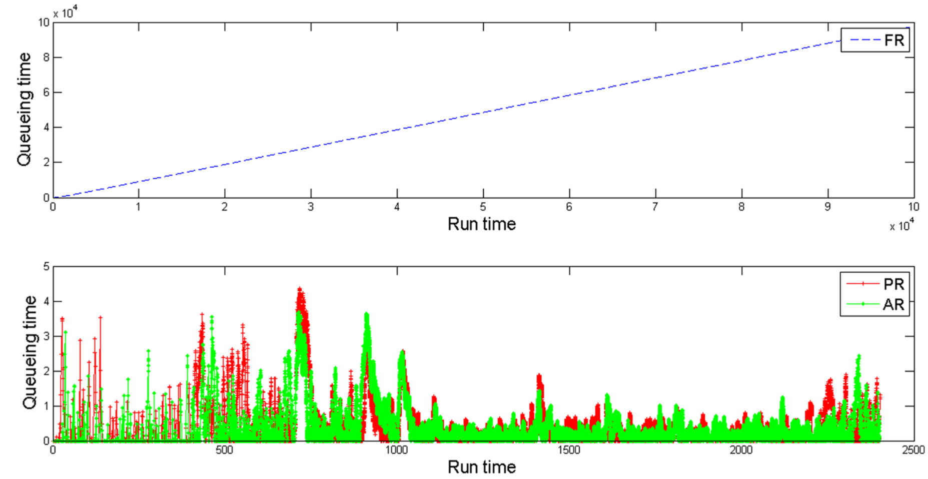

Fig. 7 Average queueing time in burst arrival rate

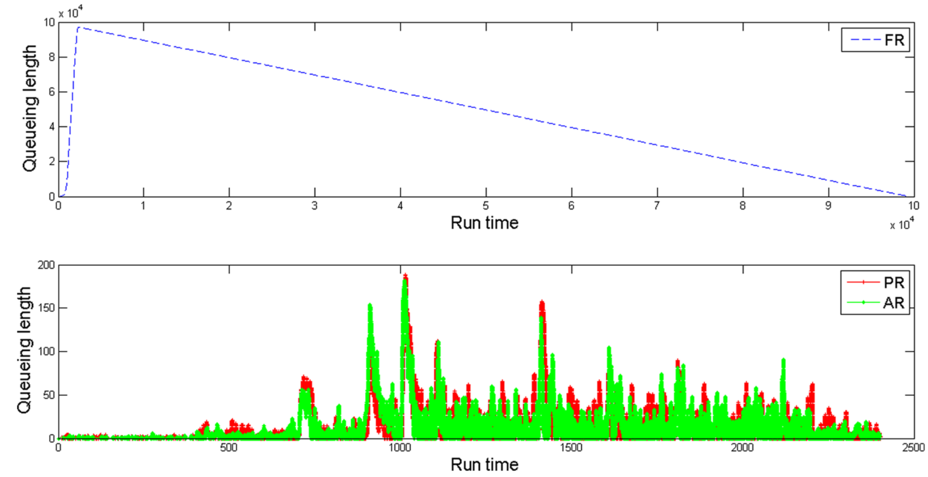

Fig. 8 Average queueing length in burst arrival rate

Table 2 Average queueing time and maximum queueing time in burst rate simulation

	Average queueing time (s)	Maximum queueing time (s)
FR	48354	97063.5
PR	0.311251	4.36831
AR	0.300435	3.67761

9 Conclusions

In this paper, a device-to-device data service providing framework in cloud radio accessing networks is proposed. The framework is based on the emerging technique, C-RAN. The queuing theory and convex optimization are used to deduce the optimal solution in this framework theoretically. Simulation is conducted on Omnet++ to evaluate the efficiency of the optimal strategy. Part of this work is presented in our previous work [38]. Our future work includes studying the relationship among the parameters of data centers, wireless communication channel, and energy/power consumption in mobile cloud and smart grid scenarios.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (Grant Nos. 61303043, 61502056, and 61572389), Hunan Provincial Natural Science Foundation of China, China (Grant Nos. 13JJ4052 and 2015JJ3010), Scientific Research Fund of Hunan Provincial Education Department (Grant Nos. 13C1022, 13C1023, and 15B009), Shenzhen S&T Innovation Project (No. JCYJ20140610151856732), and UK EPSRC Project NIRVANA (EP/L026031/1). EU H2020 Project iCIRRUS (GA-644526), Key projects of Hunan Provincial Education Department(Grant No.14A004), the Scientific Research Fund of Hunan Provincial Transportation Department (Grant No. 201446). The authors would like to appreciate the anonymous reviewers for their constructive comments.

Author details

[1]Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha, Hunan, People's Republic of China. [2]School of Computer and Communications Engineering, Changsha University of Science and Technology, Changsha, Hunan, People's Republic of China. [3]School of Computer Science and Electronic Engineering, University of Essex, Colchester Essex, UK. [4]Department of Economy and Administration, Changsha University, Changsha, Hunan, People's Republic of China. [5]School of Business, Central South University, Changsha, Hunan, People's Republic of China.

Received: 1 August 2015 Accepted: 20 December 2015
Published online: 19 January 2016

References

1. LJ Zhang, EIC Editorial: introduction to the body of knowledge areas of services computing. IEEE Trans. Services Comput. **1**(2), 62–75 (2008)
2. MP Papazoglou, P Traverso, S Dustdar, F Leymann, Service-oriented computing: state of the art and research challenges. IEEE Comput. **40**(11), 64–71 (2007)
3. A Pras, A Russu, A Pizzinat, A Gomes, C Marques, C Parada, D Vingarzan, E Schiller, I Aad, I Alyafawi, L Suciu, LS Ferreira, LM Correia, L Hendriks, MR Sama, M Santos, M Liebsch, M Corici, M Karimzadeh, O Keil, RdO Schmidt, S Ruffino, S Khatibi, T Taleb, T Braun, V Vlad, Z Yousaf, Z Zhao, MCN D4.3, algorithms and mechanisms for the mobile network cloud. European Commission, EU FP7 Mobile Cloud Networking public deliverable (2014). available: http://www.mobile-cloud-networking.eu/site/. Access date: August, 2015
4. K Morteza, Z Zhao, L Hendriks, RdO Schmidt, S la Fleur, H van den Berg, A Pras, T Braun, MJ Corici, Mobility and bandwidth prediction as a Service in virtualized LTE systems. Cloud Netw. (CloudNet), 2015 IEEE 4th Int. Conf. IEEE, Niagara Falls, ON, 132–138 (2015)
5. K Georgios, A Jamakovic, K Briggs, M Karimzadeh, C Parada, MI Corici, T Taleb, A Edmonds, TM Bohnert, Mobility and bandwidth prediction in virtualized LTE systems: architecture and challenges. Netw. Commun. (EuCNC), 2014 European Conf. IEEE, Bologna, 1–5 (2014)
6. M Alasti, B Neekzad, J Hui, R Vannithamby, Quality of service in WiMAX and LTE networks. [Topics in Wireless Communications]. Commun. Mag. IEEE. **48**(5), 104–111 (2010)
7. H Ludwig, A Keller, A Dan, RP King, R Franck, Web Service Level Agreement (WSLA) Language Specification. IBM Corporation, 815–824 (2003)
8. J Brnsted, KM Hansen, M Ingstrup, Service composition issues in pervasive computing. Pervasive Comput. **9**(1), 62–70 (2010)
9. L Zeng, B Benatallah, A Ngu, M Dumas, J Kalagnanam, H Chang, QoS-aware middleware for web service composition. IEEE Trans. Software Eng. **5**(30), 311–328 (2004)
10. L Yu, W Zhili, L Meng, Q Xue-song, Towards multi-user and network-aware Web services composition. Web Services (ICWS), 2013 IEEE 20th Int. Conf. IEEE, Santa Clara, CA, 607–608 (2013)
11. Q He, J Han, Y Yang, J Grundy, H Jin, in *2012 IEEE 5th international conference on*. QoS-driven service selection for multi-tenant SaaS. Cloud computing, (CLOUD) (IEEE, Honolulu, HI, United states, 2012), pp. 566–573
12. A Klein, F Ishikawa, S Honiden, in *Proceedings of the 21st international conference on World Wide Web*. Towards network-aware service composition in the cloud (ACM, Lyon, France, 2012), pp. 959–968
13. Y Luo, K Yang, Q Tang, J Zhang, B Xiong, A multi-criteria network-aware service composition algorithm in wireless environments. Comput. Commun. **35**, 1882–1892 (2012)
14. J Brnsted, KM Hansen, M Ingstrup, Service composition issues in pervasive computing. Pervasive Comput. **9**(1), 62–70 (2010)
15. M Chen, T Kwon, Y Choi, Energy-efficient differentiated directed diffusion for real-time traffic in wireless sensor networks. Comp. Commun. **29**(2), 231–245 (2006)
16. K Doppler, M Rinne, C Wijting, CB Ribeiro, K Hugl, Device-to-device communication as an underlay to LTE-advanced networks. Commun Mag. IEEE. **47**(12), 42–49 (2009)
17. A Sergey, A Pyattaev, K Johnsson, O Galinina, Y Koucheryavy, Cellular traffic offloading onto network-assisted device-to-device connections. Commun. Mag. IEEE. **52**(4), 20–31 (2014)
18. K-C Chen, S-Y Lien, Machine-to-machine communications: technologies and challenges. Ad Hoc Netw. **18**, 3–23 (2014)
19. C Magurawalage, M Sarathchandra, K Yang, L Hu, J Zhang, Energy-efficient and network-aware offloading algorithm for mobile cloud computing. Comput. Netw. **74**, 22–33 (2014)
20. Y Cai, FR Yu, S Bu, Cloud radio access networks (C-RAN) in mobile cloud computing systems. Comput. Commun. Workshops (INFOCOM WKSHPS), 2014 IEEE Conf. IEEE, Toronto, ON, 369–374 (2014)
21. K Sundaresan, MY Arslan, S Singh, S Rangarajan, SV Krishnamurthy, in *Proceedings of the 19th annual international conference on Mobile computing & networking*. FluidNet: a flexible cloud-based radio access network for small cells (ACM, New York, USA, 2013), pp. 99–110
22. M Peng, K Zhang, J Jiang, J Wang, W Wang, Resource allocation optimization for delay-sensitive traffic in fronthaul constrained cloud radio access networks. Trans. Veh. Tech, 1–12 (2014)
23. L Liang, B Suzhi, Z Rui, Joint power control and fronthaul rate allocation for throughput maximization in OFDMA-based cloud radio access network. Commun. IEEE Trans. **63**(11), 4097–4110 (2015)
24. CMR Institute, C-ran white paper: the road towards green ran. White Paper, ver, 2 (2011). Available: http://labs.chinamobile.com/report/59826. Access date: August, 2015
25. T Yu, Y Zhang, KJ Lin, Efficient algorithms for web services selection with end-to-end QoS constraints. ACM Trans. Web. **1**(1), 111–136 (2007)
26. X Wang, J Zhu, Y Shen, Network-aware QoS prediction for Service Composition Using Geolocation. IEEE Transactions on Services Computing, IEEE. **8**(4), 630–643 (2015)

27. Al Ridhawi, IA Yousif, K Ahmed. Service specific overlay composition and reconfiguration (IEEE, Tangier, 2012), pp. 479-484
28. D Yu, Y Lei, Z Bin, QoS-driven self-healing web service composition based on performance prediction. J Comput Science Technol. **24**(2), 250–261 (2009)
29. S Boyd, V Lieven, *Convex optimization*. (Cambridge University Press, New York, USA, 2004)
30. Omnet++ user manual, András Varga and OpenSim Ltd, 2014, https://omnetpp.org/doc/omnetpp/manual/usman.html. Access date: August, 2015
31. J Baliga, RW Ayre, K Hinton, RS Tucker, Green cloud computing: balancing energy in processing, storage, and transport. Proc. IEEE. **99**(1), 149–167 (2011)
32. QQ He, CY Wan, XH Yan, Accurate method to estimate EM radiation from a GSM base station. Progress In Electromagnetics Res. M. **34**, 19–27 (2014)
33. Y Zhang, A Ake, Understanding the characteristics of cellular data traffic. ACM SIGCOMM Comput. Commun. Rev. **42**(4), 461–466 (2012)
34. P Wang, L Rao, X Liu, Y Qi, Dynamic power management of distributed Internet data centers in smart grid environment. Global Telecommun. Conf. (GLOBECOM 2011), IEEE, Houston, TX, USA, 1-5 (2011)
35. A Qureshi, R Weber, H Balakrishnan, J Guttag, B Maggs, Cutting the electric bill for internet-scale systems. ACM SIGCOMM Comput Commun. Rev. ACM. **39**(4), 123–134 (2009)
36. P Wang, L Rao, X Liu, Y Qi, D-pro: dynamic data center operations with demand-responsive electricity prices in smart grid. Smart Grid, IEEE Trans. **3**(4), 1743–1754 (2012)
37. XG Meng, SH Wong, Y Yuan, S Lu, in *Proceedings of the 10th annual international conference on Mobile computing and networking*. Characterizing flows in large wireless data networks (ACM, New York, USA, 2004), pp. 174-186
38. YS Luo, K Yang, Q Tang, J Zhang, S Qiu, in *29th International Conference on Advanced Information Networking and Applications Workshops*. Device-to-device service selection framework in cloud radio access network (IEEE, Gwangju, Korea, March 24–27, 2015), pp. 633–637