doi:10.3772/j.issn.1006-6748.2021.02.002

Resource provisioning for computation and communication in multi-cell wireless networks①

Yang Xiumei(杨秀梅)[*,**], Chen Huaxia②[*], Zhang Mengying[*,**]
([*] Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, P. R. China)
([**] Key Laboratory of Wireless Sensor Network and Communications, Chinese Academy of Sciences, Shanghai 200050, P. R. China)

Abstract

The convergence of computation and communication at network edges plays a significant role in coping with computation-intensive and delay-critical tasks. During the stage of network planning, the resource provisioning problem for edge nodes has to be investigated to provide prior information for future system configurations. This work focuses on how to quantify the computation capabilities of access points at network edges when provisioning resources of computation and communication in multi-cell wireless networks. The problem is formulated as a discrete and non-convex minimization problem, where practical constraints including delay requirements, the inter-cell interference, and resource allocation strategies are considered. An iterative algorithm is also developed based on decomposition theory and fractional programming to solve this problem. The analysis shows that the necessary computation capability needed for certain delay guarantee depends on resource allocation strategies for delay-critical tasks. For delay-tolerant tasks, it can be approximately estimated by a derived lower bound which ignores the scheduling strategy. The efficiency of the proposed algorithm is demonstrated using numerical results.

Key words: resource provisioning, computation and communication, multi-cell wireless network, network edge

0 Introduction

Conventional wireless networks are facing dramatic challenges with the explosive growth of connected devices and emerging applications in the era of Internet of Things (IoT). Massive data are generated at network edges and need local processing for ultra-low latency. However, under existing frameworks, a large amount of redundant data have to be uploaded to remote cloud centers for central processing. Such paradigms not only consume large bandwidth but also bring unexpected processing delay. Various techniques have risen to cope with those challenges, such as mobile edge computing[1], fog computing[2] and fog-radio access network[3]. They have much in common in pushing communication, computing, control and storage to network edges.

This paradigm shift requires access points (APs) at network edges to perform computation-intensive and latency-critical applications. This is beneficial, for example, in the cyber-physical control of emergency, where data processing at the nearest AP allows the fastest response. Another example is to use local image/video processing to extract information to upload in lieu of unnecessary redundant data for bandwidth savings.

To achieve these, APs should have computation capabilities besides being traditional wireless transceivers. The resource scheduling problem on how to efficiently utilize the joint communication and computational resources has been widely investigated. To name a few, scheduling strategies for task offloading are proposed in terms of energy efficiency, task delay, or multi-criterion metrics under the cloud-fog-thing network architecture[4-6]. Various optimization methods have also been developed to obtain resource allocation solutions[7].

The resource provisioning problem in deploying a computing wireless network, however, still remains a challenging field of research[1]. One open question is how to quantify computation capabilities of APs in practical wireless networks. Questions also include where to place serving nodes and how many nodes are optimal in terms of the deployment cost. Those ques-

① Supported by the Shanghai Sailing Program (No. 18YF1427900), the National Natural Science Foundation of China (No. 61471347) and the Shanghai Pujiang Program (No. 2020PJD081).

② To whom correspondence should be addressed. E-mail: chenhuaxia@mail.sim.ac.cn

Received on May 5, 2020

tions should be answered when upgrading existing APs with co-located computing servers, or deploying new APs with computation capabilities for emerging applications such as smart factories.

Several solutions for the resource provisioning problem at network edges have been recently addressed in Refs [8-10]. Ref. [8] studied the site selection problem for fog nodes in an IoT-based logistic center, assumed the link between the fog node and the edge device was wired. The goal of Ref. [9] was to determine the location placement of serving nodes in wireless metropolitan area networks, but its heuristic algorithms ignore the impact of the underlying resource scheduling and the inter-cell interference. In Ref. [10], the demand for AP's communication-and-computation capability was analyzed in a single-cell multi-user scenario. In existing studies, the network assumption is simplified to be either wired or limited to a single-cell scenario, which weakens the dynamics of practical wireless networks.

In this work, the computation capabilities of APs are optimized under practical constraints in multi-cell wireless networks. The purpose is to provide useful guidance for wireless operators during network deployment. The main contributions are summarized as follows.

(1) A discrete and non-convex optimization problem is formulated to quantify the necessary computation capabilities of APs under heterogeneous delay requirements, discrete bandwidth allocation and dynamic inter-cell interference.

(2) The solution is developed based on decomposition theory and fractional programming (FP). The optimization problem is firstly decoupled into a computation subproblem and a communication subproblem. Then, the first subproblem is solved with a closed-form expression. The second one is solved using fractional programming.

(3) Numerical simulations in a multi-cell scenario are performed to evaluate the proposed algorithm. Practical factors are considered in simulation settings for insightful analysis. Simulation results demonstrate the efficiency of the solution.

The rest of this work is organized as follows. The system model and the problem formulation are presented in Section 1. The solution is developed in Section 2. Specifically, this section introduces the decomposition of the original problem, a closed-form solution for the computation subproblem, and an iterative update framework. The detailed algorithm for the communication subproblem is separately presented in Section 3. Simulation results are presented in Section 4. Concluding remarks are given in Section 5.

1 System model and problem formulation

A multi-cell wireless network is considered. Let $I = \{1, 2, \cdots, I\}$ and $J = \{1, 2, \cdots, J\}$ denote the sets of indices for APs and devices, respectively. AP i is capable to communicate with a group of devices within its coverage whose indexes are from the set J_i ($J_i \subset J$).

1.1 Computation model

The computation capability is measured by the number of cycles of the central processing unit (CPU) per second. Let F_i denote the computation capability of AP i. Each AP is assumed to be configured with a certain level of computation capability in order to process tasks from its serving devices, i.e.:

$$F_i > 0, \ \forall i \in I \tag{1}$$

further, the fraction of computational resources allocated to device j is denoted as β_j. Those devices without any computational requirements are simply omitted, so that β_j satisfies the following conditions.

$$\sum_{j \in J_i} \beta_j \leqslant 1, \ \beta_j \in (0, 1], \ \forall j \in J_i, \ \forall i \in I \tag{2}$$

Therefore, the amount of computational resources allocated to device $j \in J_i$ is $\beta_j F_i$.

The task is measured by the data size in bits. The computation task is assumed to be fully offloaded to AP. Let c_j denote the number of CPU cycles needed to process one single bit of device j's task. Then, the calculation time of a given task with l_j bits is $\frac{l_j c_j}{\beta_j F_i}$ seconds for each device $j \in J_i$.

1.2 Communication model

The access mode is frequency division multiple access for multiple devices served by one AP. A set $N = \{1, 2, \cdots, N\}$ of frequency subcarriers is considered. $\alpha_j^n \in \{0, 1\}$ is used to represent the subcarrier allocation indicator. Subcarrier $n \in N$ is allocated to device j if and only if $\alpha_j^n = 1$. Orthogonal subcarrier assignments are further assumed during one scheduling interval, i.e.:

$$\sum_{j \in J_i} \alpha_j^n \leqslant 1, \ \alpha_j^n \in \{0, 1\}, \ \forall j \in J_i, \ \forall i \in I \tag{3}$$

Full frequency reuse is adopted so that the inter-cell interference can not be ignored.

The device is associated to its serving AP with the strongest received signal. Let p_j^n denote the uplink transmission power at subcarrier n from device j to its serving AP. The total power of device j should be no more than the maximum uplink transmission power

$P_{\max}$, i. e. :

$$\sum_{n\in N} p_j^n \leqslant P_{\max},\ p_j^n \geqslant 0,\ \forall n \in N_i,\ \forall j \in J \tag{4}$$

Frequency selective channel is assumed due to wide-band communication. Distance-dependent path loss and shadow fading are also considered as fading components. h_{ij}^n is denoted as the uplink channel gain from device j to AP i at subcarrier n.

The communication and computation stages in processing one task is assumed to be sequential. The communication time for downlink transmission is also neglected due to the negligible amount of feedback bits. Let r_j denote the uplink transmission rate of device j associated with AP i, where

$$r_j = \sum_{n\in N} \alpha_j^n \log_2\left(1 + \frac{h_{ij}^n p_j^n}{\sigma_i^2 + \sum_{j'\in J/J_i} h_{ij'}^n p_{j'}^n}\right),\quad \forall j \in J_i,\ \forall i \in I \tag{5}$$

Then, the communication time is approximately l_j/r_j seconds. The total processing time t_j satisfies:

$$t_j = \frac{l_j}{r_j} + \frac{l_j c_j}{\beta_j F_i} \leqslant d_j,\ \forall j \in J_i,\ \forall i \in I \tag{6}$$

where d_j constrains the delay to process l_i bits.

1.3 Problem formulation

The objective of this work is to understand how many computational resources are necessary in order to satisfy the tasks' requirements in multi-cell wireless networks. Therefore, the optimization problem can be expressed as the minimization of the total computation capabilities of all APs in the network. Specifically, the problem is formulated as

$$\text{P0:}\quad \underset{\boldsymbol{F},\boldsymbol{p},\boldsymbol{\alpha},\boldsymbol{\beta}}{\text{minimize}} \sum_{i\in I} F_i \tag{7}$$
$$\text{subject to Eqs}(1)-(6)$$

where the collections of variables to be optimized are denoted as $\boldsymbol{F} = (F_i)_{i\in I}$, $\boldsymbol{p} = (p_j^n)_{j\in J,\, n\in N}$, $\boldsymbol{\alpha} = (\alpha_j^n)_{j\in J,\, n\in N}$, and $\boldsymbol{\beta} = (\beta_j)_{j\in J}$ respectively.

P0 is a challenging combinatorial and non-convex optimization problem. The formulation incorporates discrete subcarrier assignment, where subcarriers allocated by one AP are coupled with those co-channeled ones assigned by neighboring APs. It is well known that a degenerated communication problem of P0 to solely maximize the sum rate with Eq. (5) has been rather complicated[11]. It becomes more challenging when further introducing the computation problem.

2 Decomposition based solution framework

2.1 Framework overview

An overview of the procedures is firstly presented to solve the problem P0. As shown in Fig. 1, the problem P0 is solved in an iterative manner, which includes several key stages, such as problem decomposition, subproblem solving, and variable updating. Specifically, the problem P0 is firstly decomposed to the computation subproblem P1 and the communication subproblem P2. Then, the two subproblems are separately solved by its corresponding algorithms after initializing the auxiliary variable z_j. Iterations are performed by additional variable updates until the convergence condition is satisfied.

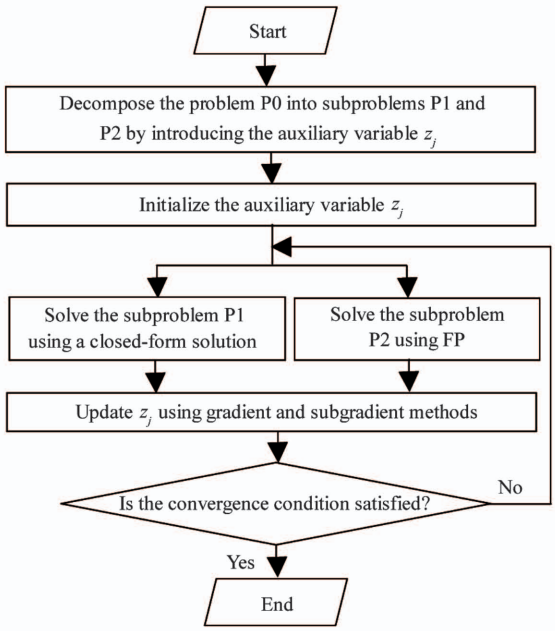


Fig. 1 Flow chart of the solution framework

2.2 Decomposition

By grouping computation variables and communication variables as $x_1 = [\boldsymbol{F}, \boldsymbol{\beta}]$ and $x_2 = [\boldsymbol{p}, \boldsymbol{\alpha}]$, P0 falls into the following structure [12],

$$\begin{aligned}&\underset{x_1,\, x_2}{\text{minimize}} \quad v_1(\boldsymbol{x}_1) + v_2(\boldsymbol{x}_2)\\ &\text{subject to} \quad u_1(\boldsymbol{x}_1) + u(\boldsymbol{x}_2) \leqslant m,\\ &\qquad\qquad \boldsymbol{x}_1 \in \boldsymbol{A}_1,\ \boldsymbol{x}_2 \in \boldsymbol{A}_2\end{aligned} \tag{8}$$

where both $v_k(\boldsymbol{x}_k)$ and $u_k(\boldsymbol{x}_k)$ are functions of $\boldsymbol{x}_k$ in its feasible region $\boldsymbol{A}_k$ for $k = 1, 2$; m is a constant. Since primal decomposition is appropriate for Eq. (8), $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be decoupled after decomposing Eq. (6) into

$$z_j \leqslant r_j \tag{9}$$

$$\frac{1}{\beta_j F_i / c_j} \leqslant \frac{d_j}{l_j} - \frac{1}{z_j} \tag{10}$$

where, the auxiliary variable z_j can be considered as a rate threshold according to Eq. (9). Detailed decomposition of P0 to two subproblems are as follows.

2.3 Computation subproblem

The objective here is to optimize computation capabilities $\boldsymbol{F}$ and allocation coefficients $\boldsymbol{\beta}$ under fixed $\boldsymbol{z} = (z_j)_{j\in\boldsymbol{J}}$. The optimization problem is

$$\text{P1}: \underset{\boldsymbol{F},\boldsymbol{\beta}}{\text{minimize}} \sum_{i\in\boldsymbol{I}} F_i \tag{11}$$
subject to Eqs(1), (2) and (10).

A closed-form solution for P1 is

$$F_i = \sum_{j\in\boldsymbol{J}_i} \frac{c_j}{\frac{d_j}{l_j} - \frac{1}{z_j}} \quad \forall i \in I \tag{12}$$

$$\beta_j = \frac{\frac{1}{\frac{d_j}{l_j} - \frac{1}{z_j}}}{\sum_{j\in\boldsymbol{J}_i} \frac{1}{\frac{d_j}{l_j} - \frac{1}{z_j}}}, \quad \forall j \in J_i, \ \forall i \in I \tag{13}$$

Proof From Eq. (10), for each $j \in J_i$, it has

$$\frac{c_j}{F_i\left(\frac{d_j}{l_j} - \frac{1}{z_j}\right)} \leqslant \beta_j \tag{14}$$

Combining Eq. (2) and Eq. (14), it obtains:

$$\frac{1}{F_i}\sum_{j\in\boldsymbol{J}_i} \frac{c_j}{\frac{d_j}{l_j} - \frac{1}{z_j}} \leqslant \sum_{j\in\boldsymbol{J}_i}\beta_j \leqslant 1 \tag{15}$$

From Eq. (15), the minimum value of F_i is obtained as shown in Eq. (12).

Evidently Eq. (10) should be satisfied with equality. Therefore, Eq. (13) can be directly derived by combining Eqs(10) and (12).

Proposition 1 F_i in Eq. (12) approaches to a lower bound, i. e., $F_i^* = \sum_{j\in J_i}(l_j c_j/d_j)$, for delay-tolerant tasks with $d_j >> l_j/z_j$.

The proof of Proposition 1 is obvious according to Eq. (12) so that it is omitted here. Further, the lower bound is no longer dependent on the rate-related parameter z_j, so it presents a computationally efficient approach to estimate AP's computation capability under weak delay demands.

2.4 Communication subproblem

The objective here is to optimize power allocation variables $\boldsymbol{p}$ and frequency allocation indicators $\boldsymbol{\alpha}$ under fixed $\boldsymbol{z}$. This is in fact the following feasibility problem.

$$\text{P2}: r_j \geqslant z_j, \ j\in J \tag{16}$$
subject to Eqs(3), (4) and (5).

For fixed $\boldsymbol{z}$, any feasible $(\boldsymbol{p}, \boldsymbol{\alpha})$ satisfying Eq. (16) is a solution.

P2 is solved in the dual domain:

$$\underset{\boldsymbol{\rho},\boldsymbol{\mu}}{\text{minimize}} \ g(\boldsymbol{\rho}, \boldsymbol{\mu}) \tag{17}$$
$$\text{subject to } \rho_j \geqslant 0, \ \forall j \in J, \quad \mu_j \geqslant 0, \ \forall j \in J.$$

where,

$$g(\rho, \mu) = \underset{\boldsymbol{p},\boldsymbol{\alpha}}{\text{maxmize}} \ L(\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\mu}) \tag{18}$$

is the dual function, and

$$L(\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\mu}) = \sum_{j\in J}\sum_{n\in N}(\rho_j r_j^n - \mu_j p_j^n) + \sum_{j\in J}(-\rho_j z_j + \mu_j P_{\max}) \tag{19}$$

is the Lagrangian function. Here, $\boldsymbol{\rho} = (\rho_j)_{j\in\boldsymbol{J}}$ and $\boldsymbol{\mu} = (\mu_j)_{j\in\boldsymbol{J}}$ are Lagrange multipliers for rate and power constraints in Eq. (16) and Eq. (4), respectively; $r_j^n (j \in \boldsymbol{J}_i)$ is the rate at subcarrier n, where

$$r_j^n = \alpha_j^n \log_2\left(1 + \frac{h_{ij}^n p_j^n}{\sigma_i^2 + \sum_{j'\in J/J_i} h_{ij'}^n p_{j'}^n}\right) \tag{20}$$

Then r_j can be simply expressed as $r_j = \sum_{n\in\boldsymbol{N}} r_j^n$. The dual problem in Eq. (17) is solved by using Algorithm 2 in Section 3.

2.5 Iterative algorithm

The auxiliary variable z_j is iteratively updated according to the gradient and subgradient methods.

$$z_j \leftarrow z_j - \delta\left(\sum_{i\in\boldsymbol{I}}\omega(z_j) + \varphi(z_j)\right), \ j \in J \tag{21}$$

where δ is the positive step-size; the gradient $\omega(z_j)$ from P1 and the subgradient $\varphi(z_j)$ from P2 are as follows.

$$\omega(z_j) = \frac{\partial F_i}{\partial z_j} = -\frac{c_j}{(d_j z_j/l_j - 1)^2}$$

$$\varphi(z_j) = z_j - r_j^*$$

r_j^* is computed by Eq. (5) using the solution of P2.

To sum up, the solution of P0 can be achieved by solving two subproblems P1 and P2 iteratively and updating z_j using Eq. (21), as shown in Algorithm 1.

Algorithm 1 Solution for P0

Input: configuration parameters $P_{\max}$, c_j, l_j, d_j, and channel gain h_{ij}^n, $\forall n \in \boldsymbol{N}$, $\forall j \in \boldsymbol{J}_i$, $\forall i \in \boldsymbol{I}$;

Output: $\boldsymbol{F}$, $\boldsymbol{p}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$;

Initialization: $z_j = \frac{l_i}{d_j} + \varepsilon$, $\forall j \in \boldsymbol{J}$;

Repeat

1) update $\boldsymbol{p}$, $\boldsymbol{\alpha}$ by calling Algorithm 2;

2) update $\boldsymbol{F}$, $\boldsymbol{\beta}$ by Eq. (12) and Eq. (13);

3) update $\boldsymbol{z}$ by Eq. (21);

until convergence

* ε is a very small positive value to guarantee an initial feasible solution of P2.

3 Fractional programming solution for P2

In this section, the dual problem of P2 is solved through FP. Firstly, it is shown that the dual problem has a similar FP structure as Ref. [11]. Then, the baseline FP in Ref. [11] is extended to a general multi-subcarrier FP with Gauss-Seidel iterations. Finally, the solution is presented in Algorithm 2.

Algorithm 2 Solution for P2

Input: configuration parameters $P_{\max}$, f_j, l_j, d_j, and channel gain h_{ij}^n, $\forall n \in \boldsymbol{N}$, $\forall j \in \boldsymbol{J}_i$, $\forall i \in \boldsymbol{I}$;

Output: $\boldsymbol{p}$, $\boldsymbol{\alpha}$, $\boldsymbol{\rho}$, $\boldsymbol{\mu}$;

Initialization: set $(\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\mu})$ to feasible values;

Repeat

  Repeat for fixed $(\boldsymbol{\rho}, \boldsymbol{\mu})$

    1) update $\boldsymbol{\gamma}$, $\boldsymbol{y}$, $\boldsymbol{p}$, $\boldsymbol{Q}$, $\boldsymbol{\alpha}$ by Eq. (28);

  until convergence

  2) update $\boldsymbol{\rho}$, $\boldsymbol{\mu}$ by Eq. (29);

until convergence

In the following, the equivalent form of the dual function in Eq. (18) is discussed. For fixed $\boldsymbol{\rho}$, $\boldsymbol{\mu}$ and z, Eq. (18) is in fact

$$\underset{\boldsymbol{p},\,\boldsymbol{\alpha}}{\text{maximize}} \quad \sum_{j\in J}\sum_{n\in N}(\rho_j r_j^n - \mu_j p_j^n) \tag{22}$$

where r_j^n has the same FP structure as Ref. [11]. Therefore, Eq. (22) can be solved by referring to the baseline FP algorithm based on the quadratic transform and Lagrangian dual transform (Theorem 1 and 3 in Ref. [11]). The difference is that an additional linear term is introduced, i. e., $\mu_j p_j^n$ in Eq. (22). Following the same line for the proof of Theorem 1 and 3 in Ref. [11], it is sure that this linear term does not change the properties stated therein (the details are omitted here). As a result, Eq. (22) can be reformulated to a parallel optimization problem for each subcarrier $n \in N$.

$$\underset{\boldsymbol{p},\,\boldsymbol{\alpha}}{\text{maximize}} \quad \sum_{i\in I}\sum_{j\in J_i} Q_{ij}^n(\alpha_j^n, p_j^n) \tag{23}$$

where,

$$\begin{aligned} Q_{ij}^n(\cdot) = {} & \rho_j \log_2(1+\gamma_j^n) - \rho_j\gamma_j^n - (y_j^n)^2\sigma_i^2 \\ & - \mu_j p_j^n + 2y_j^n\sqrt{\rho_j(1+\gamma_j^n)h_{ij}^n p_j^n} \\ & - \sum_{i'\in \boldsymbol{I}}\sum_{j'\in \boldsymbol{J}_{i'}}(y_{j'}^n)^2 h_{i'j}^n p_j^n \end{aligned} \tag{24}$$

Accordingly, a new term, i. e., $\mu_j p_j^n$, is included in Eq. (24) compared to the baseline FP in Ref. [11]. For each $j \in J_i$, γ_j^n and y_j^n are auxiliary variables defined as follows, respectively.

$$\gamma_j^n = \frac{h_{ij}^n p_j^n}{\sigma_i^2 + \sum_{j'\in \boldsymbol{J}/\boldsymbol{J}_i} h_{ij'}^n p_{j'}^n} \tag{25a}$$

$$y_j^n = \frac{\sqrt{\rho_j(1+\gamma_j^n)h_{ij}^n p_j^n}}{\sigma_i^2 + \sum_{j'\in \boldsymbol{J}} h_{ij'}^n p_{j'}^n} \tag{25b}$$

The approach to solve Eq. (23) is presented as follows. Suppose subcarrier n is assigned to task j by AP i. The optimal p_j^n that maximizes the object of the max operation in Eq. (23) can be obtained by letting $\frac{\partial Q_{ij}^n}{\partial p_j^n} = 0$. Therefore, p_j^n can be obtained as

$$p_j^n = \min\left\{\frac{\rho_j(y_j^n)^2(1+\gamma_j^n)h_{ij}^n}{(\mu_j + \sum_{i'\in \boldsymbol{I}}\sum_{j'\in \boldsymbol{J}_{i'}}(y_{j'}^n)^2 h_{i'j}^n)^2},\ P_{\max}\right\} \tag{26}$$

By substituting Eq. (26) into Eq. (24) and comparing with all the possible assignments of this subcarrier, α_j^n can be obtained as

$$\alpha_j^n = \begin{cases} 0, & \text{if } \max\limits_{j\in \boldsymbol{J}_i} Q_{ij}^n \leqslant 0, \text{ or } j \neq \arg\max\limits_{j\in \boldsymbol{J}_i} Q_{ij}^n \\ 1, & \text{for } j = \arg\max\limits_{j\in \boldsymbol{J}_i} Q_{ij}^n, \text{ otherwise} \end{cases} \tag{27}$$

So far, the solution of $(p_j^n, \alpha_j^n)_{j\in \boldsymbol{J},\, n\in \boldsymbol{N}}$ can be obtained by iteratively updating Eqs (24) – (27) among neighboring cells for a multi-subcarrier system. In the proposed algorithm, Gauss-Seidel iterations[12] are used to update the above variables. In this way, AP i uses the recent results of its neighboring cells to update its own variables so that the interference status becomes more accurate. Specifically, the variables are updated in the following circular fashion.

$$\boldsymbol{\gamma}_j^{n(k+1)} = f_\gamma(p_j^{n(k)}, \boldsymbol{\alpha}_{i+}^{(k)}, \boldsymbol{\alpha}_{i-}^{(k+1)}) \tag{28a}$$

$$y_j^{n(k+1)} = f_y(p_j^{n(k)}, \boldsymbol{\gamma}_j^{n(k+1)}, \boldsymbol{\alpha}_{i+}^{(k)}, \boldsymbol{\alpha}_{i-}^{(k+1)}) \tag{28b}$$

$$p_j^{n(k+1)} = f_p(\boldsymbol{\gamma}_j^{n(k+1)}, y_j^{n(k+1)}, \boldsymbol{b}_{i+}^{(k)}, \boldsymbol{b}_{i-}^{(k+1)}) \tag{28c}$$

$$Q_{ij}^{n(k+1)} = f_Q(p_j^{n(k+1)}, \boldsymbol{\gamma}_j^{n(k+1)}, y_j^{n(k+1)}, \boldsymbol{b}_{i+}^{(k)}, \boldsymbol{b}_{i-}^{(k+1)}) \tag{28d}$$

$$\alpha_j^{n(k+1)} = f_\alpha(Q_{ij}^{n(k+1)}) \tag{28e}$$

where,

$\boldsymbol{\alpha}_{i+}^{(k)} = [\alpha_{j'}^{n(k)}, p_{j'}^{n(k)} | j' \in J_{i'}, i' \in I, i' > i]$,

$\boldsymbol{\alpha}_{i-}^{(k+1)} = [\alpha_{j'}^{n(k+1)}, p_{j'}^{n(k+1)} | j' \in J_{i'}, i' \in I, i' < i]$,

$\boldsymbol{b}_{i+}^{(k)} = [y_j^{n(k)} | j' \in J_{i'}, i' \in I, i' > i]$,

$\boldsymbol{b}_{i-}^{(k+1)} = [y_j^{n(k+1)} | j' \in J_{i'}, i' \in I, i' < i]$,

and $f_\gamma(\cdot)$, $f_y(\cdot)$, $f_p(\cdot)$, $f_Q(\cdot)$, $f_\alpha(\cdot)$ correspond to Eqs (25a), (25b), (26), (24), (27), respectively; superscript k (or $k+1$) represents the index of iterations.

Once Eq. (22) is solved for all n, Lagrange multipliers $\boldsymbol{\rho}$, $\boldsymbol{\mu}$ can be updated using subgradient methods.

$$\rho_j \leftarrow \rho_j - \tau_1(\tilde{r}_j - z_j) \tag{29a}$$

$$\mu_j \leftarrow \mu_j - \tau_2(P_{\max} - \sum_{n\in N}\tilde{p}_j^n) \tag{29b}$$

where τ_1 and τ_2 are the positive step-sizes; $\tilde{p}_j^n$ is the so-

lution of Eq. (22); and $\tilde{r}_j$ is calculated by Eq. (5).

The solution of P2 is summarized in Algorithm 2, where the collections of variables γ_j^n, y_j^n and Q_{ij}^n are denoted as $\boldsymbol{\gamma}$, $\boldsymbol{y}$ and $\boldsymbol{Q}$, respectively. Compared to the baseline FP, the general multi-subcarrier FP further: (1) performs joint power allocation among multiple subcarriers under the total power constraint, with corresponding updates in Eq. (24) and Eq. (26); and (2) updates variables by Gauss-Seidel iterations in Eq. (28) to better handle the interlinked resource allocation due to the interference.

4 Performance evaluation

In this section, simulations are performed to evaluate the proposed algorithm. The simulation setting follows a pico-cell scenario of the 3rd Generation Partnership Project (3GPP)[13-14]. The general assumptions are as follows. A homogeneous deployment with three APs is considered, where each AP serves three devices in its coverage. Each AP is located at the center of the cell and the inter-AP distance is 100 m. All APs fully reuse 10 MHz frequency band, which consists of 50 resource blocks (RBs) with each RB consisting of 12 continuous subcarriers. The unit for frequency scheduling is one RB. The maximum uplink transmission power is 23 dBm. The power spectral density of noise at each AP is -174 dBm/Hz. A 3GPP probability-based path loss model is considered. The standard deviations of shadow fading are 3 dB and 4 dB for line of sight (LOS) and non-line of sight (NLOS), respectively. The small-scale fading is modeled as Rayleigh fading. The channel is assumed to be frequency flat within one RB but independent for different RBs. The data size of each task is set to 10 Mbits, and the number of CPU cycles needed to process one bit of each task to 100 cycles/bit.

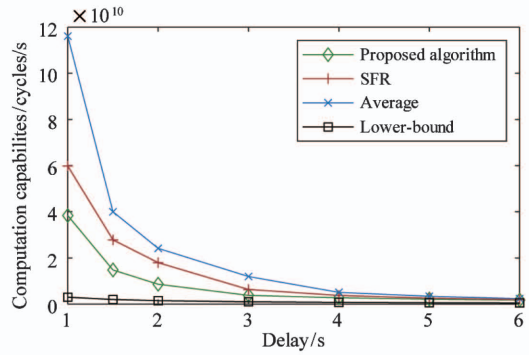
The effects of three practical factors on the required computation capabilities are empirically evaluated. These are tasks' delay requirements, resource allocation strategies, and the interference status. First, the task's delay is varied, denoted as d, from 1 s to 6 s to indicate both delay-critical and delay tolerant tasks. Second, three benchmarks are used to evaluate the proposed algorithm. These are the soft frequency reuse (SFR) scheme, which is one of the representative techniques for frequency scheduling to handle the inter-cell interference [15], a generic average resource allocation scheme (average), and the 'lower-bound' performance based on Proposition 1. Finally, two scenarios are considered for the interference status characterized by devices' locations, where at the 'cell edge', the minimum distance from the device to its serving AP is 45 m, and in the 'cell area', it is 5 m.

In Fig. 2, the necessary computation capabilities of the network are compared using different resource allocation (RA) algorithms under the above settings. Simulation results show that, as the delay becomes large, the required computational resources become sharply reduced. Intuitively, this effect is consistent with the fact that the delay-critical tasks demand much more computational resources than the delay-tolerant tasks. Furthermore, the demand approximately approaches to the lower bound under weak delay demand, e.g., $d > 4$ s. This indicates that the resource provisioning for computation capabilities can be simplified by using the lower-bound estimation instead of running an iterative algorithm.

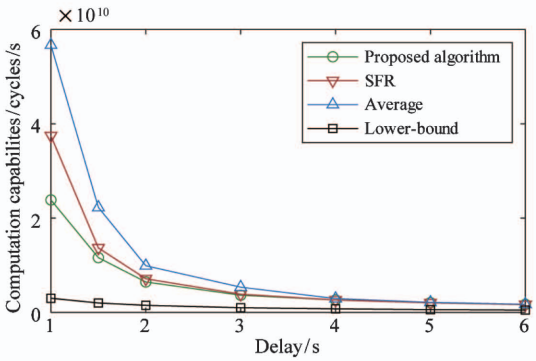
It is observed that, for those delay-critical tasks, the scheduling strategy has a strong impact on the necessary computation capabilities of the network. Take $d = 1$ s as an example. The demand for computation capabilities with the proposed algorithm reduces to about one third of the average scheme for cell-edge devices, as shown in Fig. 2(a). The reason of this effect is that the proposed algorithm is more efficient in dealing with the inter-cell interference and thus achieves higher communication rates. Similarly, SFR also brings benefits to reduce computation costs compared to the average scheme, but is still inferior to the proposed algorithm.

There exists a trade-off between the transmission rate and the demand for computational resources. When devices are randomly located in the whole cell area, some locations near to the cell center increase communication rates, due to the strong received signal strength and the negligible interference. Hence, as shown in Fig. 2(b), the demand for computational resources under each scheduling algorithm is highly reduced compared to that in the cell-edge scenario (Fig. 2(a)).

Fig. 3 further illustrates the experimental cumulative distribution function (CDF) of the required computation capability for each AP when devices are located at the cell edge. Roughly speaking, the computational resources are varying in a certain range due to the inter-cell interference and the received signal strength, which finally leads to dynamic communication rates. The ranges are rather wider for both the average allocation scheme and SFR than the proposed algorithm. Thus, it is necessary to balance the worst case and the deployment cost during the computational resource planning for APs in practical wireless networks.

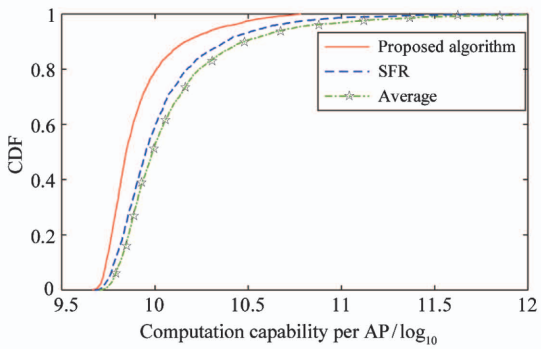


(a) Devices are uniformly distributed at the cell edge

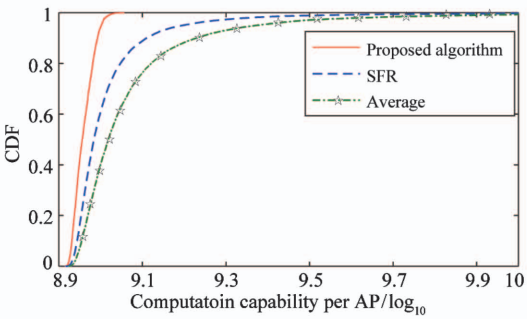


(b) Devices are uniformly distributed in the cell area

Fig. 2 The minimum demand for the total computation capabilities of the network under varying delays and different devices' locations



(a) Task's delay is 1 s



(b) Task's delay is 4 s

Fig. 3 Experimental CDF of computation capabilities of each AP. Devices are randomly located at the cell edge

The degree of the demand for computational resources is summarized in Table 1 under the previously mentioned constraints according to the analysis of the simulation results. It indicates the significance of the different factors during the resource provisioning in the practical multi-cell network deployment.

Table 1 Demand for computational resources under key factors

Key factors \ Resource demand		high	low
Delay requirement	critical	√	
	tolerant		√
Inter-cell interference	strong	√	
	weak		√
RA algorithms	optimized		√
	general	√	

5 Conclusion

The resource provisioning problem is investigated for APs with communication-and-computation capabilities in multi-cell wireless networks. Specifically, the minimum demand for computational resources is quantified to guarantee the tasks' requirements under practical concerns. The analysis shows that the resource provisioning depends on key factors such as delay requirements, scheduling strategies, and the interference status. In particular, delay-critical tasks demand much more computational resources and are more sensitive to scheduling strategies than delay-tolerant tasks. Further, for delay-tolerant tasks, a lower bound is derived to estimate the demand for computation capabilities, which is more computationally efficient. The conclusions are beneficial for wireless operators when upgrading an existing network or deploying a new one for emerging applications.

References

[1] Mao Y Y, You C S, Zhang J, et al. A survey on mobile edge computing: the communication perspective [J]. *IEEE Communications Surveys Tutorials*, 2017, 19(4): 2322-2358

[2] Zhou Y Q, Tian L, Liu L, et al. Fog computing enabled future mobile communiation networks: a convergence of communication and computing[J]. *IEEE Communications Magazine*, 2019, 57(5): 20-27

[3] Ku Y J, Lin D Y, Lee C F, et al. 5G radio access network design with the fog paradigm: confluence of communications and computing[J]. *IEEE Communications Magazine*, 2017, 55(4): 46-52

[4] You C S, Huang K B, Chae H, et al. Energy-efficient resource allocation for mobile-edge computation offloading

[J]. *IEEE Transactions on Wireless Communications*, 2017, 16(3): 1397-1411

[5] Zheng J C, Cai Y M, Wu Y, et al. Dynamic computation offloading for mobile cloud computing: a stochastic game-theoretic approach [J]. *IEEE Transactions on Mobile Computing*, 2019, 18(4): 771-786

[6] Liu L Q, Chang Z, Guo X J, et al. Multiobjective optimization for computation offloading in fog computing [J]. *IEEE Internet of Things Journal*, 2018, 5(1): 283-294

[7] Moura J, Hutchison D. Game theory for multi-access edge computing: survey, use cases, and future trends [J]. *IEEE Communications Surveys Tutorials*, 2019, 21(1): 260-288

[8] Lin C C, Yang J W. Cost-efficient deployment of fog computing systems at logistics centers in industry 4.0 [J]. *IEEE Transactions on Industrial Informatics*, 2018, 14(10): 4603-4611

[9] Jia M, Cao J N, Liang W F. Optimal cloudlet placement and user to cloudlet allocation in wireless metropolitan area networks [J]. *IEEE Transactions on Cloud Computing*, 2017, 5(4): 725-737

[10] Yang X M, Liu Z N, Yang Y. Minimization of weighted bandwidth and computation resource of fog servers under per-task delay constraint [C] // Proceeding of the 2018 IEEE International Conference on Communications, Kansas City, USA, 2018: 1-6

[11] Shen K M, Yu W. Fractional programming for communication systems part II: uplink scheduling via matching [J]. *IEEE Transactions on Signal Processing*, 2018, 66(10): 2631-2644

[12] Palomar D P, Chiang M. A tutorial on decomposition methods for network utility maximization [J]. *IEEE Journal on Selected Areas in Communications*, 2006, 24(8): 1439-1451

[13] 3GPP. Further enhancements to LTE time division duplex for downlink-uplink interference management and traffic adaptation [S]. TR 36.828, v11.0.0, 2012

[14] Ding M, Lopez-Perez D, Claussen H, et al. On the fundamental characteristics of ultra-dense small cell networks [J]. *IEEE Network*, 2018, 32(3): 92-100

[15] Chang S H, Park H G, Kim S H, et al. Study on coverage of full frequency reuse in FFR systems based on outage probability [J]. *IEEE Transactions on Communications*, 2018, 66(11): 5828-5843

Yang Xiumei, born in 1979. She received the Ph. D. degree from the Shanghai Institute of Microsystem and Information Technology (SIMIT), Chinese Academy of Sciences (CAS), Shanghai, China, in 2011. She also received the B. S. and M. S. degrees in communications and information systems from Shandong University, Jinan, China, in 2001 and 2004, respectively. She is currently a professor at SIMIT, CAS. During 2018-2019, she was a visiting scholar at Northwestern University, Evanston, USA. Her current research interests include fog computing, Internet of Things and network intelligence.