基于终端能耗和系统时延最小化的边缘计算卸载及资源分配机制

代美玲 刘周斌 郭少勇 邵苏杰 邱雪松*
(北京邮电大学网络与交换技术国家重点实验室 北京 100876)

摘 要：通过移动边缘计算下移云端的应用功能和处理能力支撑计算密集或时延敏感任务的执行成为当前的发展趋势。但面对众多移动终端用户时，如何有效利用计算资源有限的边缘节点来保障终端用户服务质量(QoS)成为关键问题。为此，该文融合边缘云与远端云构建了一种分层的边缘云计算架构，以此架构为基础，以最小化移动设备能耗和任务执行时间为目标，将问题形式化描述为资源约束下的最小化能耗和时延加权和的凸优化问题，并提出基于乘子法的计算卸载及资源分配机制解决该问题。实验结果表明，在计算任务量很大的情况下，提出的计算卸载及资源分配机制能够有效降低移动终端能耗，并在任务执行时延方面较局部计算与计算卸载机制分别降低最高60%与10%，提高系统性能。

关键词：边缘计算；计算卸载；资源分配；终端能耗；系统时延
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1009-5896(2019)11-2684-07
DOI: 10.11999/JEIT180970

A Computation Offloading and Resource Allocation Mechanism Based on Minimizing Devices Energy Consumption and System Delay

DAI Meiling LIU Zhoubin GUO Shaoyong SHAO Sujie QIU Xuesong
(*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract: To support the execution of computation-intensive, delay-sensitive computing task by moving down the computing and processing capability in mobile edge computing becomes the current trend. However, when serving a large number of mobile users, how to use effectively the edge nodes with limited computing resources to ensure Quality of service (QoS) of end-user has become a key issue. To solve this problem, the edge cloud and remote cloud are combined to build a layered edge cloud computing architecture. Based on this architecture, with the goal of minimizing mobile device energy consumption and task execution time, the problem which is proved to be convex is formulated to minimize the weight sum of energy and delay. A computation offloading and resource allocation mechanism based on multiplier method is proposed. Simulations are conducted to evaluate the proposed mechanism. Compared with local computing and computation offloading mechanism, the proposed mechanism can effectively reduce the energy consumption of mobile device and the delay of system by up to 60% and 10%, respectively.

Key words: Edge computing; Computing offloading; Resource allocation; Energy consumption; Delay of system

1 引言

移动边缘计算通过无线接入网络“增强”移动终端用户计算能力，实现如视频监测，语音识别，增强现实[1]等计算密集、时延敏感任务的执行。但面对用户数量持续增加，计算任务量不断增大的情况，边缘节点有限的计算能力依旧无法很好地满足用户需求，如何利用计算资源有限的边缘节点保障终端用户服务质量成为当前研究的热点问题。

目前，边缘云与远端云相结合的协同计算模式是解决边缘节点计算能力有限，无法有效满足用户需求问题的有效方法[2–6]。在该协同计算模式下，移动终端计算过载时，大量的任务负载将被卸载至边缘节点和远端云节点，从而在有效缓解边缘节点负载压力的同时，节省用户设备的资源消耗和任务处理时间，实现任务的高效计算。在单用户场景[7–11]下，为了获得最优的卸载策略，文献[7–9]考虑了任

收稿日期：2018-10-17；改回日期：2019-03-13；网络出版：2019-04-01
*通信作者： 邱雪松 xsqiu@bupt.edu.cn
基金项目：国家电网公司科技项目(52110118001H)
Foundation Item: The State Grid Technology Project (52110118001H)

务划分和任务分配问题，通过设定的指标决定任务是否被迁移，迁移量最优为多少。文献[10]考虑任务在通信和计算上的不同需求，定义一个新的名为计算能耗效率的变量，用于通信-计算调度，解决时延约束的计算卸载问题，文献[11]研究了一种特殊类型的面向数据划分的应用程序，设计该类应用程序被部分卸载，结合考虑传输功率的最优化，实现移动用户的能耗最小化。但单用户场景与现实实际多用户场景存在较大差异，使得该类卸载策略在实际应用中存在问题。面对现实场景下的多用户问题，如何在考虑边缘节点资源分配的条件下制定最优卸载决策是面临的难点问题。文献[12–14]中利用博弈论的方法在多用户的云计算环境下得到最优化卸载决策，文献[12]研究多无线信道环境中的计算卸载问题，文献[13]通过设定单一无线接入点描述多用户计算卸载问题为计算卸载博弈，并分别设计算法获得计算卸载决策，但忽略了边缘节点资源分配对计算卸载的影响。文献[15]中对资源分配和卸载决策制定进行了联合优化，达到在满足用户时延约束的同时节省能耗的目标，为了简化计算，将移动设备能耗设定为常量，忽略了时间变化对用户能耗的影响。

综上，边缘云与远端云相结合的多用户多任务场景下，终端用户的服务质量保障方面仍有诸多问题有待解决，其中主要问题之一是如何实现高效的计算卸载和资源分配。本文融合边缘云与远端云构建了一种面向多移动终端的分层边缘云计算架构。依此，以最小化终端能耗为基础目标，最小化任务执行时延为关键指标，采取集中式卸载决策方式，构建凸优化问题模型。针对该问题，基于拉格朗日乘子法，在设定平均资源分配的条件下，满足Karush-Kuhn-Tucker(KKT)条件，进行初始求解，获得多终端用户场景下的初始计算卸载决策。同时，针对资源分配的问题，提出了基于乘子法的计算卸载及资源分配机制，在进行计算卸载决策的同时，实现边缘节点的最优资源分配。

2　系统模型及问题描述

2.1　系统模型

分层边缘云计算架构如图1所示，由一个远端云节点、多个边缘计算节点以及多个移动终端设备组成。其中，远端云节点与边缘节点、边缘节点与移动终端设备分别存在一对多的映射关系，移动终端设备通过无线网络接入到无线接入点，无线接入点可以通过互联网将任务卸载至远端云节点并接受远端云节点响应的计算结果，返回给移动终端设备。

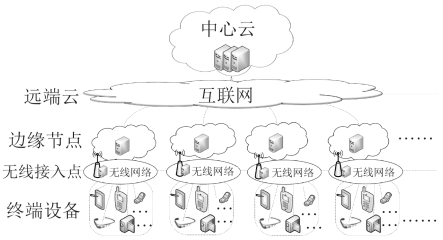

图 1 分层边缘云计算架构

系统最优化目标主要包括两部分：系统时延与移动终端能耗。系统时延由计算时延与通信时延两部分组成。从终端用户卸载任务到边缘节点和远端云节点有益于接入充足的计算资源从而减少计算时延，但同时，额外的通信时延被加入到系统时延中。所以，高效的计算卸载算法需要综合考虑计算负载调度和相关通信限制，平衡计算时延和通信花销；移动终端设备能耗主要考虑计算能耗和传输能耗。计算能耗主要指移动终端本地执行任务所需考虑的计算能耗，传输能耗则为卸载任务时传输数据所产生的通信方面的能耗。

2.2　用户模型

每个无线接入点管理N个移动终端，表示为$N=[1,2,\cdots,n]$，其中每个移动终端拥有计算密集、时延要求严格的任务需要执行，将终端用户U_i的计算任务定义为元组$W_i=\left(B_i^{\text{in}},V_i,B_i^{\text{out}}\right)(i\in\{1,2,\cdots,n\})$，并假设单用户一次请求仅有一种类型任务。$B_i^{\text{in}}$表示任务输入数据大小，$V_i$表示单任务执行所需CPU周期数，$B_i^{\text{out}}$表示响应数据大小。类似于已存在的一些相关工作[7]，本文将考虑一种准静态场景，在任务卸载过程中，所有的移动终端设备和无线网络状态不发生改变，设备不切换接入点。本文将假设终端i当前有一批任务量为λ_i的任务需要被执行，则终端i的迁移决策约束为

$$x_i+x_i^{\text{edge}}+x_i^{\text{cloud}}=\lambda_i \tag{1}$$

(1)计算能耗

基于用户模型定义，移动终端i执行任务的时间t_i^{local}及任务计算能耗E_i^{local}为

$$t_i^{\text{local}}=\frac{v_i}{C_i}\cdot x_i \tag{2}$$

$$E_i^{\text{local}}=P_i^{\text{c}}\cdot t_i^{\text{local}} \tag{3}$$

其中，定义终端i的计算能力为C_i，即单位时间运行CPU周期数。P_i^{c}为用户执行浮点计算的能耗功率。

(2)传输能耗

基于以上定义与假设，移动终端i的数据传输能耗E_i^{up}可如下定义

$$E_i^{\text{up}}=P_i^{\text{up}}\cdot t_i^{\text{tran}} \tag{4}$$

P_i^{up}为用户通过无线网络传输数据到无线接入点的信号传输功率，t_i^{tran}为用户数据传输时间。

2.3 边缘计算模型

传输模型部分主要包括无线传输与互联网传输两部分。设为用户分配的信道带宽为B, h_i^{up}为终端i与边缘节点间的信道增益，假设移动终端i上传传输速率为r_i，则可达到的传输速率为

$$r_i = B \cdot \log_2\left(1 + \frac{h_i^{\mathrm{up}} \cdot P_i^{\mathrm{up}}}{B \cdot N_0}\right) \tag{5}$$

定义边缘节点为移动终端设备i分配的计算资源为C_i^{edge}，为边缘节点单位时间可执行的CPU周期次数。可得到数据从终端传输至边缘节点的传输时延t_i^{te}及当前卸载至边缘的任务执行时间t_i^{ee}为

$$t_i^{\mathrm{te}} = \frac{B_i^{\mathrm{in}}}{r_i} \cdot x_i^{\mathrm{edge}} \tag{6}$$

$$t_i^{\mathrm{ee}} = \frac{v_i}{C_i^{\mathrm{edge}}} \cdot x_i^{\mathrm{edge}} \tag{7}$$

其中计算资源C_i^{edge}满足如下约束

$$\sum_{i=1}^{n} C_i^{\mathrm{edge}} \le C_{\mathrm{edge}} \tag{8}$$

故边缘节点执行卸载任务的时延及用户能耗分别为

$$t_i^{\mathrm{edge}} = t_i^{\mathrm{te}} + t_i^{\mathrm{ee}} \tag{9}$$

$$E_i^{\mathrm{edge}} = P_i^{\mathrm{up}} \cdot t_i^{\mathrm{te}} \tag{10}$$

2.4 云计算模型

互联网传输在无线接入点与远端云节点间进行，定义远端云节点任务数据传播时延为t_i^{tc}，包括移动终端节点传输数据至无线接入点时间t_i^{ue}及无线接入点至远端云节点传输时间t_i^{ec}两部分

$$t_i^{\mathrm{tc}} = t_i^{\mathrm{ue}} + t_i^{\mathrm{ec}} \tag{11}$$

假设边缘节点至远端云节点的互联网传输速率为R_i，则有

$$t_i^{\mathrm{ue}} = \frac{B_i^{\mathrm{in}}}{r_i} \cdot x_i^{\mathrm{cloud}} \tag{12}$$

$$t_i^{\mathrm{ec}} = \frac{B_i^{\mathrm{in}}}{R_i} \cdot x_i^{\mathrm{cloud}} \tag{13}$$

远端云节点的任务量为x_i^{cloud}，设定远端云节点执行能力为C_{cloud}。平稳状态下，鉴于远端云节点的计算资源充足，不考虑资源分配。对于远端云节点计算模型，其计算时间可考虑为

$$t_i^{\mathrm{exec}} = \frac{v_i}{C_{\mathrm{cloud}}} \cdot x_i^{\mathrm{cloud}} \tag{14}$$

考虑边缘节点到远端云节点的传播时延为$\triangle t$，则远端云节点执行卸载任务的时延及用户能耗分别为

$$t_i^{\mathrm{cloud}} = t_i^{\mathrm{tc}} + t_i^{\mathrm{exec}} + \triangle t \tag{15}$$

$$E_i^{\mathrm{cloud}} = P_i^{\mathrm{up}} \cdot t_i^{\mathrm{ue}} \tag{16}$$

根据式(2)—式(16)，可得终端i的任务执行能耗和时延期望为

$$E_i = E_i^{\mathrm{local}} + E_i^{\mathrm{edge}} + E_i^{\mathrm{cloud}} \tag{17}$$

$$T_i = \frac{x_i}{\lambda_i} \cdot t_i^{\mathrm{local}} + \frac{x_i^{\mathrm{edge}}}{\lambda_i} \cdot t_i^{\mathrm{edge}} + \frac{x_i^{\mathrm{cloud}}}{\lambda_i} \cdot t_i^{\mathrm{cloud}} \tag{18}$$

2.5 问题描述

对于移动终端i，假设当前需要执行任务量为λ_i，卸载决策得到本地执行，边缘节点执行，远端云节点执行任务量为$< x_i, x_i^{\mathrm{edge}}, x_i^{\mathrm{cloud}} > (i \in \{1, 2, \cdots, n\})$。忽略无线接入点执行决策的计算时延与返回决策数据的传输时延，则针对边缘节点管理范围内的移动终端，处理任务的时延期望T及全体移动终端的总能耗E为

$$T = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \cdot T_i \tag{19}$$

$$E = \sum_{i=1}^{n} E_i \tag{20}$$

其中

$$\lambda = \sum_{i=1}^{n} \lambda_i \tag{21}$$

该问题将被描述为一个最优化问题，在2.1节系统模型中介绍了系统在请求阶段于无线接入点处实现卸载决策及资源分配，故针对卸载决策和资源分配，本文将最优化问题的解定义为$S =< X, C >$, $X \triangleq < x_1, x_2, \cdots, x_n, x_1^{\mathrm{edge}}, x_2^{\mathrm{edge}}, \cdots, x_n^{\mathrm{edge}}, x_1^{\mathrm{cloud}}, x_2^{\mathrm{cloud}}, \cdots, x_n^{\mathrm{cloud}} >$, $C \triangleq < C_1^{\mathrm{edge}}, C_2^{\mathrm{edge}}, \cdots, C_n^{\mathrm{edge}} >$。

系统时延是系统的QoS基本指标，最小化终端能耗是基础目标。故将问题形式化描述为

$$\left.\begin{aligned} &\min_S f(S) = \min_S \left(E(S) + V \cdot T(S)\right) \\ &\text{s.t. } x_i \ge 0, x_i^{\mathrm{edge}} \ge 0, x_i^{\mathrm{cloud}} \ge 0 \\ &C_i^{\mathrm{edge}} \ge 0, \sum_{i=1}^{n} C_i^{\mathrm{edge}} \le C_{\mathrm{edge}} \\ &x_i + x_i^{\mathrm{edge}} + x_i^{\mathrm{cloud}} = \lambda_i, i \in \{1, 2, \cdots, n\} \end{aligned}\right\} \tag{22}$$

其中，V为权重常量，具体取值与时延和能耗相关联，依据移动设备当前状态确定。

3　基于终端能耗和系统时延最小化的边缘计算卸载及资源分配机制

3.1　多用户计算卸载机制

为简化计算，设定单用户单任务场景。在该类场景，即$n=1$场景下，该问题简化为如下最优化问题

$$\left.\begin{aligned}&\min_{x_1,x_1^{\text{edge}},x_1^{\text{cloud}}} E_1+V\cdot T_1\\&\text{s.t.}\ \ x_1\geq 0,x_1^{\text{edge}}\geq 0,x_1^{\text{cloud}}\geq 0\\&x_1+x_1^{\text{edge}}+x_1^{\text{cloud}}=\lambda_1,i\in\{1,2,\cdots,n\}\end{aligned}\right\}\tag{23}$$

定理 1　最优化问题式(23)为凸优化问题。

定理1可根据凸函数判别的2阶条件进行证明，目标函数的海赛矩阵在取值空间上处处半正定，且问题式(23)的所有约束均为线性函数，故问题式(23)为凸优化问题。

由定理1知，该问题为凸优化问题，为求最优解，引入拉格朗日函数如下

$$L(X,\mu,\eta)=f(X)+\mu\cdot h_1(X)+\sum_{j=1}^{3}\eta_j\cdot g_j(X)\tag{24}$$

考虑其不等式约束及等式约束，对于该凸优化问题，满足KKT条件的点$<X,\mu,\eta>$是该问题的最优解。

针对多用户计算卸载决策的制定，在不考虑资源分配时，终端不存在相互关联关系，可将目标函数转化为分别针对单用户求解。分别找到每个用户满足KKT条件的最优解，求出的解集合X即为多用户最优解

$$\begin{aligned}&\min_X\left(E(X)+V\cdot T(X)\right)\\&=\min_X\sum_{i=1}^{n}\left(E_i+V\cdot\frac{\lambda_i}{\lambda}\cdot T_i\right)\\&=\sum_{i=1}^{n}\min_X\left(E_i+V\cdot\frac{\lambda_i}{\lambda}\cdot T_i\right)\end{aligned}\tag{25}$$

该问题引入拉格朗日函数求得非整数最优解，未满足实际需求，求解后对最优解$<x_i,x_i^{\text{edge}},x_i^{\text{cloud}}>$分别向下取整，再将3组可行解代入目标函数，获得目标最优解。求解算法于表1中详细说明。

3.2　多用户计算卸载及资源分配机制

定义问题式(22)解空间为$S=<X,C>$。

定理 2　问题式(22)为凸优化问题。

定理2同定理1的证明类似，通过计算得到目标函数的海赛矩阵，可计算该矩阵的行列式在取值上大于等于0，该矩阵在取值空间上处处半正定。问题式(22)的约束条件均为线性约束，该问题为凸优化问题。

表 1　多用户计算卸载

初始化：各移动终端数量n及计算能力C_i，边缘节点计算能力C_{edge}，远端云节点计算能力C_{cloud}，无线带宽资源B，权值V，$S=\varnothing$;

输入：各用户终端计算任务请求REQ($[\lambda_1,\lambda_2,\cdots,\lambda_n]$);

输出：最优卸载决策$S=X^*$;

$C_i^{\text{edge}}=C_{\text{edge}}/n$;

while TRUE do;

接收用户计算卸载请求REQ，提取请求中的对应任务信息：$B_i^{\text{in}},V_i,B_i^{\text{out}},P_i^{\text{c}},P_i^{\text{up}},\lambda_i$;

for each $i\in\{1,2,\cdots,n\}$ do;

引入拉格朗日函数，求得满足KKT条件的最优解$<x_i,x_i^{\text{edge}},x_i^{\text{cloud}}>$;

最优解向下取整，得整数解$<x'+1_i,x_i^{'\text{edge}},x_i^{'\text{cloud}}>$，$<x'_i,x_i^{'\text{edge}}+1,x_i^{'\text{cloud}}>$，$<x'_i,x_i^{'\text{edge}},x_i^{'\text{cloud}}+1>$;

将整数可行解代入目标函数，取使目标函数最小的整数解为最优整数解;

end for;

回传最优解X^*，移动终端接收卸载决策，执行任务;

end while.

定理2证明问题式(22)为凸优化问题，该问题联合考虑终端计算卸载及边缘节点资源分配，其约束条件为$C_i^{\text{edge}}\geq 0,\sum_{i=1}^{n}C_i^{\text{edge}}\leq C_{\text{edge}}$。本文采用乘子法进行求解，设定初始解为平均资源分配可行解，即3.1节中得到的可行解$S_0=<X_0,C_0>$。

定义其乘子罚函数为

$$\begin{aligned}\varphi(S,\mu,\eta,M)&=f(S)-\sum_{i=1}^{n}\mu_i\cdot h_i(S)\\&+\frac{M}{2}\cdot\sum_{i=1}^{n}{h_i}^2(S)+\frac{1}{2\cdot M}\sum_{j=1}^{4n+1}\left\{\left[\max(0,\eta_j\right.\right.\\&\left.\left.-M\cdot g_j(S))\right]^2-\eta_j^2\right\}\end{aligned}\tag{26}$$

利用拟牛顿法求解，求得BFGS$(\varphi(S,\mu,\eta,M))$为最优解，得到最优S_k^*。此时S_k^*为非整数最优解，再对其取整处理，可求得该最优化问题的最优整数解。求解算法于表2中详细说明。

4　实验与性能分析

利用MATLAB进行实验仿真，模拟分层边缘云计算系统，包含1个云服务器，1个基于WiFi的无线接入点和对应的边缘计算节点，以及多个移动终端设备。设定$N=10$个移动终端设备，随机分布于50 m × 50 m的范围内，无线接入点与边缘计算节点位于区域中心，用户移动终端与边缘节点间信道增益$h_i^{\text{up}}=d_n^{-\varsigma}$，设定$\varsigma=4$，$d_n$取值在[0, 50]之

表 2 多用户计算卸载及资源分配机制

初始化：n, C_i, C_{edge}, C_{cloud}, B，权值V, $S=\varnothing$

输入：各用户终端计算任务请求REQ($[\lambda_1,\lambda_2,\cdots,\lambda_n]$)

输出：最优卸载决策$S=X^*$

$C_i^{\text{edge}}=C_{\text{edge}}/n$, $C_0=<C_1^{\text{edge}},C_2^{\text{edge}},\cdots,C_n^{\text{edge}}>$;

while TRUE do;

接收用户计算卸载请求REQ，提取任务信息：

$B_i^{\text{in}}, V_i, B_i^{\text{out}}, P_i^{\text{c}}, P_i^{\text{up}}, \lambda_i$;

for each $i\in\{1,2,\cdots,n\}$ do;

引入拉格朗日函数，求得满足KKT条件的最优解

$<x_i,x_i^{\text{edge}},x_i^{\text{cloud}}>$;

end for;

得到平均资源分配条件下的初始最优解X^*, $X_0=X^*$;

$S_0=<X_0,C_0>$;

$\mu^{(1)}=(1,1,\cdots,1)$, $\eta^{(1)}=(1,1,\cdots,1)$, $\varepsilon=10^{-5}$, $M=2$,

$\theta=0.8$, $\alpha=2$;

$k=k+1$;

$S_1=\text{BFGS}\left(\varphi\left(S,\mu,\eta,M\right)\right)$;

$$\beta_k=\left\{\sum_{i=1}^{n}h_i^{\,2}\left(S_k\right)+\sum_{j=1}^{4n+1}\left[\left(\min g_j\left(S_k\right),\frac{\left(\eta^{(K)}\right)_j}{M}\right)\right]^2\right\}^{1/2};$$

while $\beta_k>\varepsilon$ do;

更新罚函数：若$\beta_k>\theta\cdot\beta_k$，则$M=\alpha\cdot M$，否则$M$不变；

更新乘子向量$\mu^{(k)}$, $\eta^{(k)}$;

$k=k+1$;

$S_k=\text{BFGS}\left(\varphi\left(S,\mu,\eta,M\right)\right)$;

依据上述公式计算β_k值；

end while;

对$<x_i,x_i^{\text{edge}},x_i^{\text{cloud}}>$求最优整数解，返回$S_k^*=<X_k^*,C_k^*>$，

按X_k^*进行计算卸载，按C_k^*进行计算资源分配；

end while.

间。移动终端P_i^{up}=0.1 W, P_i^{c}取值在[0.1, 0.5]之间，背景噪声$N_0=-174$ dBm/Hz，$B=1$ MHz，终端计算能力C_i取值[0.5, 1.5]之间，设定终端任务量为[500 k, 3 M]之间。而边缘节点与远端云节点计算能力设定为$C_{\text{edge}}=10$ Gcycles/s，$C_{\text{cloud}}=100$ Gcycles/s。与局部执行策略、平均卸载策略，分布式计算卸载策略[13]、共享CAP卸载策略[15]、混合云卸载策略(采用本文3.1机制)和最优化策略(采用本文3.2机制)进行对比分析。

(1)算法性能分析

设定用户任务量及任务复杂度与用户量正相关，即增加的用户任务量及任务复杂度均高于原系统用户。设定目标函数权值V= 0.05，系统在移动终端能耗如图2所示。相较于其它策略本文所提策略能够极大地降低移动终端用户能耗。

设定权值V= 5.00，得到如图3所示实验结果。

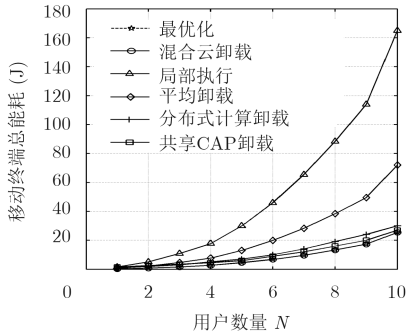

图 2 不同策略下移动终端总能耗变化

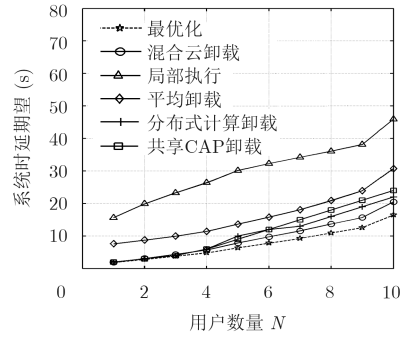

图 3 不同策略下系统时延期望变化

随任务量不断增多，任务计算难度不断提高，本文提出的混合云卸载策略及最优化策略较其它策略增长较为缓慢。在实现计算卸载的情况下，对边缘节点的资源进行合理分配能够有效地减少系统的任务执行时间，充分利用边缘节点计算资源，提高系统性能。

(2)资源分配影响因素研究

为研究混合云卸载及资源分配策略在资源分配上的相关影响因素，考虑5个用户的4种实验场景。场景1任务复杂度线性增长，任务量线性增长；场景2任务复杂度线性增长，任务量相同；场景3任务复杂度相同，任务量线性增长；场景4任务复杂度、任务量均相同。如图4所示，在其它参数相同时，任务当前所需要的计算负载决定用户在边缘节点分配到的资源。

(3)系统时延和能耗均衡

对于无线接入点的决策管理器来说，系统时延和移动终端能耗是主要考虑指标。图5展示了不同权值V下的系统移动终端总能耗变化情况。在最优化策略中，系统移动终端总能耗随权值增长。当V增加时，减少系统时延成为优化的主要目标。随着用户数量及任务量的增加，边缘节点及云节点计算任务加重，计算时间增加，移动终端的执行时间相对变短，故移动终端在任务分配上的比例会对应的有部分增长。

图6为不同权重情况下系统时延与用户数量的

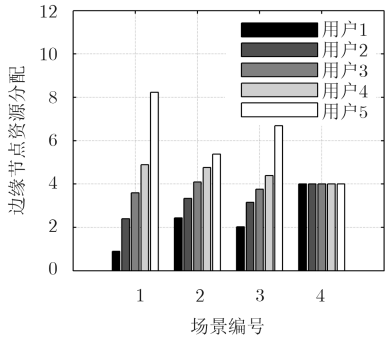

图 4 不同场景下边缘节点资源分配情况

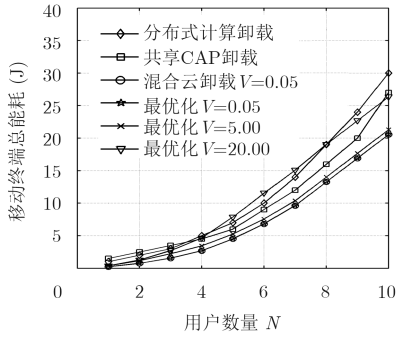

图 5 权重对移动终端总能耗的影响

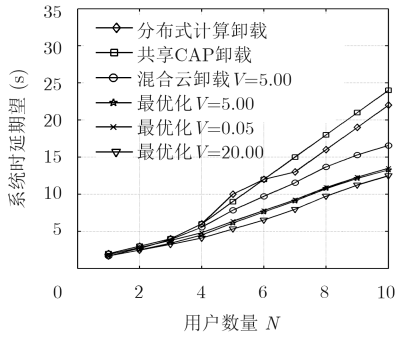

图 6 权值对系统时延期望的影响

变化。随着权值V的增加，在系统用户量相同的情况下，系统时延期望减小。当权值V增加时，减少系统时延成为主要的优化目标，最优化策略适应于减小时延的目标更优化地选择任务在局部设备进行计算或者卸载至边缘及云端。

(4)计算需求与通信需求间比例对卸载决策的影响

本实验在计算需求与通信需求不同比例z下，分析卸载状况，结果如图7所示。当$z<4$时，无卸载发生；当z增大时，卸载量逐渐增多，如$z>20$后，任务基本卸载至边缘端和云端。是由于比值较小时，移动终端可满足资源需求。当比值逐渐增大，移动终端无法满足资源需求时将卸载至远端执行。

5　结束语

融合边缘云与远端云构建了一种边缘云计算

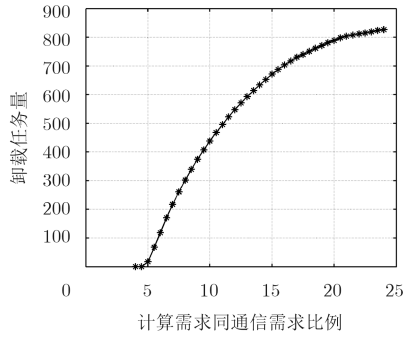

图 7 z的变化对卸载决策的影响

3层架构，利用边缘云及远端云资源增强移动终端的计算能力。依此，分析多用户多任务计算卸载问题，设计了基于乘子法的计算卸载及资源分配机制，卸载计算任务到混合云的同时完成边缘节点的资源分配，实现移动终端能耗及任务执行时延的最小化。通过实验验证最优化策略能够有效地提高系统性能，在降低移动终端能耗及系统时延方面优于局部执行及平均执行策略，并在任务执行时延方面分别降低最高60%与10%。实验结果表明用户计算负载影响边缘节点计算资源的分配，任务计算需求和通信需求之间的比率影响用户计算卸载决策。下一步将考虑融合无线资源分配因素的计算卸载策略，分析移动性对计算卸载的影响，降低移动设备能耗及系统时延。

参 考 文 献

[1] CHEN T Y H, RAVINDRANATH L, DENG Shuo, *et al*. Glimpse: Continuous, real-time object recognition on mobile devices[C]. Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, Seoul, South Korea, 2015: 155–168.

[2] LEE H S and LEE J W. Task offloading in heterogeneous mobile cloud computing: Modeling, analysis, and cloudlet deployment[J]. *IEEE Access*, 2018, 6: 14908–14925. doi: 10.1109/ACCESS.2018.2812144.

[3] VAN DEN BOSSCHE R, VANMECHELEN K, and BROECKHOVE J. Cost-optimal scheduling in hybrid IaaS clouds for deadline constrained workloads[C]. Proceedings of the IEEE 3rd International Conference on Cloud Computing, Miami, USA, 2010: 228–235.

[4] TONG Liang, LI Yong, and GAO Wei. A hierarchical edge cloud architecture for mobile computing[C]. Proceedings of the IEEE INFOCOM 2016- the 35th Annual IEEE International Conference on Computer Communications, San Francisco, USA, 2016: 1–9.

[5] DU Jianbo, ZHAO Liqiang, FENG Jie, *et al*. Computation offloading and resource allocation in mixed fog/cloud computing systems with Min-Max fairness guarantee[J].

IEEE Transactions on Communications, 2018, 66(4): 1594–1608. doi: 10.1109/TCOMM.2017.2787700.

[6] AHMAD A, PAUL A, KHAN M, *et al*. Energy efficient hierarchical resource management for mobile cloud computing[J]. *IEEE Transactions on Sustainable Computing*, 2017, 2(2): 100–112. doi: 10.1109/TSUSC.2017.2714344.

[7] KAO Y H, KRISHNAMACHARI B, RA M R, *et al*. Hermes: Latency optimal task assignment for resource-constrained mobile computing[J]. *IEEE Transactions on Mobile Computing*, 2017, 16(11): 3056–3069. doi: 10.1109/TMC.2017.2679712.

[8] WU Huaming, KNOTTENBELT W, WOLTER K, *et al*. An Optimal Offloading Partitioning Algorithm in Mobile Cloud Computing[M]. Cham, Springer, 2016: 311–328.

[9] DINH T Q, TANG Jianhua, LA Q D, *et al*. Offloading in mobile edge computing: task allocation and computational frequency scaling[J]. *IEEE Transactions on Communications*, 2017, 65(8): 3571–3584. doi: 10.1109/TCOMM.2017.2699660.

[10] MENG Xianling, WANG Wei, and ZHANG Zhaoyang. Delay-constrained hybrid computation offloading with cloud and fog computing[J]. *IEEE Access*, 2017, 5: 21355–21367. doi: 10.1109/ACCESS.2017.2748140.

[11] WANG Yanting, SHENG Min, WANG Xijun, *et al*. Mobile-edge computing: partial computation offloading using dynamic voltage scaling[J]. *IEEE Transactions on Communications*, 2016, 64(10): 4268–4282. doi: 10.1109/TCOMM.2016.2599530.

[12] CHEN Xu, JIAO Lei, LI Wenzhong, *et al*. Efficient multi-user computation offloading for mobile-edge cloud computing[J]. *IEEE/ACM Transactions on Networking*, 2016, 24(5): 2795–2808. doi: 10.1109/TNET.2015.2487344.

[13] CHEN Xu. Decentralized computation offloading game for mobile cloud computing[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2015, 26(4): 974–983. doi: 10.1109/TPDS.2014.2316834.

[14] CARDELLINI V, DE NITTO PERSONÉ V, DI VALERIO V, *et al*. A game-theoretic approach to computation offloading in mobile cloud computing[J]. *Mathematical Programming*, 2016, 157(2): 421–449. doi: 10.1007/s10107-015-0881-6.

[15] CHEN Menghsi, DONG Min, and LIANG Ben. Joint offloading decision and resource allocation for mobile cloud with computing access point[C]. Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 2016: 3516–3520.

代美玲：女，1995年生，博士生，研究方向为移动边缘计算、区块链.

刘周斌：男，1972年生，高级工程师，研究方向为信息安全、能源互联网和分布式系统.

郭少勇：男，1985年生，讲师，研究方向为电力物联网与区块链.

邵苏杰：男，1985年生，讲师，研究方向为网络管理与智能电网，边缘计算.

邱雪松：男，1973年生，教授，博士生导师，研究方向为网络与业务管理.