Compution Offloading Strategy Based on Joint Allocation in Mobile Device Cloud

Yi-di XIE, Yong ZHANG, Ying-lei TENG, Ruo-fan XIA and Da GUO
Beijing University of Posts and Telecommunications, Beijing 100876, China

**Keywords:** Computation offloading, Mobile device cloud, Task allocation, Communication competition, Simulated annealing.

**Abstract.** The contradiction between mobile terminal limited energy and users’ high requirements is more and more intensified, to provide satisfactory computation performance for real time applications, offloading workloads to Mobile Device Cloud (MDC) will be adopted. In this paper, we consider the problem of communication competition in the network, and put forward the offloading strategy based on the Joint Allocation Scheme (JAS) for lower latency. In the physical layer, the channel bandwidth allocation mode of OFDMA is used to make transmission rate realistic, combined with the task assignment, a delay model belongs to the NP problem can be established. We use the step-by-step solution, firstly consider the allocation of continuous task, and then use the simulated annealing algorithm (SAA) to get the optimal solution.

**Introduction**

Nowadays, cloud or cloudlet can’t meet the requirements of some time-sensitive procedures, delay performance still need to be improved. Through MDC technology, users can cooperate with the surrounding idle computing resources to solve this problem.

The latency of computation offloading consists of two parts: data transmission and task execution, which involves the type of communication mode, blocking and non-blocking, the parallel processor begins calculation immediately after receiving its assigned fraction of load, which is called non-blocking mode in [1]. In contrast, there is a blocking mode which introduces transmission and computing start time in literature[2], and taking system delay and energy consumption into account, but does not consider the communication competition.

Every offloading target corresponds to a respective subtask in the network, and the dependency among subtasks affects the task scheduling. [3] describes the cloud computation offloading architecture of the mobile device with multi-level dependency, where cloudlet is used as the middle layer task scheduling. There is research investigating the involvement of user processor in communication and its impact on task scheduling[4]. Literature[5] introduces middleware. The task allocation of large-scale offloading is often based on the optimization of scheduling algorithm. In literature[6], there are optimal results for single-round algorithms and the design of an asymptotically optimal multi-round algorithm. In addition to system latency, there are some equally important factors that can be considered, such as energy consumption and network lifetime[7].

In this paper, we consider communication competition, propose JAS, which considers the joint dynamic allocation[8] of tasks and bandwidth. Define task as a continuous variable and assign subcarriers to offloading targets. The main contributions of our work are summarized as follows:

The dynamic channel allocation in OFDMA is introduced, where bandwidth resources are divided into orthogonal subcarriers. Establish system latency model belonging to a kind of NP problem, considering the influence of bandwidth and task allocation on delay.

Propose JAS to analyze the latency. Once the channel allocation is determined, task partition with minimum latency can be seen, and the global optimal subcarriers allocation will be further determined. "task-node" mapping matrix is used to represent the delay generated by each node.
The simulated annealing algorithm is applied to search bandwidth allocation scheme, and latency is used as the objective function while establishing the annealing process. Compared with other distribution methods, the optimal solution of the search shows advantage on reducing delay.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_j$</td>
<td>computation of subtask $j$</td>
</tr>
<tr>
<td>$d_j$</td>
<td>data size of subtask $j$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>computation of unit data</td>
</tr>
<tr>
<td>$c_i$</td>
<td>computation speed of slave $i$</td>
</tr>
<tr>
<td>$r_i$</td>
<td>transmission speed of link $i$</td>
</tr>
<tr>
<td>$D$</td>
<td>total offloading data</td>
</tr>
<tr>
<td>$B$</td>
<td>total bandwidth</td>
</tr>
<tr>
<td>$R$</td>
<td>total communication resource</td>
</tr>
<tr>
<td>$X$</td>
<td>task-node correspondence matrix</td>
</tr>
<tr>
<td>$L^1_i$</td>
<td>transmission latency of slave $i$</td>
</tr>
<tr>
<td>$L^2_i$</td>
<td>execution latency of slave $i$</td>
</tr>
<tr>
<td>$N$</td>
<td>total number of subcarriers</td>
</tr>
<tr>
<td>$n$</td>
<td>number of subtasks/slaves</td>
</tr>
</tbody>
</table>

**System Model**

This paper selects the blocking mode to simulate the communication among devices, analyzes the behavior of a single user $N_0$ (master) who has the right to dynamically allocate the subcarriers $[w_1,w_2,w_3,...,w_n]$, divides the total task into $n$ subtasks, corresponding to appropriate nodes (slave), parameters are introduced in Table 1. Figure 1 reflects the nodes collaboration relationship in MDC.

In the OFDMA system, in view of the fairness of bandwidth allocation, channel gain and system throughput, the subcarriers corresponding to each slave node are a subset of subcarriers complete set $F$, which can be written as $F_i=\{f_1,f_2,f_3,...,f_N\}$, have $m_i$ subcarriers with same size, the bandwidth allocated to slave $i$ is $W_i=m_i\frac{B}{N}$ [9]. The model considers the heterogeneity of subcarrier channel transmission conditions, according to the Shannon formula we define

$$B = \sum_{i=1}^{n} W_i \cdot$$ (1)

and the sum of average communication rates[10] can be defined as (2).

$$R = \sum W_i \left[ \sum_k \lambda \cdot \log_2 \left( 1 + p_{ik} |h_{ik}|^2 \eta \right) \right] = \sum_{i=1}^{n} r_i$$ (2)

where $p_{ik}|h_{ik}|^2\eta$ denotes the signal-noise ratio (SNR) of master $i$ on subcarrier $k$. It is possible to set the ratio of the average transmission rate to the maximum rate to $\lambda$. System delay consists of two parts.
the computation delay $L_c$, the data transmission delay $L_t$. The slave node from which the result of the computation is to be processed, determines the overall delay of offloading

$$K = \max \left[ L_i + L'_i \right] = \max \left[ \frac{w_i}{c_i} + \frac{d_{ij}}{r_i} \right],$$

where subtask $j$ is assigned to slave node $i$.

With the minimum delay target, all slave nodes deliver tasks at the same time, the maximum delivery time of the system is equal to the delivery delay of each slave node [11]. In simplified model, the task of processing unit data is an unit task [12], that is $w_j=\alpha d_j$, $j=(1,2,\ldots,n)$, overall delay is

$$K = d_{ij} \left[ \frac{\alpha}{c_i} + \frac{1}{r_j} \right] \Rightarrow d > 0.$$  \hspace{1cm} (4)

Set vectors $\vec{d}=\left(d_1, d_2, \ldots, d_n\right)$, $\vec{c}=\left(\frac{1}{c_1}, \frac{1}{c_2}, \ldots, \frac{1}{c_n}\right)$, $\vec{r}=\left(\frac{1}{r_1}, \frac{1}{r_2}, \ldots, \frac{1}{r_n}\right)$, then define $\vec{q}=\alpha \cdot \vec{c} + \vec{r}$, with the total amount of data $D$ and transmission resource $R$ unchanged, the problem can be expressed as:

$$\begin{align*}
\min_{\mathbf{X}q} & \quad d^T \mathbf{X} q \\
\text{s.t.} & \quad x_{ij} \in \{0,1\} \quad \sum_{i} x_{ij} = 1 \\
& \quad \sum_{j} x_{ij} = 1 \\
& \quad \sum_{j} d_j = D \\
& \quad \sum_{j} r_j = R
\end{align*}$$

where $\mathbf{X}$ is mapping relation matrix between nodes and tasks, containing element $x_{ij}$ which is 1 when subtask $j$ is assigned to node $i$ and 0 in other cases. Problem (5) can be summarized as NP problem[13], distribution scheme involves allocation of tasks and bandwidth of the OFDMA system.

**Joint Allocation Scheme**

**Optimal Subtask Assignment**

The amount of data varies according to the characteristics of task. At the same time, the link between the master node and each slave node has been formed. When the transmission rate is determined, each task allocation scheme corresponds to a set of subtask delivery time.

In the $L \times n \times n$ matrix below, an initial matrix $X_0$ is chosen as initial correspondence state. At this time, the subtask $T_j$ and the slave node $N_i$ have an association relation, the initial delivery time matrix $d^T X q$ can be obtained. Since all node delivered time is the same, that is $L_{ij} = L_{pq}$,

<table>
<thead>
<tr>
<th></th>
<th>$N_1$</th>
<th>$N_2$</th>
<th>$\ldots$</th>
<th>$N_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>$L_{11}$</td>
<td>$L_{12}$</td>
<td>$L_{1n}$</td>
<td></td>
</tr>
<tr>
<td>$T_2$</td>
<td>$L_{21}$</td>
<td>$L_{22}$</td>
<td>$L_{2n}$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td></td>
</tr>
<tr>
<td>$T_n$</td>
<td>$L_{n1}$</td>
<td>$L_{n2}$</td>
<td>$L_{nn}$</td>
<td></td>
</tr>
</tbody>
</table>

and according to formulation (3), define $z_i$ in (6), and the best task allocation can be obtained under the above bandwidth condition.

$$\frac{\alpha}{c_i} + \frac{1}{r_j} = z_i \Rightarrow d_i \cdot z_i = K \Rightarrow d_i = \frac{D}{z_i \sum_j \frac{1}{z_j}}.$$  \hspace{1cm} (6)
Bandwidth Allocation Based on Simulated Annealing

The SAA firstly set a highest initial temperature, corresponding to the initial solution to the problem and the objective function. With the decrease of temperature, the search is carried out around the initial solution. According to the Metropolis criterion[14], a new scheme is accepted with a probability, and the global range is advanced to the local range to obtain the global optimal solution. SAA mainly identifies the following questions:

1. Objective function
   In the process of simulated annealing, the delay $K$ of offloading is taken as the optimization target, and each bandwidth allocation mode is $S_q$, and the set of all reasonable states is $S$, and the system delay under a specific state is
   \[ K_q = \frac{w_i + d_j}{c_i - r_j} \]  
   (7)

2. Neighborhood selection
   In the subcarrier allocation state $S_q$, we define the number of elements contained in the available subset of the nodes $F_q$, the allocation mode satisfying the condition
   \[ \sum \text{length}(F_q) \leq N, \ S_q \in S. \]  
   (8)
   the current state neighborhood is $S' = S - S_q$.

3. Cooling process
   Select an appropriate highest temperature $T_h$ (initial temperature) to ensure the quality of the final solution. The cooling coefficient is $\beta$, and there is $T_{q+1} = T_q \beta$ in every time slot until reaching the lowest temperature $T_l$ (termination temperature), and each time slot corresponds to a current subcarrier allocation scheme $K_q$. Then calculate $\Delta K = K_{best} - K_q$. If the current solution is superior to the optimal solution $K_{best}$, replace it, otherwise replace it with probability $f_{chg} = \exp(\Delta K/T_p)$.

![Figure 2. Simulated annealing algorithm flow chart.](image)

4. Stop condition
   Reaching the end temperature $T_f$ or the set ideal latency value $K_{fin}$.

Simulation Analysis

In this section, we evaluate the performance of the JAS proposed by comparing it with Average Allocation (AA) and Random Allocation (RA). In AA scheme, master equally allocates the available bandwidth resources to each link, and the amount of data held by the master node is evenly distributed. In RA scheme, master node uses Flooding mechanism to find available nodes, then sorts nodes in the queue with the order of discovery, all subtasks will be assigned to the various nodes in this order, while the channel subcarriers will not be re-planned.

297
In the simulation platform, the total data value is 500k, the total bandwidth of the network is 1 MHz, \( \lambda \) is set to \( 2^{-1/2} \), the SNR is in the range of [10, 50]dB randomly, the total number of subcarriers is 512, and the computation rate of the slave node is in the range of [100k, 300k] randomly, \( T_h \) is 1000, \( \beta \) is 0.95. \( T_l \) is 0.001. Figure 3 expresses the evolution of objective function. We set \( \alpha \) which can be considered to be task volume /data volume ratio to 50, the total task volume maintains unchanged, as Fig.4 shows, with the number of participating increasing, the system delay has a downward trend.

![Figure 3. SAA evolution.](image)

![Figure 4. Number of slaves – latency.](image)

![Figure 5. Task/data – latency.](image)

In Figure 5, the number of slaves is 8, observe data/task proportion factor, when \( \beta \) increase, the total delay has a rising trend, the above two figures show that JAS achieves better performance than the other two disciplines.

**Summary**

This paper presents a computational offloading strategy aiming to reduce latency in MDC and establishes mathematical model. Then propose the JAS, which takes the task allocation and communication competition into account, and handle the NP problem step-by-step.

From the simulation results, the JAS is superior to other allocation schemes. How to set the subtracting strategy of sub-task interdependence and the incentive mechanism of cost optimization in the system are the focus in the future study.

**Acknowledgement**

This work is supported by National Natural Science Foundation of China (61171097), it’s a great honor to get the help.

**References**


