基于最优运输理论的蜂窝网边缘卸载时延优化研究

吕翔宇，肖泳，钟祎，李强，葛晓虎
（华中科技大学电子信息与通信学院，湖北 武汉 430074）

摘 要：随着物联网的发展，蜂窝网络中接入了大量的用户设备。由于用户设备空间分布和应用需求的变化，需要对用户设备卸载决策进行动态调整。综合考虑网络中用户设备空间分布、应用需求、基站侧边缘服务器的处理能力等参数信息，从分布角度出发，优化用户设备的卸载决策。基于最优运输理论，提出一种时延优化算法。通过合理规划网络中用户设备的卸载基站，降低用户设备计算任务卸载过程的平均时延。仿真结果表明，所提基于时延优化的卸载机制能使平均时延降低 81.06%，并能均衡各基站之间处理的业务量。

关键词：边缘卸载；最优运输理论；时延优化；物联网
中图分类号：TN92
文献标志码：A
doi: 10.11959/j.issn.2096−3750.2023.00352

Research on edge offloading delay optimization of cellular networks based on optimal transport theory

LYU Xiangyu, XIAO Yong, ZHONG Yi, LI Qiang, GE Xiaohu
School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: With the development of the internet of things, a large number of user device (UD) were connected to cellular network. Since the changes in the spatial distribution of UD and application requirements, it is necessary to dynamically adjust the UD' offloading decision. Comprehensively considering various parameter information in the networks such as the spatial distribution of UD, application requirements, and the processing capability of the edge servers on the base station (BS) side, the offloading decision of UD were optimized from the perspective of distribution. Based on the optimal transport theory, a delay optimization algorithm was proposed to reduce the average delay of the UD' computing tasks offloading process by reasonably planning the offloading BS of the UD in the networks. The simulation results show that the average delay can be reduced by 81.06% using the proposed offloading mechanism based on delay optimization, and the traffic handled by each BS is balanced.

Key words: edge offloading, optimal transport theory, delay optimization, IoT

0 引言

近年来，随着无线网络技术的发展，各种对业务时延较为敏感的新兴应用不断涌现，如互动式在线游戏[1]、远程医疗[2]、无人驾驶[3]等。新应用在丰富人们生产生活的同时，也带来了大量的计算密集型和时延敏感型任务。受限于终端的计算资源和电池容量，在时延约束范围内完成对计算任务的处理对终端来说是一个巨大的挑战。计算卸载通过将计算密集型的任务从用户设备迁移到云端服务器处

收稿日期：2023−04−11；**修回日期：**2023−07−02
通信作者：葛晓虎，xhge@mail.hust.edu.cn
基金项目：国家自然科学基金资助项目（No.U2001210，No.62071190）；国家重点研发计划（No.2020YFB1807700）；湖北省重点研发计划（No.2021BAA015）
Foundation Items: The National Natural Science Foundation of China (No.U2001210, No.62071190), The Key Research and Development Program of China (No.2020YFB1807700), The Key Research and Development Program of Hubei Province (No.2021BAA015)

理[4]，可有效解决上述问题。尽管云计算中心具有强大的计算和存储能力，但是用户设备到云计算中心一般具有较长的传输距离，这会增加计算任务处理过程的传输时延[5]。此外，随着接入蜂窝网的用户设备数量急剧增长，将海量数据从用户设备传到云计算中心还可能造成骨干网的拥塞。因此，作为一种新的架构，移动边缘计算（MEC, mobile edge computing）被引入无线网络。与云计算相比，MEC 采用分布式部署的方式，将 MEC 服务器部署在靠近用户设备的网络边缘侧，在提供计算资源的同时，降低了用户设备的计算任务传输时延和传输能耗。

针对计算任务边缘卸载中的时延优化问题，前期已有大量研究工作发表。文献[1]研究了带有能量获取的单用户边缘卸载问题，在传输功率约束下，最小化所有计算任务的平均执行时延。文献[6]研究了基于时分多址的多用户边缘卸载问题，通过联合优化分配给各用户的传输时长和计算任务卸载比例来最小化系统总时延。为了解决用户移动性不可预测对计算任务卸载造成的迁移影响，文献[7]提出了一种基于李雅普诺夫优化的方法，可在长期迁移成本预算约束下，最小化计算任务执行时延。文献[8]研究了多用户设备在存在窃听者的情况下，将计算任务卸载到基站（BS, base station）的问题。通过联合优化用户设备发射功率、MEC 服务器的计算资源分配和用户关联，在安全和计算资源约束下，最小化所有用户设备计算和传输的总时延。

然而上述研究工作仅考虑了计算任务的传输时延和计算时延，考虑计算任务随机到达 MEC 服务器的情况，研究者将排队论引入边缘卸载的研究，文献[9]基于李雅普诺夫优化和随机任务到达模型，研究了用户设备功耗和计算任务执行时延之间的权衡。文献[10]根据马尔可夫链理论研究了双时间尺度随机优化问题，在移动设备平均功率约束下，最小化计算任务的平均时延。文献[11]研究了支持光纤−无线宽带接入的边缘计算架构，其中通信队列被建模成 $M/G/1$ 的轮询模式，计算队列被建模为 $M/M/c$ 的排队模式。面对区块链中视频转码计算量大且耗时的问题，文献[12]提出一种基于区块大小的自适应卸载方案，联合考虑资源分配、卸载调度和自适应块大小，最大化转码器的平均奖励，其中计算队列被建模为 $M/G/c$ 队列。文献[13]研究了具有云计算功能的小蜂窝基站超密集部署情况下，系统长期时延最小化问题，其中通信队列和计算队列被建模为 $M/M/1$ 队列。

上述研究工作中，通常对用户设备卸载决策采用逐个优化的方式求解最优值。然而，面对用户设备数量快速增长带来的高计算复杂度、大量的通信开销以及用户设备空间分布不均匀的特性，部分研究者从分布的角度对边缘卸载问题进行了研究，并且把用户设备的空间分布建模成连续分布。由于用户设备的计算任务卸载过程可以看作数据的运输过程，优化的目标可以看作运输过程的代价，因此将最优运输理论引入无线网络。最优运输理论是研究供给侧和需求侧之间资源最佳匹配的理论，即寻找最优运输方案，使资源以最小代价从供给侧转移到需求侧。最优运输理论在经济学[14-15]、交通运输[16]、人工智能[17-18]、无线网络[19-23]等方面有广泛应用。考虑无人机携带的电池容量对飞行时间的影响以及无人机为网络中多个子区域提供服务，文献[19]通过联合优化无人机服务区域和无人机运行轨迹，最小化无人机总能量。文献[20]研究了在满足地面用户设备负载需求时，通过合理分配带宽和无人机间服务区域，最大限度地减少无人机平均悬停时间。文献[22]研究了三维空间中，无人机基站服务无人机用户的传输数据量最大问题，通过合理划分无人机基站的服务区域，最大化区域内无人机用户传输给无人机基站的数据量。文献[23]研究了多无人机辅助的边缘卸载问题，利用无人机的中继作用，将用户设备的计算任务传输到远处的基站进行处理，基于最优运输理论，合理规划无人机之间的服务范围，最小化上行传输中用户设备和无人机的总功耗。

上述研究虽然将最优运输理论应用于计算任务边缘卸载中的决策优化，但是仅考虑了计算任务的上传时延，忽略了计算任务在基站侧的排队处理时延和回传时延。此外，上述研究均假设用户设备产生的计算任务相同，这与当前用户应用需求的多样化发展趋势不匹配。因此，为了应对上述挑战，本文考虑在多基站提供的服务下，综合考虑用户设备的空间分布、应用需求、MEC 服务器的任务处理能力等系统参数，通过合理规划区域内用户设备的卸载决策，有效降低网络中所有用户设备计算任务卸载过程的平均时延，即系统平均时延，包括计算任务从用户设备到基站的上传时延、计算任务在基站侧被处理的计算时延，以及处理结果从基站侧返回用户设备的回传时延。需要指出的是，在完全卸载的蜂窝网场景下，对于空间分布为离散分布的用

户设备的卸载决策，其表现形式为用户设备与卸载基站的对应。本文考虑空间分布为连续分布的用户设备将计算任务完全卸载到基站侧 MEC 服务器的场景，因此卸载决策表现为用户设备所在的位置与卸载基站的对应。当卸载到同一基站的不同位置连接起来时，形式上表现为基站之间服务区域的划分。本文在以下描述中，用户设备的卸载决策和基站服务区域的划分等价。因此，用户设备的卸载决策最优等价于基站服务区域达到最优划分。

本文的主要贡献如下。

1) 建立了用户设备计算任务卸载过程的时延分析与量化模型。针对用户设备计算任务卸载过程的上传、计算和回传阶段，分别建立相应的时延模型，并且在计算阶段考虑计算任务排队处理情况，根据排队论，建立排队模型。

2) 基于最优运输理论中半连续代价函数性质，提出了蜂窝网中边缘卸载的时延分析和优化算法，所提算法可根据用户设备空间分布、应用需求、MEC 服务器的任务处理能力等系统参数信息，合理划分基站的服务区域，使系统平均时延最小。

3) 仿真结果表明，所提算法具有很好的优化性能和收敛性。与基于最近距离的卸载机制相比，所提卸载机制使系统平均时延降低了 81.06%；与基于随机接入的卸载机制相比，所提卸载机制使系统平均时延降低了 79.28%。

1　系统模型及问题描述

1.1　系统模型

系统模型如图 1 所示，在区域 $\mathcal{D}$ 内分布着 M 个基站和 U 个单天线用户设备，其中 M 个基站随机分布，每个基站配备 N 根天线[24]。用 $\mathcal{M}=\{1,\cdots,i,\cdots,M\}$ 表示 M 个基站构成的集合，其中 BS_i 的位置表示为 (X_i,Y_i)。基站侧部署 MEC 服务器为附近的用户设备提供数据处理服务。BS_i 的任务处理能力由 MEC 服务器内部装载的 CPU 频率 f_i 表征，并用 $\mathcal{F}=\{f_i\}_{i\in\mathcal{M}}$ 表示区域 $\mathcal{D}$ 内所有基站的任务处理能力。用 $\mathcal{U}=\{1,2,\cdots,U\}$ 表示区域 $\mathcal{D}$ 内用户设备的集合。由于用户设备在区域 $\mathcal{D}$ 内分布并不一定是均匀的，部分位置附近用户设备数量较多，而部分位置附近用户设备数量较少。因此从整个区域来看，区域内某位置附近存在较多的用户设备意味着在以该位置为中心的一个极小范围内用户设备密度高，某位置附近存在较少的用户设备意味着在以该位置为中心的一个极小范围内用户设备密度低。为了更具一般性，将区域 $\mathcal{D}$ 内用户设备分布建模为连续分布，并用一个二维分布 $f(x,y)$ 描述区域 $\mathcal{D}$ 内用户设备的分布情况，满足 $\iint_{\mathcal{D}} f(x,y)\mathrm{d}x\mathrm{d}y=1$。换句话说，用 $\mathcal{D}_r(x,y)$ 表示以 (x,y) 为圆心、$r>0$ 为半径的区域，$f(x,y)$ 表示 $\lim\limits_{r\to 0}\mathcal{D}_r(x,y)$ 内的用户设备数占区域 $\mathcal{D}$ 内用户设备总数的比例。取区域 $\mathcal{D}$ 内的典型位置 (x_0,y_0)，用 u_0 表示 (x_0,y_0) 处的用户设备组，用 $u(x_0,y_0)$ 表示 $\mathcal{D}_r(x_0,y_0)$ 内用户设备的个数，所以

$$u(x_0,y_0)=Uf(x_0,y_0) \tag{1}$$

根据基站服务用户设备区域的不同，区域 $\mathcal{D}$ 可以被划分成 M 个子区域，用 $\mathcal{D}_i$ 表示 BS_i 服务的子区域，则 $\mathcal{D}_1\cup\mathcal{D}_2\cup\cdots\cup\mathcal{D}_M=\mathcal{D}$。

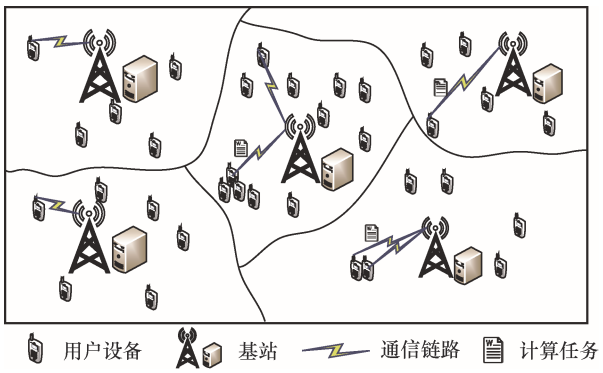


图 1　系统模型

1.2　任务模型

本文考虑完全卸载的任务模式[25]。用户设备的计算任务由单位时间内产生的数据包构成，本文假设用户设备的数据包产生服从泊松过程[26]，且不同的用户设备产生的计算任务优先级相同。用 $\lambda(x_0,y_0)$ 表示用户设备组 u_0 中用户设备产生数据包的平均速率，其中 $\lambda(x_0,y_0)$ 从 $[1,\lambda_{\max}]$ 中随机抽取，$\lambda_{\max}$ 表示用户设备单位时间内产生数据包个数的最大值。用 π_i 表示 BS_i 处数据包的到达速率。因为 BS_i 为区域 $\mathcal{D}_i$ 内的用户设备提供服务，所以 BS_i 处数据包到达速率[13]为

$$\pi_i=\iint_{\mathcal{D}_i}\lambda(x,y)u(x,y)\mathrm{d}x\mathrm{d}y \tag{2}$$

用 $\boldsymbol{\pi}=\{\pi_i\}_{i\in\mathcal{M}}$ 表示区域 $\mathcal{D}$ 内所有基站的数据包到达速率。

由于用户设备可以请求不同类型的任务，所以这些数据包的数据量大小可能各不相同。为了简化系统模型，假设用户设备产生数据包中数据量为服从参数为$\frac{1}{\gamma}$的负指数分布，则数据包的期望大小为γ（单位为 bit）。

1.3 信道模型

由于本文通过对区域$\mathcal{D}$内基站服务区域的合理划分实现系统平均时延最小，所以同一位置用户设备组会整体卸载。为了方便处理，用同一位置每个用户设备到基站的信道增益的平均值代表该位置用户设备到基站的信道增益。具体来说，用$\boldsymbol{h}_i^j(x_0,y_0)$表示用户设备组$u_0$中第$j$个用户设备到$\mathrm{BS}_i$的信道增益，用$\boldsymbol{h}_i(x_0,y_0)$表示$(x_0,y_0)$处的用户设备与$\mathrm{BS}_i$之间的信道增益，则

$$\left\|\boldsymbol{h}_i(x_0,y_0)\right\|_2^2=\frac{1}{u\left(x_0,y_0\right)}\sum_{j=1}^{u(x_0,y_0)}\left\|\boldsymbol{h}_i^j(x_0,y_0)\right\|_2^2 \tag{3}$$

考虑用户设备与基站之间是瑞利信道，接收端采用等增益组合接收，所以用户设备组u_0中第j个用户设备与BS_i之间的信道增益为[27]

$$\left\|\boldsymbol{h}_i^j(x_0,y_0)\right\|_2^2=\boldsymbol{\varphi}_{i0}^j\boldsymbol{\beta}_{i0}^j \tag{4}$$

其中，$\boldsymbol{\varphi}_{i0}^j=\left[\varphi_{i0}^{j,1},\cdots,\varphi_{i0}^{j,n},\cdots\varphi_{i0}^{j,N}\right]\in\mathbb{R}^{1\times N}$是小尺度衰落系数矩阵，$\varphi_{i0}^{j,n}$表示用户设备组$u_0$中第$j$个用户设备到$\mathrm{BS}_i$的第$n$根天线的小尺度衰落系数，大小服从参数为$\vartheta$的瑞利分布[28]，$\varphi_{i0}^{j,n}$的概率密度函数可表示为

$$f(\varphi_{i0}^{j,n})=\frac{\varphi_{i0}^{j,n}}{\vartheta^2}\mathrm{e}^{-\frac{(\varphi_{i0}^{j,n})^2}{2\vartheta^2}},\varphi_{i0}^{j,n}>0 \tag{5}$$

其中，$\boldsymbol{\beta}_{i0}^j=\left[{d_{i0,1}^j}^{-\alpha},\cdots,{d_{i0,n}^j}^{-\alpha},\cdots,{d_{i0,N}^j}^{-\alpha}\right]^{\mathrm{T}}\in\mathbb{R}^{N\times 1}$是大尺度衰落系数矩阵，其中，$\alpha$是路径损耗指数，$d_{i0,n}^j$是用户设备组$u_0$中第$j$个用户设备到$\mathrm{BS}_i$的第$n$根天线的欧氏距离。由于天线之间的距离相较于用户设备到基站之间的距离可忽略不计，所以用户设备到同一基站的所有天线距离相同，即${d_{i0}^j}^{-\alpha}={d_{i0,1}^j}^{-\alpha}=\cdots={d_{i0,N}^j}^{-\alpha}$，所以，用户设备组$u_0$中第$j$个用户设备与$\mathrm{BS}_i$之间的信道增益可写为

$$\left\|\boldsymbol{h}_i^j\left(x_0,y_0\right)\right\|_2^2=\boldsymbol{\varphi}_{i0}^j\boldsymbol{\beta}_{i0}^j=\left[\varphi_{i0}^{j,1},\cdots,\varphi_{i0}^{j,n},\cdots,\varphi_{i0}^{j,N}\right]\begin{bmatrix}{d_{i0}^j}^{-\alpha}\\ \vdots\\ {d_{i0}^j}^{-\alpha}\end{bmatrix} \tag{6}$$

1.4 问题描述

用户设备将计算任务卸载到基站进行处理的过程包括上传、计算和回传 3 个阶段，由于基站为每个用户设备分别提供了正交的上下行信道，所以上下行传输可以同时进行。也就是说，当用户设备通过带宽为B_{up}的上行信道将计算任务传输到基站的同时，基站可以通过带宽为B_{do}的下行信道将之前的任务处理结果无等待地传回用户设备。本文假设用户设备之间采用正交频分多址接入的方式实现卸载过程[29]，则位于(x_0,y_0)处的用户设备组u_0中第j个用户设备将计算任务卸载到BS_i处理过程的平均总时延为

$$\begin{aligned}t_{\mathrm{total},i}^j(x_0,y_0)=&t_{\mathrm{up},i}^j(x_0,y_0)+\\&t_{\mathrm{com},i}^j(x_0,y_0)+t_{\mathrm{do},i}^j(x_0,y_0)\end{aligned} \tag{7}$$

其中，$t_{\mathrm{up},i}^j(x_0,y_0)$表示用户设备组$u_0$中第$j$个用户设备将计算任务传输到$\mathrm{BS}_i$的上传时延，$t_{\mathrm{com},i}^j(x_0,y_0)$表示用户设备组$u_0$中第$j$个用户设备的计算任务在$\mathrm{BS}_i$处理的计算时延，$t_{\mathrm{do},i}^j(x_0,y_0)$表示用户设备组$u_0$中第$j$个用户设备的计算结果从$\mathrm{BS}_i$的回传时延。

在计算任务从用户设备到基站的上传阶段，假设每个用户设备以固定功率P_{u}发送，根据香农定理和式(3)，用户设备组u_0中用户设备到BS_i的平均信道容量为

$$C_i(x_0,y_0)=B_{\mathrm{up}}\mathrm{lb}\left(1+\frac{P_{\mathrm{u}}\left\|\boldsymbol{h}_i(x_0,y_0)\right\|_2^2}{N_0}\right) \tag{8}$$

其中，N_0是噪声功率。

用户设备组u_0中第j个用户设备将计算任务传输到BS_i的上传阶段的平均时延为[13]

$$t_{\mathrm{up},i}^j(x_0,y_0)=\frac{\lambda(x_0,y_0)\gamma}{C_i(x_0,y_0)} \tag{9}$$

当计算任务到达基站后，考虑处理单比特数据需要ω次 CPU 周期。由于单个数据包的数据量大小服从参数为$\frac{1}{\gamma}$的负指数分布，在 MEC 服务器恒定处理速率下，BS_i侧的 MEC 服务器处理单个数据包的时间服从参数为$\frac{f_i}{\gamma\omega}$的负指数分布。由于数据包达到基站的过程是泊松过程，将每个 MEC 服

务器的任务处理方式建模为 $M/M/1$ 排队过程[30]。根据式(2)，BS_i 侧的 MEC 服务器处理单个数据包的平均时延为

$$t_{i,\mathrm{delay}}=\frac{1}{\dfrac{f_i}{\gamma\omega}-\pi_i} \tag{10}$$

用户设备组 u_0 中第 j 个用户设备的计算任务在计算阶段被处理的平均时延为

$$t^j_{\mathrm{com},i}(x_0,y_0)=\lambda(x_0,y_0)t_{i,\mathrm{delay}} \tag{11}$$

当计算任务在基站被处理后，假设 BS_i 的下行传输功率为 P^i_{d}，回传结果的数据量与上传任务的数据量之间的比例为 δ，则结果从 BS_i 回传给用户设备组 u_0 中第 j 个用户设备的平均时延为

$$t^j_{\mathrm{do},i}(x_0,y_0)=\frac{\delta\lambda(x_0,y_0)\gamma}{B_{\mathrm{do}}\mathrm{lb}\left(1+\dfrac{P^i_{\mathrm{d}}\left\|\boldsymbol{h}_i(x_0,y_0)\right\|_2^2}{N_0}\right)} \tag{12}$$

根据式(7)、式(9)、式(11)和式(12)，用户设备组 u_0 中第 j 个用户设备将计算任务卸载到 BS_i 处理过程的平均总时延为

$$t^j_{\mathrm{total},i}(x_0,y_0)=\frac{\lambda(x_0,y_0)\gamma}{B_{\mathrm{up}}\mathrm{lb}\left(1+\dfrac{P_{\mathrm{u}}\left\|\boldsymbol{h}_i(x_0,y_0)\right\|_2^2}{N_0}\right)}+\frac{\lambda(x_0,y_0)}{\dfrac{f_i}{\gamma\omega}-\pi_i}+\frac{\delta\lambda(x_0,y_0)\gamma}{B_{\mathrm{do}}\mathrm{lb}\left(1+\dfrac{P^i_d\left\|\boldsymbol{h}_i(x_0,y_0)\right\|_2^2}{N_0}\right)} \tag{13}$$

根据式(13)，位于 (x_0,y_0) 处的用户设备组 u_0 中所有用户设备将计算任务卸载到 BS_i 处理过程的总时延为

$$t_{\mathrm{total},i}(x_0,y_0)=t^j_{\mathrm{total},i}(x_0,y_0)Uf(x_0,y_0) \tag{14}$$

根据式(14)，被 BS_i 服务的用户设备计算任务卸载过程的总时延为

$$t_i=\iint_{\mathcal{D}_i}t_{\mathrm{total},i}(x,y)\mathrm{d}x\mathrm{d}y \tag{15}$$

区域 $\mathcal{D}$ 内所有用户设备计算任务卸载过程的总时延为

$$t_{\mathrm{total}}=\sum_{i=1}^{M}t_i \tag{16}$$

根据式(16)，系统平均时延最小化问题可表述为

P1:

$$\min_{\mathcal{D}_i}\sum_{i=1}^{M}\iint_{\mathcal{D}_i}\left(\frac{\lambda(x,y)\gamma}{B_{\mathrm{up}}\mathrm{lb}\left(1+\dfrac{P_{\mathrm{u}}\left\|\boldsymbol{h}_i(x,y)\right\|_2^2}{N_0}\right)}+\frac{\lambda(x,y)}{\dfrac{f_i}{\gamma\omega}-U\iint_{\mathcal{D}_i}\lambda(x,y)f(x,y)\mathrm{d}x\mathrm{d}y}+\frac{\delta\lambda(x,y)\gamma}{B_{\mathrm{do}}\mathrm{lb}\left(1+\dfrac{P^i_d\left\|\boldsymbol{h}_i(x,y)\right\|_2^2}{N_0}\right)}\right)f(x,y)\mathrm{d}x\mathrm{d}y \tag{17}$$

$$\text{s.t.}\quad \mathcal{D}_p\bigcap\mathcal{D}_q=\varnothing\quad,\forall p\neq q\in\mathcal{M} \tag{17a}$$

$$\bigcup_{i\in\mathcal{M}}\mathcal{D}_i=\mathcal{D} \tag{17b}$$

约束式(17a)保证划分后的基站服务区域不会相互重叠，即用户设备选择一个基站进行任务卸载；约束式(17b)保证区域内的用户设备能被完全服务。求解问题 P1 是有难度的，首先优化变量 $\mathcal{D}_i(\forall i\in\mathcal{M})$ 是相互依赖的连续变量，其次为了更加准确地拟合用户设备的空间分布，$f(x,y)$ 被认为是 x 和 y 的泛型函数，这些因素导致给定二重积分的复杂性。由于最优运输理论研究从一个分布到另一个分布的最小代价问题，所以，本文引入最优运输理论对其进行求解。

2　基于最优运输理论的时延优化

2.1　最优运输理论

最优运输问题（也称蒙日问题）最早由法国数学家加斯帕德 • 蒙日于 1781 年提出[31]，主要研究当沙堆和沙坑体积相同的情况下，如何用最小的代价将沙子从沙堆搬到沙坑。从数学角度来看，最优运输问题可以抽象为一个概率分布转换成另一个概率分布所需的最小代价。其中这两个概率分布可以是离散的，也可以是连续的，并根据分布不同分成 3 种情况，其中从一个连续分布转换成一个离散分布的最优运输问题称为半离散最优运输问题。

蒙日问题用数学语言可表述为，给定 polish 空间上的两个可分的、完备的度量空间 $X\subset R^n$ 、

$Y \subset R^m$，R^m、R^n 代表 m、n 维空间，以及对应空间上的概率分布 $f_1 \in \mathcal{P}(X)$、$f_2 \in \mathcal{P}(Y)$，其中 $\mathcal{P}(X)$ 代表 X 上所有概率分布构成的空间，$\mathcal{P}(Y)$ 代表 Y 上所有概率分布构成的空间。对于任意可测集 $A \subset Y$，寻找从 $X \to Y$ 传输变换 F，使服从分布 f_1 的随机向量 $\boldsymbol{x}$ 变换成服从 f_2 的随机向量 $\boldsymbol{y}$，同时使变换的总代价最小。数学表达式为

$$\min_F \int_X c\left(\boldsymbol{x}, F\left(\boldsymbol{x}\right)\right) f_1\left(\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} \tag{18}$$

$$\text{s.t.} \quad \int_A f_2\left(\boldsymbol{y}\right) \mathrm{d}\boldsymbol{y} = \int_{F^{-1}\left(A\right)} f_1\left(\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} \quad , \ \forall A \subset Y \tag{18a}$$

其中，$c\left(\boldsymbol{x}, F\left(\boldsymbol{x}\right)\right)$ 表示从源随机向量 $\boldsymbol{x}$ 变换到对应目的随机向量 $\boldsymbol{y} = F\left(\boldsymbol{x}\right)$ 的单位代价，$F^{-1}(A) = \left\{\boldsymbol{x} \mid \boldsymbol{x} \in X, F(\boldsymbol{x}) \in A\right\}$，约束条件通常简记为 $F_{\#} f_1 = f_2$，并将 $F_{\#}$ 称为推前测度[32]。

解决蒙日问题是具有挑战性的，首先蒙日问题具有高度的非线性[33]，其次蒙日问题要求源分布中的每个点只映射到目的分布的一个位置，因此最优映射不一定存在。后来经过列奥尼德·康托罗维奇的研究[20]，将映射方案松弛为运输方案，使源分布中的每个点都可以变换到目的分布的多个点。松弛后的蒙日问题称为康托罗维奇问题。康托罗维奇问题用数学语言可表述为，给定 polish 空间上的两个可分的、完备的度量空间 $X \subset R^n$、$Y \subset R^m$，R^m、R^n 代表 m、n 维空间，以及对应空间上的概率分布 $f_1 \in \mathcal{P}(X)$、$f_2 \in \mathcal{P}(Y)$，其中 $\mathcal{P}(X)$ 代表 X 上所有概率分布构成的空间，$\mathcal{P}(Y)$ 代表 Y 上所有概率分布构成的空间。两个概率分布之间的所有运输方案即联合概率分布可统一表示为 $\boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{y})$。寻找最优运输方案 $\boldsymbol{\pi}^*(\boldsymbol{x}, \boldsymbol{y})$，使得从源分布 f_1 变换到目的分布 f_2 的总代价最小。数学表达式为

$$\min_{\pi} \int_{X \times Y} c(\boldsymbol{x}, \boldsymbol{y}) \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \tag{19}$$

$$\text{s.t.} \int_Y \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} = f_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \tag{19a}$$

$$\int_X \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} = f_2(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \tag{19b}$$

其中，$c(\boldsymbol{x}, \boldsymbol{y})$ 表示从源随机向量 $\boldsymbol{x}$ 变换到目标随机向量 $\boldsymbol{y}$ 的单位代价。可以注意到，运输方案总是非空的。

和蒙日问题相比，康托罗维奇问题有 3 个好处。首先由于康托罗维奇问题是蒙日问题的弱化和推广，所以当蒙日问题的最优解存在时，该解也是康托罗维奇问题最优解。其次，康托罗维奇问题对任何半连续代价函数存在最优解决方案[20]。最后，康托罗维奇问题存在对偶公式，可以得到一个易于处理的解。值得注意的是，当源分布函数与代价函数连续时，蒙日问题和康托罗维奇问题等价[34]。

2.2　时延优化算法

本文通过对基站之间服务区域进行优化，来降低系统平均时延。区域 $\mathcal{D}$ 内，用户设备的空间分布被建模成连续分布，基站的空间分布呈离散分布，因此，用户设备将计算任务卸载到基站处理的过程，可以建模成半离散最优运输过程，而用户设备的计算任务卸载时延可以看作运输过程中的代价。作为源分布的用户设备空间分布函数为 $f(x,y)$，作为目的分布的基站空间分布函数可表示为

$$\Theta = \sum_{i=1}^{M} \kappa_i \delta_i \tag{20}$$

其中，κ_i 代表 BS_i 的归一化计算任务处理能力，即 BS_i 处理区域 $\mathcal{D}_i$ 内计算任务数据包量占区域 $\mathcal{D}$ 内计算任务数据包总量的值，满足 $\sum_{i=1}^{M} \kappa_i = 1$；$\delta_i$ 即 $\delta_i(x,y)$，为狄拉克函数。

$$\delta_i(x,y) = \begin{cases} 1 & ,(x,y) = (X_i, Y_i) \\ 0 & ,(x,y) \neq (X_i, Y_i) \end{cases} \tag{21}$$

首先对问题 P1 给出最优基站服务区域存在定理。

对问题 P1 进行等价变形，可转变如下。

P1-1:

$$\min_{\mathcal{D}_i} \sum_{i=1}^{M} \iint_{\mathcal{D}_i} \mathcal{C}_i(x,y) \mathcal{O}(x,y) \mathrm{d}x \mathrm{d}y \tag{22}$$

$$\text{s.t.} \quad \mathcal{D}_p \bigcap \mathcal{D}_q = \varnothing \quad , \forall p \neq q \in \mathcal{M} \tag{22a}$$

$$\bigcup_{i \in \mathcal{M}} \mathcal{D}_i = \mathcal{D} \tag{22b}$$

$$\mathcal{C}_i(x,y) = \frac{\gamma}{B_{\mathrm{up}} \mathrm{lb}\left(1 + \dfrac{P_{\mathrm{u}} \left\| \boldsymbol{h}_i(x,y) \right\|_2^2}{N_0}\right)} + \frac{1}{\dfrac{f_i}{\gamma \omega} - U s_i} + \frac{\delta \gamma}{B_{\mathrm{do}} \mathrm{lb}\left(1 + \dfrac{P_{\mathrm{d}}^i \left\| \boldsymbol{h}_i(x,y) \right\|_2^2}{N_0}\right)} \tag{22c}$$

$$\mathcal{O}(x,y) = \lambda(x,y) f(x,y) \tag{22d}$$

$$s_i = \iint_{\mathcal{D}_i} \lambda(x,y) f(x,y) \mathrm{d}x \mathrm{d}y \tag{22e}$$

引理 1：考虑 polish 空间上两个概率分布 f 和 Θ，f 为连续概率分布，$\Theta=\sum_{i=1}^{M}\kappa_i\delta_i$ 为离散概率分布，其中 δ_i 是狄拉克函数，κ_i 表示取到 i 的概率。对于任何半连续代价函数，存在从 $f\to\Theta$ 的最优映射，使得 $\int_{\mathcal{D}}c\left(x,T(x)\right)f\left(x\right)\mathrm{d}x$ 最小化[20,35]。

定理 1：对于问题 P1-1，区域 $\mathcal{D}$ 内，空间分布服从连续分布 $f(x,y)$ 的用户设备将计算任务卸载到空间分布服从离散分布 $\Theta=\sum_{i=1}^{M}\kappa_i\delta_i$ 的基站处理时，存在最优基站服务区域。

证明：记 $\kappa_i=\dfrac{\iint_{\mathcal{D}_i}\mathcal{O}\left(x,y\right)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}}\mathcal{O}\left(x,y\right)\mathrm{d}x\mathrm{d}y}$。定义一个单位单纯形如

$$\mathcal{K}=\left\{\boldsymbol{\kappa}=\left(\kappa_1,\kappa_2,\cdots,\kappa_{\mathcal{M}}\right)\in\mathbb{R}^M;\sum_{i=1}^{M}\kappa_i=1,\kappa_i\geqslant 0,\forall i\in\mathcal{M}\right\} \tag{23}$$

因为基站的最优服务区域划分是在确定用户设备空间分布 $f(x,y)$ 和用户设备计算任务数据包产生速率分布 $\lambda\left(x,y\right)$ 的情况下进行的，所以对于给定的 $\lambda(x,y)$、$f(x,y)$，$\iint_{\mathcal{D}}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y$ 是一个定值。对于任意给定的 $\boldsymbol{\kappa}$，$s_i=\kappa_i\iint_{\mathcal{D}}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y$（$\forall i\in\mathcal{M}$），也就是说对任意的 i，s_i 和 κ_i 一一对应。

由于 $\boldsymbol{\kappa}$ 给定，也就是 $s_i(\forall i\in\mathcal{M})$ 确定，所以问题 P1-1 中的代价函数 $\mathcal{C}_i(x,y)$ 是关于 (x,y) 的连续函数。又因为所有连续函数都是半连续函数，所以根据引理 1，存在从 $f\to\Theta$ 的最优映射，使问题 P1-1 有解，所以问题 P1-1 存在最优基站服务区域，即问题 P1 存在最优基站服务区域。证毕。

在求解问题 P1 前，首先形式进行变化如下。

P1-2:

$$\min_{\mathcal{D}_i}\sum_{i=1}^{M}\left(\iint_{\mathcal{D}_i}Q_i(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+\frac{s_i}{\dfrac{f_i}{\gamma\omega}-Us_i}\right) \tag{24}$$

$$\text{s.t.}\quad \mathcal{D}_p\cap\mathcal{D}_q=\varnothing\quad,\forall p\neq q\in\mathcal{M} \tag{24a}$$

$$\bigcup_{i\in\mathcal{M}}\mathcal{D}_i=\mathcal{D} \tag{24b}$$

$$Q_i(x,y)=\frac{\gamma}{B_{\mathrm{up}}\mathrm{lb}\left(1+\dfrac{P_{\mathrm{u}}\left\|\boldsymbol{h}_i(x,y)\right\|_2^2}{N_0}\right)}+\frac{\delta\gamma}{B_{\mathrm{do}}\mathrm{lb}\left(1+\dfrac{P_{\mathrm{d}}^i\left\|\boldsymbol{h}_i(x,y)\right\|_2^2}{N_0}\right)} \tag{24c}$$

$$\mathcal{O}(x,y)=\lambda(x,y)f(x,y) \tag{24d}$$

$$s_i=\iint_{\mathcal{D}_i}\lambda(x,y)f(x,y)\mathrm{d}x\mathrm{d}y \tag{24e}$$

定理 2：对于问题 P1-2，区域 $\mathcal{D}$ 内，空间分布服从连续分布 $f(x,y)$ 的用户设备将计算任务卸载到空间分布服从离散分布 $\Theta=\sum_{i=1}^{M}\kappa_i\delta_i$ 的基站处理时，系统平均时延最小为

$$t_{\mathrm{OT}}=\sum_{i=1}^{M}\left(\iint_{\mathcal{D}_i^*}Q_i(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+\frac{s_i^*}{\dfrac{f_i}{\gamma\omega}-Us_i^*}\right) \tag{25}$$

其中，$\mathcal{D}_i^*$ 是最优基站服务区域中 BS_i 所服务的区域，由式(26)给出。

$$\mathcal{D}_i^*=\left\{\begin{array}{l}(x,y)\,|\,Q_i(x,y)+\dfrac{\dfrac{f_i}{\gamma\omega}}{\left(\dfrac{f_i}{\gamma\omega}-Us_i\right)^2}\\ <Q_j(x,y)+\dfrac{\dfrac{f_j}{\gamma\omega}}{\left(\dfrac{f_j}{\gamma\omega}-Us_j\right)^2},\forall i\neq j\in\mathcal{M}\end{array}\right\} \tag{26}$$

证明：根据定理 1，问题 P1 存在最优基站服务区域 $\mathcal{D}_i(\forall i\in\mathcal{M})$。考虑其中两个最优分区 $\mathcal{D}_k$ 和 $\mathcal{D}_l$，以及区域 $\mathcal{D}_k$ 内的一个点 e_0，坐标为 $\left(x_{k0},y_{k0}\right)$。以 e_0 为圆心，作半径为 r 的圆，并限制圆域完全在区域 $\mathcal{D}_k$ 内。用 $\mathcal{D}_r(e_0)$ 表示以 e_0 为圆心、半径为 r 的圆域，则 $\mathcal{D}_r(e_0)\subset\mathcal{D}_k$。根据选取的圆域，生成新的小区划分如下

$$\begin{cases}\widetilde{\mathcal{D}_k}=\mathcal{D}_k\setminus\mathcal{D}_r(e_o)\\ \widetilde{\mathcal{D}_l}=\mathcal{D}_l\cup\mathcal{D}_r(e_o)\\ \widetilde{\mathcal{D}_i}=\mathcal{D}_i,i\neq l,k\end{cases} \tag{27}$$

用$s_r=\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y$和$\tilde{s}_i=\iint_{\tilde{\mathcal{D}}_i}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y$，并记$g(s)=\dfrac{s}{\dfrac{f_i}{\gamma\omega}-Us}$。由于$\mathcal{D}_i(\forall i\in\mathcal{M})$是最优基站服务区域，所以

$$\begin{aligned}&\sum_{i=1}^{M}\left(\iint_{\mathcal{D}_i}Q_i(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+g(s_i)\right)\\&\overset{(\Delta)}{\leqslant}\sum_{i=1}^{M}\left(\iint_{\tilde{\mathcal{D}}_i}Q_i(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+g\left(\tilde{s}_i\right)\right)\end{aligned}$$

$$\begin{aligned}&\Leftrightarrow\iint_{\mathcal{D}_l}Q_l(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+g(s_l)+\\&\iint_{\mathcal{D}_k}Q_k(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+g(s_k)\leqslant\\&\iint_{\tilde{\mathcal{D}}_l}Q_l(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+g\left(\tilde{s}_l\right)+\\&\iint_{\tilde{\mathcal{D}}_k}Q_k(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+g\left(\tilde{s}_k\right)\\&\Leftrightarrow\iint_{\mathcal{D}_r(e_0)}Q_k(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+g(s_k)-g\left(\tilde{s}_k\right)\leqslant\\&\iint_{\mathcal{D}_r(e_0)}Q_l(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+g\left(\tilde{s}_l\right)-g(s_l)\end{aligned}\tag{28}$$

其中，$\overset{(\Delta)}{\leqslant}$表示$(\mathcal{D})_{i=1,\cdots,\mathcal{M}}$是最优基站服务区域划分，因此其他区域划分$\left(\tilde{\mathcal{D}}\right)_{i=1,\cdots,\mathcal{M}}$所产生的代价都不小于最优基站服务区域划分。不等式两边同时除以s_r，并求出当$r\to 0$的极限。

$$\begin{aligned}&\lim_{r\to0}\frac{1}{s_r}\iint_{\mathcal{D}_r(e_0)}Q_k(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y=\\&\lim_{r\to0}\frac{\iint_{\mathcal{D}_r(e_0)}Q_k(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}\end{aligned}\tag{29a}$$

$$\begin{aligned}&\lim_{r\to0}\frac{1}{s_r}\iint_{\mathcal{D}_r(e_0)}Q_l(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y=\\&\lim_{r\to0}\frac{\iint_{\mathcal{D}_r(e_0)}Q_l(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}\end{aligned}\tag{29b}$$

$$\lim_{r\to0}\frac{1}{s_r}\left(g(s_k)-g\left(\tilde{s}_k\right)\right)=\lim_{r\to0}\frac{g(s_k)-g\left(\tilde{s}_k\right)}{s_r}\tag{29c}$$

$$\lim_{r\to0}\frac{1}{s_r}\left(g\left(\tilde{s}_l\right)-g(s_l)\right)=\lim_{r\to0}\frac{g\left(\tilde{s}_l\right)-g(s_l)}{s_r}\tag{29d}$$

引理 2：如果$f(x,y)$在有界闭区域$\mathcal{D}$上连续，函数$g(x,y)$在$\mathcal{D}$上可积且不变号，则存在一个点$(\xi,\eta)\in\mathcal{D}$使式(30)成立[36]。

$$\iint_{\mathcal{D}}f(x,y)g(x,y)\mathrm{d}x\mathrm{d}y=f(\xi,\eta)\iint_{\mathcal{D}}g(x,y)\mathrm{d}x\mathrm{d}y\tag{30}$$

根据引理 2，即二重积分中值定理的推广，式(29a)、式(29b)可得到

$$\begin{aligned}&\lim_{r\to0}\frac{\iint_{\mathcal{D}_r(e_0)}Q_k(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}=\\&\lim_{r\to0}\frac{Q_k(\mu_1,\eta_1)\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}\end{aligned}\tag{31a}$$

$$\begin{aligned}&\lim_{r\to0}\frac{\iint_{\mathcal{D}_r(e_0)}Q_l(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)dxdy}=\\&\lim_{r\to0}\frac{Q_l(\mu_2,\eta_2)\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}\end{aligned}\tag{31b}$$

其中，$(\mu_1,\eta_1)\in\mathcal{D}_r(e_0)$，$(\mu_2,\eta_2)\in\mathcal{D}_r(e_0)$。

根据函数$Q_i(x,y)(i\in\{k,l\})$的连续性和式(31a)、式(31b)可得

$$\begin{aligned}&\lim_{r\to0}\frac{Q_k(\mu_1,\eta_1)\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}=\\&Q_k(x_{k0},y_{k0})\end{aligned}\tag{32a}$$

$$\begin{aligned}&\lim_{r\to0}\frac{Q_l(\mu_2,\eta_2)\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}{\iint_{\mathcal{D}_r(e_0)}\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y}=\\&Q_l(x_{k0},y_{k0})\end{aligned}\tag{32b}$$

根据$g(s)$是连续可微、非递减的凸函数和式(29c)、式(29d)可得

$$\lim_{r\to0}\frac{g(s_k)-g\left(\tilde{s}_k\right)}{s_r}=g'(s_k)\tag{32c}$$

$$\lim_{r\to 0}\frac{g\left(\tilde{s}_l\right)-g(s_l)}{s_r}=g'(s_l) \tag{32d}$$

将式(32a)、式(32b)、式(32c)和式(32d)代入式(28)，得到

$$Q_k(x_{k0},y_{k0})+g'\left(s_k\right)<Q_l(x_{k0},y_{k0})+g'(s_l) \tag{33}$$

也就是说，在最优分配下，点 e_0 被分配给区域 $\mathcal{D}_k$ 而非区域 $\mathcal{D}_l$ 时应满足式(33)。

因此， BS_i 最优服务区域可被表述为

$$\mathcal{D}_i^*=\left\{\begin{array}{l}(x,y)\,|\,Q_i(x,y)+\dfrac{\dfrac{f_i}{\gamma\omega}}{\left(\dfrac{f_i}{\gamma\omega}-Us_i\right)^2}\\ <Q_j(x,y)+\dfrac{\dfrac{f_j}{\gamma\omega}}{\left(\dfrac{f_j}{\gamma\omega}-Us_j\right)^2},\forall i\neq j\in\mathcal{M}\end{array}\right\} \tag{34}$$

把式(34)代入问题 P1-2 可得区域 $\mathcal{D}$ 内系统平均时延最小为

$$t_{\mathrm{OT}}=\sum_{i=1}^{M}\left(\iint_{\mathcal{D}_i^*}Q_i(x,y)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y+\frac{s_i^*}{\dfrac{f_i}{\gamma\omega}-Us_i^*}\right) \tag{35}$$

证毕。

从式(26)中看到，对 $\forall i\in\mathcal{M}$， s_i 和 $\mathcal{D}_i$ 之间存在相互依赖，所以求解式(26)没有明确的形式，因此根据文献[35]提出一种在有限迭代次数内就能收敛到最优解的迭代算法。通过求解式(26)，找到最优的基站服务区域和系统平均时延的最小值。基于最优运输理论的时延优化算法如算法 1 所示。

算法 1　基于最优运输理论的时延优化算法

输入　用户设备分布 $f(x,y)$，用户设备数量 U，基站数量 M，不同位置用户设备产生的计算任务数据包量 $\lambda(x,y)$，计算任务数据包中数据量 γ，用户设备发射功率 P_{u}，处理单比特数据所需的 CPU 周期数 ω，基站的位置 $\left\{(X_i,Y_i)\right\}_{i\in\mathcal{M}}$，基站的发射功率 $\left\{P_{\mathrm{d}}^i\right\}_{i\in\mathcal{M}}$，基站的计算任务处理能力 $\left\{f_i\right\}_{i\in\mathcal{M}}$，路径损耗指数 α，瑞利分布标准差 ϑ，上下行信道传输带宽 B_{up}、B_{do}，有效电容系数 ξ，返回结果比例 δ，设置迭代标识 $z=1$ 和误差 ε

输出　最优区域划分 $\mathcal{D}_i^*(\forall i\in\mathcal{M})$，系统平均时延最小值 t_{OT}

初始化　子区域 $\mathcal{D}_i^{(z)}$，设置参数 $\Phi_i^{(z)}(x,y)=0$；

while

对于每一个 $\mathrm{BS}_i(i\in\mathcal{M})$，如果上一次该点在 BS_i 服务范围内，则利用 $\Phi_i^{(z+1)}(x,y)=\left(1-\dfrac{1}{z}\right)\Phi_i^{(z)}(x,y)$ 更新；如果不在该 BS_i 服务范围内，则利用 $\Phi_i^{(z+1)}(x,y)=1-\left(1-\dfrac{1}{z}\right)\left(1-\Phi_i^{(z)}(x,y)\right)$ 更新；

根据更新的参数 $\Phi_i^{(z+1)}(x,y)$ 计算每一个基站的任务量 $s_i=\iint_{\mathcal{D}}\left(1-\Phi_i^{(z+1)}(x,y)\right)\mathcal{O}(x,y)\mathrm{d}x\mathrm{d}y\ (\forall i\in\mathcal{M})$；

迭代次数更新 $z=z+1$；

利用式(26)更新子区域的范围；

利用式(25)计算系统平均时延 $t_{\mathrm{OT}}^{(z)}$；

Until $\left|t_{\mathrm{OT}}^{(z)}-t_{\mathrm{OT}}^{(z-1)}\right|<\varepsilon$；

得到的最优基站服务区域 $\mathcal{D}_i^*=\mathcal{D}_i^{(z)}(\forall i\in\mathcal{M})$ 和系统平均时延最小值 $t_{\mathrm{OT}}=t_{\mathrm{OT}}^{(z)}$。

算法 1 的核心思想利用了用户设备的谨慎自利性，即关注整个系统代价最小。对于本文所研究的用户设备计算任务卸载而言，关注系统平均时延最小。在算法的第一步，对不同位置的用户设备组选择卸载的基站进行随机初始化，本文选择的初始化是用户设备基于最近距离的卸载机制进行计算任务卸载。该过程受用户设备的空间分布和计算任务的空间分布的影响，造成部分基站负载过大，导致计算任务处理时间过长，因此根据式(26)，用户设备会选择使自身代价更小的基站进行卸载。但是这可能会造成一个结果，即负载较轻的基站由于被大量用户设备卸载任务而成为负载较重的基站，而原有的负载较重的基站因大量用户设备离开而变成负载较轻的基站。为了避免自利性使用户设备在基站之间来回选择，引入了谨慎参数 $\boldsymbol{\Phi}(x,y)$，即用户设备在下一次选择中对改变的谨慎。迭代算法对于区域内每一个基站，当用户设备在基站服务范围内时，用户离开该基站的意愿为 $\boldsymbol{\Phi}_i^{(z+1)}(x,y)=\left(1-\dfrac{1}{z}\right)\boldsymbol{\Phi}_i^{(z)}(x,y)$；当用户设备不在基站服务范围内时，用户不会卸载到该基站的意愿为 $\boldsymbol{\Phi}_i^{(z+1)}(x,y)=1-\left(1-\dfrac{1}{z}\right)\left(1-\boldsymbol{\Phi}_i^{(z)}(x,y)\right)$。根据谨慎参数，接着，对区域内的基站逐个求解整个区域内愿

意卸载到它的任务量，并利用式(26)更新每个基站的服务范围。利用式(25)计算在本次迭代中的系统平均时延。然后，进行收敛判断，当前后两次迭代所得的系统平均时延差小于误差时，即达到最优基站服务区域划分。最后，将最优基站服务区域划分和系统平均时延最小值输出。

3　仿真分析

本节针对基于时延优化的卸载机制进行仿真分析，并将本文提出的用户设备卸载机制与基于最近距离的卸载机制以及基于随机接入的卸载机制进行对比。

本节仿真分析选取一个 $L_x \times L_y$ 的矩形区域，其中 L_x、L_y 分别为矩形区域的边长，5 个基站被部署在区域内。由于区域内的用户设备可以服从任意的二维连续分布，所以为了模拟用户设备分布不均匀情况，本文选择带有热点特征的二维截断式高斯分布，其表达式为

$$f(x,y)=\frac{1}{\varpi}\exp\left[-\left(\frac{x-\mu_x}{\sqrt{2}\sigma_x}\right)^2\right]\exp\left[-\left(\frac{y-\mu_y}{\sqrt{2}\sigma_y}\right)^2\right] \tag{36}$$

其中，$\varpi=2\pi\sigma_x\sigma_y\mathrm{erf}\left(\frac{L_x-\mu_x}{\sqrt{2}\sigma_x}\right)\mathrm{erf}\left(\frac{L_y-\mu_y}{\sqrt{2}\sigma_y}\right)$，$\mu_x$、$\mu_y$ 分别表示 x、y 方向上的均值，σ_x、σ_y 分别表示 x、y 方向上的标准差，$\mathrm{erf}(s)=\frac{2}{\sqrt{\pi}}\int_0^s \mathrm{e}^{-t^2}\mathrm{d}t$。根据二维截断高斯分布的性质，热点坐标为 (μ_x,μ_y)，σ_x、σ_y 分别代表热点周围在 x、y 方向上用户设备密度的倒数，因此定义热点周围 x、y 方向上的用户设备密度为 $\rho_x=\frac{1}{\sigma_x}$、$\rho_y=\frac{1}{\sigma_y}$，并且在后续的仿真中设定 $\rho_x=\rho_y$。本节设定上下行传输信道带宽 B_{up}、B_{do} 相同，均为 B。此外，在基站侧 MEC 服务器任务处理能力方面，本节仿真采用两种场景，其中场景一为区域内基站侧 MEC 服务器任务处理能力相同，即 $\{f_i\}_{i\in\mathcal{M}}=f$；场景二为 BS_3 侧 MEC 服务器处理能力是区域其他基站的 2 倍，即 $\{f_i\}_{i\in\mathcal{M}/\{3\}}=f$，$f_3=2f$。如果没有特殊说明，本节基于场景一进行仿真，仿真参数见表 1。

区域内用户设备分布情况和不同卸载机制下区域划分与基站负载情况如图 2 所示。区域内用户

表 1　　**仿真参数**

参数	说明	数值
L_x、L_y	矩形区域边长	1 000 m
μ_x、μ_y	二维截断高斯分布的均值	330 m、350 m
σ_x、σ_y	二维截断高斯分布的标准差	300、300
γ	数据包中数据量的均值	10 kbit
$\{(X_i,Y_i)\}_{i\in\mathcal{M}}$	基站位置坐标	(200,200)(200,800) (400,400)(800,200) (800,800)
f	基站侧 MEC 服务器的频率	30 GHz
ω	处理单比特数据所需 CPU 周期数	400 cycle/bit
U	用户设备数量	8 000
N_0	噪声功率密度	10^{-11} W
ϑ	瑞利分布标准差	0.5
N	天线根数	4
δ	回传结果数据量与上传任务数据量之间的比例	0.2
$\lambda_{\max}$	用户设备单位时间内产生数据包个数的最大值	4
α	路径损耗指数	4
$\{P_{\mathrm{d}}^i\}_{i\in\mathcal{M}}$	基站发射功率	1 W
P_{u}	用户设备发射功率	0.3 W
B	上、下行信道传输带宽	10 MHz

设备分布情况如图 2(a)所示，可以看到在(330,350)周围，用户设备分布比例最高，沿径向向外用户设备分布比例逐步减少。区域内基站侧 MEC 服务器任务处理能力相同，且卸载机制与计算任务生成速率不同的情况下，基站处理数据包量与区域内数据包总生成量的关系如图 2(b)所示。根据图线可知，在区域内用户设备的数据包生成速率一定的情况下，基于最近距离的卸载机制会造成基站之间被卸载数据包量出现显著差别，其中 BS_1、BS_3 被卸载任务量过大，分别是 BS_5 处理数据包量的 2.4 倍和 2.7 倍，这导致 BS_1、BS_3 对应的服务区域内用户设备计算任务卸载平均时延增加。而采用基于时延优化的卸载机制则会使每个基站处理数据包量保持相对均衡，使得整个区域内的用户设备被公平地服务。当卸载机制一定时，对应不同的用户设备数据包生成速率，基于时延优化的卸载机制能够在传输时延和计算时延之间进行权衡，动态调整每个基站处理的数据包量，使区域内用户设备能够在短时间内被公平地服务。图 2(c)、图 2(d)分别展示了在 λ_{max}=4 时基于最近距离的卸载机制的区域划分情况和基于时延优化的卸载机制的区域划分情况。根据图 2(b)可得，基于最近距离的卸载机制中分配给 BS_1、BS_3 的部分用户设备在基于时延优化的卸载机制分配给远离热点的 BS_2、BS_4、BS_5，在图 2(c)、图 2(d)中的体现是基站服务范围的变化。

采用不同卸载机制且用户设备发射功率不同时，系统平均时延与传输信道带宽的关系如图 3 所示。在同一区域中，在用户设备发射功率相同的情况下，采用基于最近距离的卸载机制与基于时延优化的卸载机制得到的系统平均时延均会随传输信道带宽的增大而降低。在用户设备发射功率和传输信道带宽一定时，采用基于时延优化的卸载机制得到的系统平均时延比采用基于最近距离的卸载机制得到的系统平均时延更低。当传输信道带宽为 14 MHz、用户设备发射功率为 0.1 W 时，系统平均时延降低了 32.49%；当传输信道带宽为 14 MHz、用户设备发射功率为 0.3 W 时，系统平均时延降低了 42.17%。对比用户设备不同的发射功率可得，当传输信道带宽一定时，用户设备发射功率越高，系统平均时延降低越大。

不同卸载机制下，系统平均时延和区域内用

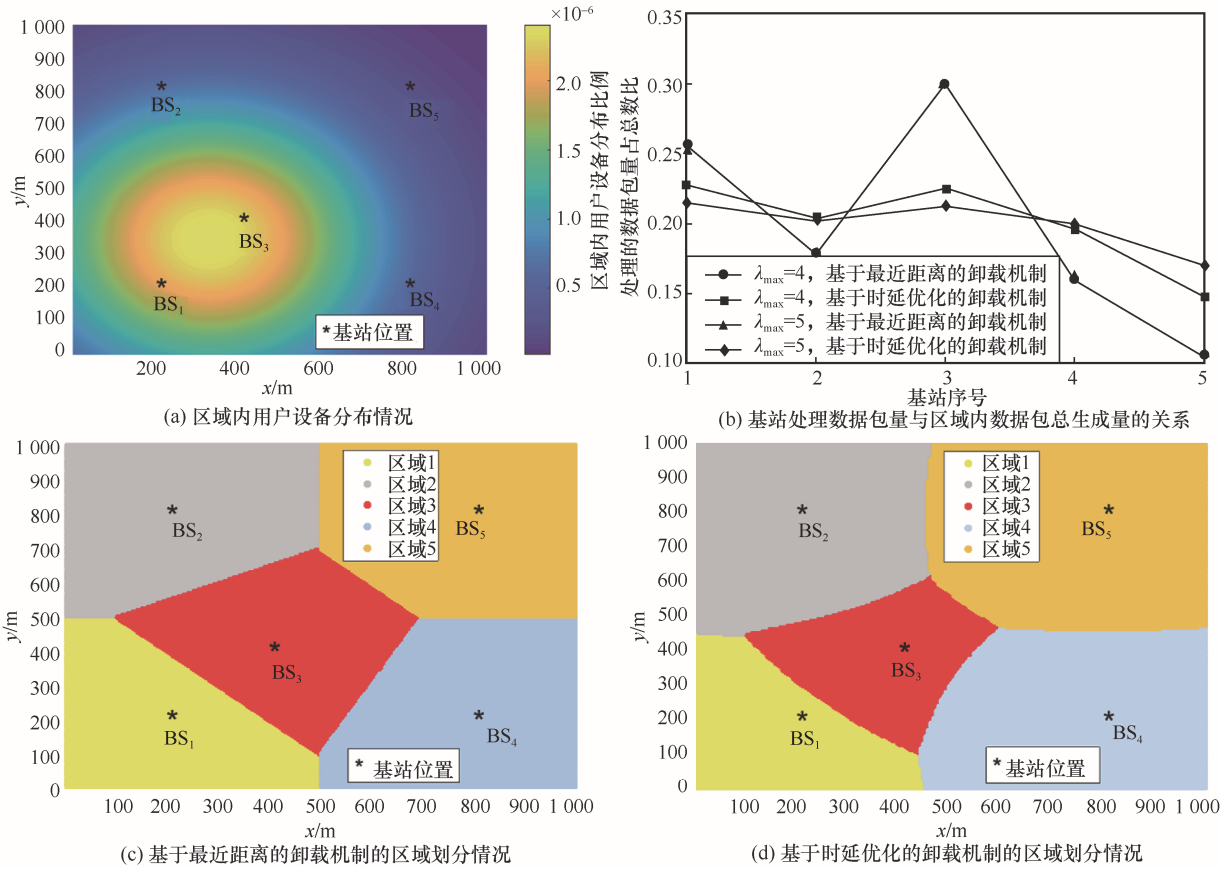


图 2　区域内用户设备分布情况和不同卸载机制下区域划分与基站负载情况

户设备数的关系如图 4 所示。根据图线可知，当用户设备卸载机制确定时，随着区域内用户设备数的提高，系统平均时延增加，而且基于最近距离的卸载机制得到的系统平均时延增长更快。当区域内用户设备数一定时，采用基于时延优化的卸载机制得到的系统平均时延更低。当区域内用户设备数为 8 800 时，采用基于时延优化的卸载机制得到的系统平均时延比采用基于最近距离的卸载机制得到的系统平均时延降低了 81.06%，比采用基于随机接入的卸载机制得到的系统平均时延降低了 79.28%。

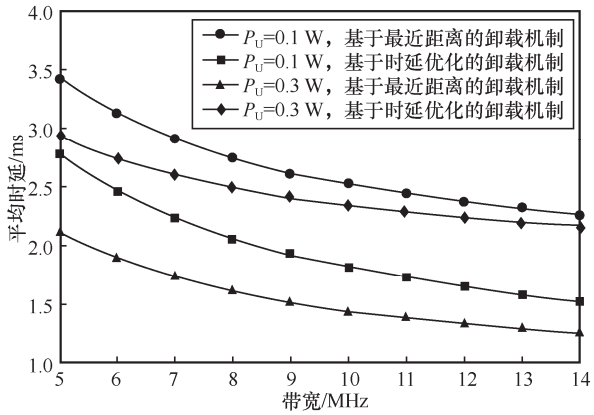


图 3　系统平均时延与传输信道带宽关系

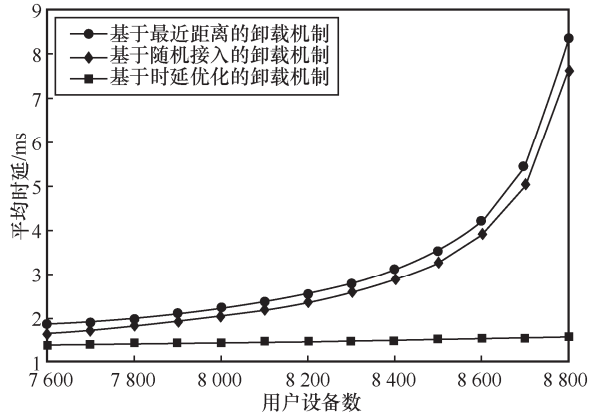


图 4　系统平均时延和区域内用户设备数的关系

不同场景下基站侧 MEC 服务器任务处理能力如图 5 所示，展示了采用不同卸载机制时，每个基站处理数据包量与区域内数据包总生成量之比之间的关系。由图线可知，在场景一中，基站侧 MEC 服务器任务处理能力相同，基于时延优化的卸载机制比基于最近距离的卸载机制具有更好的均衡性。而在场景二中，由于 BS_3 侧 MEC 服务器的任务处理能力较其他基站提高了一倍，所以采用基于时延优化的卸载机制得到的 BS_3 被卸载的数据包量大大超过其余基站。相较于场景二中采用基于时延优化的卸载机制能够对基站处理数据包量进行均衡，使用户设备得到公平的服务，采用基于最近距离的卸载机制对基站处理数据包量调整幅度较低，此时整个区域内用户设备不能被公平服务。

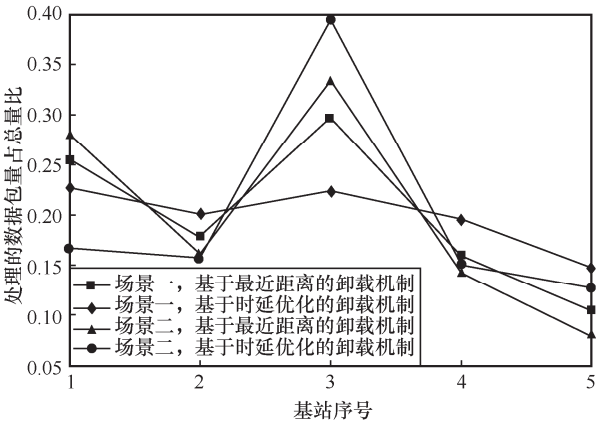


图 5　不同场景下基站侧 MEC 服务器任务处理能力

场景二下系统平均时延与传输信道带宽关系如图 6 所示。基于图 6 的结果，当卸载机制确定时，随着传输信道带宽的增加，系统平均时延降低。当传输信道带宽一定时，采用基于时延优化的卸载机制得到的系统平均时延更低。当传输信道带宽为 14 MHz 时，采用基于时延优化的卸载机制得到的系统平均时延比采用基于最近距离的卸载机制得到的系统平均时延降低了 33.02%，比采用基于随机接入的卸载机制得到的系统平均时延降低了 19.96%。

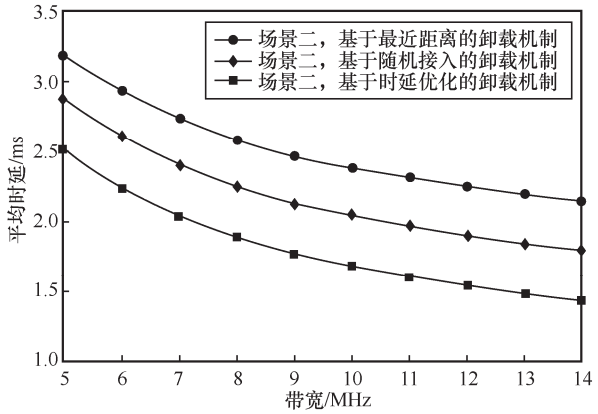


图 6　场景二下系统平均时延与传输信道带宽关系

不同卸载机制下，系统平均时延和用户设备热点处密度关系如图 7 所示。由图线可知，当卸载机

制一定时，随着用户设备热点处密度的增加，系统平均时延增加，而且基于最近距离的卸载机制得到的系统平均时延增长更快。当用户设备热点处密度一定时，采用基于时延优化的卸载机制得到的系统平均时延比采用基于最近距离的卸载机制得到的系统平均时延更低，且当热点处密度为 0.003 3 时，系统平均时延降低 55.24%。

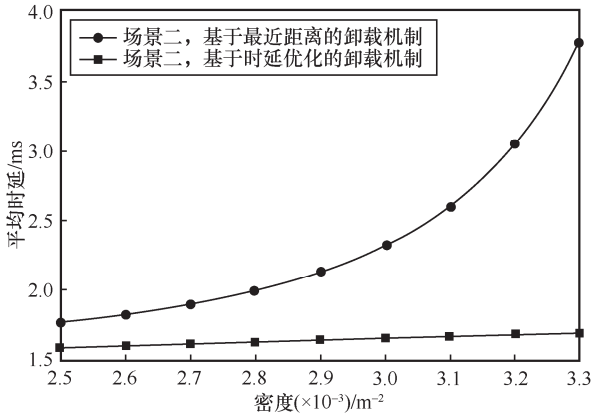


图 7　系统平均时延和用户设备热点处密度关系

传输信道带宽不同的情况下，采用基于时延优化的卸载机制时，系统平均时延与迭代次数的关系如图 8 所示。可以看到，在迭代次数相同的情况下，系统平均时延会随着传输信道带宽的增加而减小。在传输信道带宽一定时，系统平均时延会随着迭代次数的增加而趋于减小，并当到达一定迭代次数后，保持平稳。在迭代次数小于 5 时，随迭代次数增加，系统平均时延变化较为明显，此时系统内进行计算任务传输时延和计算时延的博弈。当迭代次数大于 5 时，随迭代次数增加，系统平均时延趋于平稳，达到最优的基站服务区域。此外，仿真结果表明，该算法在经过合理次数的迭代后能达到收敛。

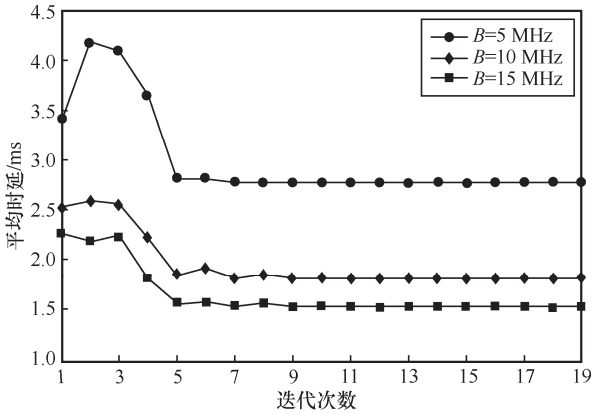


图 8　系统平均时延与迭代次数的关系

结合式(9)，定义区域内用户设备上传阶段的平均能耗为

$$E=\sum_{i=1}^{M}\iint_{\mathcal{D}_i} t_{\mathrm{up},i}^{j}(x,y)f(x,y)P_{\mathrm{u}}\mathrm{d}x\mathrm{d}y \tag{37}$$

根据式(37)，定义用户设备的贡献率 η，用来表示用户设备为了降低系统平均时延做出的能耗贡献，用数学表达式为

$$\eta=\frac{E_{\mathrm{OT}}-E_{\mathrm{nearest}}}{E_{\mathrm{OT}}}\times 100\% \tag{38}$$

其中，E_{OT} 表示基于时延优化的卸载机制中，用户设备计算任务上传阶段的平均能耗；E_{nearest} 表示基于最近距离的卸载机制中，用户设备计算任务上传阶段的平均能耗。

不同卸载机制下，用户设备上传计算任务时的平均能耗和贡献率与传输信道带宽的关系如图 9 所示。根据左边的 y 轴可知，随着传输信道带宽的提高，在不同卸载机制下用户设备上传计算任务时产生的平均能耗不断降低，且基于最近距离的卸载机制产生的上传能耗低于基于时延优化的卸载机制产生的上传能耗。结合图 2(c)、图 2(d)可知，这是因为部分用户设备将计算任务卸载到距离更远的基站。根据右边的 y 轴可知，用户设备的贡献率 η 随着传输信道带宽的提高而增加，当传输信道带宽为 14 MHz 时，用户设备的贡献率为 12%。结合图 3 可知，用户设备通过多消耗 12%的能量，换取系统平均时延降低 42.17%。

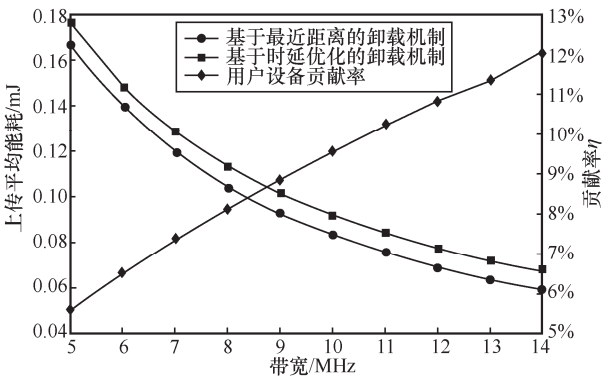


图 9　用户设备上传计算任务时的平均能耗和贡献率与传输信道带宽的关系

4　结束语

本文研究了多基站协同为区域内用户设备提供计算任务卸载服务的问题，具体来说，当区域内存在任意分布的用户设备且应用需求多样化时，通过合理划分基站服务区域，使系统平均时延最小。考虑优化区域是相互依赖的连续变量，本文引入最优

运输理论进行分析求解，提出一种基于最优运输理论的时延优化算法。该算法能够根据用户设备的空间分布、应用需求、MEC 服务器的任务处理能力等系统参数动态调整用户设备卸载的基站，也即调整基站服务区域，使系统平均时延最小。仿真结果表明，和基于最近距离的卸载机制相比，本文提出的基于时延优化的卸载机制使系统平均时延降低了81.06%，并能使各基站处理的业务量更加均衡。

参考文献：

[1] MAO Y Y, ZHANG J, LETAIEF K B. Dynamic computation offloading for mobile-edge computing with energy harvesting devices[J]. IEEE Journal on Selected Areas in Communications, 2016, 34(12): 3590-3605.

[2] ARCHANA R, VAISHNAVI C, PRIYANKA D S, et al. Remote health monitoring using IoT and edge computing[C]//Proceedings of 2022 International Conference on IoT and Blockchain Technology (ICIBT). Piscataway: IEEE Press, 2022: 1-6.

[3] ZHENG J B, YANG T Y, LIU H W, et al. Accurate detection and localization of unmanned aerial vehicle swarms-enabled mobile edge computing system[J]. IEEE Transactions on Industrial Informatics, 2021, 17(7): 5059-5067.

[4] WANG K. Migration strategy of cloud collaborative computing for delay-sensitive industrial IoT applications in the context of intelligent manufacturing[J]. Computer Communications, 2020(150): 413-420.

[5] LIN B, ZHU F N, ZHANG J S, et al. A time-driven data placement strategy for a scientific workflow combining edge computing and cloud computing[J]. IEEE Transactions on Industrial Informatics, 2019, 15(7): 4254-4265.

[6] REN J K, YU G D, CAI Y L, et al. Latency optimization for resource allocation in mobile-edge computation offloading[J]. IEEE Transactions on Wireless Communications, 2018, 17(8): 5506-5519.

[7] TAO O Y, ZHI Z, XU C. Follow me at the edge: mobility-aware dynamic service placement for mobile edge computing[J]. IEEE Journal on Selected Areas in Communications, 2018, 36(10): 2333-2345.

[8] ZHOU Y, YEOH P L, PAN C H, et al. Offloading optimization for low-latency secure mobile edge computing systems[J]. IEEE Wireless Communications Letters, 2020, 9(4): 480-484.

[9] MAO Y Y, ZHANG J, SONG S H, et al. Power-delay tradeoff in multi-user mobile-edge computing systems[C]//Proceedings of 2016 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2017: 1-6.

[10] LIU J, MAO Y Y, ZHANG J, et al. Delay-optimal computation task scheduling for mobile-edge computing systems[C]//Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT). Piscataway: IEEE Press, 2016: 1451-1455.

[11] RIMAL B P, VAN D P, MAIER M. Cloudlet enhanced fiber-wireless access networks for mobile-edge computing[J]. IEEE Transactions on Wireless Communications, 2017, 16(6): 3601-3618.

[12] LIU M T, YU F R, TENG Y L, et al. Distributed resource allocation in blockchain-based video streaming systems with mobile edge computing[J]. IEEE Transactions on Wireless Communications, 2019, 18(1): 695-708.

[13] CHEN L X, ZHOU S, XU J. Computation peer offloading for energy-constrained mobile edge computing in small-cell networks[J]. IEEE/ACM Transactions on Networking, 2018, 26(4): 1619-1632.

[14] 沈银芳. 多元 Monge-Kantorovich 运输问题研究[D]. 上海: 华东师范大学, 2009.
SHEN Y F. Study on multi-monge-kantorovich transportation problem[D]. Shanghai: East China Normal University, 2009.

[15] MASHKIN A L, TELUSHKINA E K, ULITSKAYA N M, et al. Digital technologies of public administration in transport[C]//Proceedings of 2021 Intelligent Technologies and Electronic Devices in Vehicle and Road Transport Complex (TIRVED). Piscataway: IEEE Press, 2021: 1-6.

[16] LI S Q, LANG M X, LI S Y, et al. Optimization of high-speed railway line planning with passenger and freight transport coordination[J]. IEEE Access, 2022(10): 110217-110247.

[17] OH G, SIM B, CHUNG H, et al. Unpaired deep learning for accelerated MRI using optimal transport driven CycleGAN[J]. IEEE Transactions on Computational Imaging, 2020(6): 1285-1296

[18] AKBARI A, AWAIS M, FATEMIFAR S, et al. Deep order-preserving learning with adaptive optimal transport distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 313-328.

[19] WANG D, TIAN J, ZHANG H X, et al. Task offloading and trajectory scheduling for UAV-enabled MEC networks: an optimal transport theory perspective[J]. IEEE Wireless Communications Letters, 2022, 11(1): 150-154.

[20] MOZAFFARI M, SAAD W, BENNIS M, et al. Wireless communication using unmanned aerial vehicles (UAVs): optimal transport theory for hover time optimization[J]. IEEE Transactions on Wireless Communications, 2017, 16(12): 8052-8066.

[21] SILVA A, TEMBINE H, ALTMAN E, et al. Optimum and equilibrium in assignment problems with congestion: mobile terminals association to base stations[J]. IEEE Transactions on Automatic Control, 2013, 58(8): 2018-2031.

[22] WANG Y, HU Z Q, WEN X M, et al. Three-dimensional aerial cell partitioning based on optimal transport theory[C]//2020 IEEE International Conference on Communications Workshops (ICC Workshops). Piscataway: IEEE Press, 2020: 1-6.

[23] WANG L Y, ZHANG H X, GUO S S, et al. Deployment and association of multiple UAVs in UAV-assisted cellular networks with the knowledge of statistical user position[J]. IEEE Transactions on Wireless Communications, 2022, 21(8): 6553-6567.

[24] AREDO S C, NEGASH Y, MARYE Y W, et al. Hardware efficient massive MIMO systems with optimal antenna selection[J]. Sensors, 2022, 22(5): 1743.

[25] SHAN X Y, ZHI H X, LI P, et al. A survey on computation offloading for mobile edge computing information[C]//Proceedings of 2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDa-

taSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS). Piscataway: IEEE Press, 2018: 248-251.

[26] 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.

[27] NGUYEN T T, LE L B, LE-TRUNG Q. Computation offloading in MIMO based mobile edge computing systems under perfect and imperfect CSI estimation[J]. IEEE Transactions on Services Computing, 2021, 14(6): 2011-2025.

[28] GÓMEZ-DÉNIZ E, GÓMEZ-DÉNIZ L. A generalisation of the Rayleigh distribution with applications in wireless fading channels[J]. Wireless Communications and Mobile Computing, 2013, 13(1): 85-94.

[29] 崔高峰, 徐媛媛, 张尚宏, 等. 基于最小能耗的多无人机无线网络安全数据卸载策略[J]. 通信学报, 2021, 42(5): 51-62.
CUI G F, XU Y Y, ZHANG S H, et al. Secure data offloading strategy for multi-UAV wireless networks based on minimum ene-rgy consumption[J]. Journal on Communications, 2021, 42(5): 51-62.

[30] COOPER R B. Introduction to queueing theory[M]. London: Edward Arnold, 1981.

[31] MONGE G. Mémoire sur la théorie des déblais et des remblais[J]. Mem. Math. Phys. Acad. Royale Sci., 1781: 666-704.

[32] 沈雪姣. 建立在偏微分方程/概率理论基础上 Monge-Kantorovich 问题的快速算法[D]. 上海: 华东师范大学, 2012.
SHEN X J. A fast algorithm for Monge-Kantorovich problem based on partial differential equation/probability theory[D]. Shanghai: East China Normal University, 2012.

[33] VILLANI C. Topics in optimal transportation[M]. Providence: American Mathematical Society, 2003

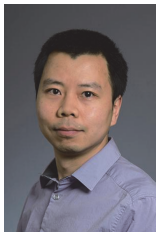
[34] AMBROSIO L, GIGLI N. A user's guide to optimal transport[M]// Heidelberg: Springer, 2013: 1-155.

[35] CRIPPA G, JIMENEZ C, PRATELLI A. Optimum and equilibrium in a transport problem with queue penalization effect[J]. Advances in Calculus of Variations, 2009, 2(3): 207-246.

[36] 殷凤, 王鹏飞. 二重积分中值定理的推广[J]. 忻州师范学院学报, 2011, 27(2): 15-16, 30.
YIN F, WANG P F. The extension of double integral mean value theorem[J]. Journal of Xinzhou Teachers University, 2011, 27(2): 15-16, 30.

[作者简介]

吕翔宇（1996– ），男，华中科技大学电子信息与通信学院硕士生，主要研究方向为无线通信、边缘卸载、最优运输理论。



肖泳（1980– ），男，博士，华中科技大学教授，主要研究方向为网络人工智能、边缘计算、通信网络博弈理论等。

钟祎（1989– ），男，博士，华中科技大学副教授，主要研究方向为无线干扰管理、资源分配等。

李强（1984– ），男，博士，华中科技大学教授，主要研究方向为无线协作通信、认知无线电/协作频谱共享、无线信息能量同传、物联网、边缘计算、边缘缓存等。

葛晓虎（1972– ），男，博士，华中科技大学教授，主要研究方向为移动通信、无线网络中的流量建模、绿色通信等。