移动性感知的边缘服务迁移策略

吴大鹏[1,2,3]，吕吉[1,2,3]，李职杜[1,2,3]，王汝言[1,2,3]

（1. 重庆邮电大学通信与信息工程学院，重庆 400065；2. 重庆高校市级光通信与网络重点实验室，重庆 400065；
3. 泛在感知与互联重庆市重点实验室，重庆 400065）

摘 要：针对移动边缘计算网络中由于用户位置动态变化而导致边缘服务器间负载不均衡、用户服务质量降低的问题，提出了一种移动性感知的边缘服务迁移算法。首先，以最小化用户服务请求感知时延为目标，将优化问题建模为混合整数非线性规划问题。其次，基于 Lyapunov 优化方法将时延优化问题解耦为边缘服务迁移子问题与无线接入子问题。再次，提出快速边缘决策算法求解出给定无线接入策略情况下最优的资源分配与边缘服务迁移方案。最后，提出异步最佳响应算法迭代出最优无线接入策略。仿真结果表明，与现有服务迁移策略相比较，所提算法能够在保证服务迁移成本稳定的情况下降低用户服务请求的感知时延。

关键词：移动边缘计算网络；边缘服务迁移；迁移成本；感知时延

中图分类号：TP393

文献标识码：A

doi: 10.11959/j.issn.1000−436x.2020085

Mobility aware edge service migration strategy

WU Dapeng[1,2,3], LYU Ji[1,2,3], LI Zhidu[1,2,3], WANG Ruyan[1,2,3]

1. School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2. Key Laboratory of Optical Communication and Networks in Chongqing, Chongqing 400065, China
3. Key Laboratory of Ubiquitous Sensing and Networking in Chongqing, Chongqing 400065, China

Abstract: To address the problem of load imbalance among edge servers and quality of service degradation caused by dynamic changes of user locations in mobile edge computing networks, a mobility aware edge service migration algorithm was proposed. Firstly, the optimization problem was formulated as a mix integer nonlinear programming problem, with the goal of minimizing the perceived delay of user service request. Then, the delay optimization problem was decoupled into the edge service migration and edge node selection sub-problems based on the Lyapunov optimization approach. Thereafter, the fast edge decision algorithm was proposed to optimize the resource allocation and edge service migration under a given radio access strategy. Finally, the asynchronous optimal response algorithm was proposed to iterate out the optimal radio access strategy. Simulation results validate the proposed algorithm can reduce the perceived delay under the service migration cost constraint while comparing with other existing algorithms.

Key words: mobile edge computing network, edge service migration, migration cost, perceived delay

1 引言

随着移动互联网的不断发展，虚拟现实、智能家居等资源密集型应用在人们的生产和生活中发挥着越来越重要的作用[1-2]。这些应用通常具有超低时延和海量数据处理等需求，给储能、计算及缓存能力有限的移动设备带来了极大的挑战。移动边缘计算网络把移动边缘计算（MEC, mobile edge

收稿日期：2020–01–06；**修回日期**：2020–03–26

通信作者：李职杜，lizd@cqupt.edu.cn

基金项目：国家自然科学基金资助项目（No.61771082, No.61871062, No.61901078）；重庆市教委科学技术研究基金资助项目（No.KJQN201900609）；重庆市高校创新团队建设计划基金资助项目（No.CXTDX201601020）

Foundation Items: The National Natural Science Foundation of China (No.61771082, No.61871062, No.61901078), The Science and Technology Research Program of Chongqing Municipal Education Commission (No.KJQN201900609), Chongqing Funded Project of Chongqing University Innovation Team Construction (No.CXTDX201601020)

computing）与移动云计算（MCC, mobile cloud computing）的各自优势相结合，为这些应用需求提供了新的解决思路[3-4]。然而，如何为移动中的设备（如车辆、轨道交通等）提供连续不间断的服务是制约移动边缘计算网络进一步提高用户服务质量的瓶颈之一。

在移动边缘计算网络中，由于用户的移动性、服务请求的多样性及区域请求量的差异性，容易造成边缘服务器负载不平衡、热点地区网络拥塞等问题，从而严重降低用户的服务质量。虽然边缘服务迁移技术能够保证为用户提供连续化服务，而且还能平衡边缘服务器之间的工作负担，但受限于移动边缘计算网络的通信能力、计算资源及存储容量，如何根据用户的实时位置信息快速、高效地进行边缘服务迁移，成为移动边缘计算网络动态服务迁移中待解决的关键问题之一。

本文从实际情况出发，提出了一种移动性感知的边缘服务迁移策略，考虑到用户的运动轨迹具有部分可预测性及服务请求类别的多样性，对用户的服务迁移决策与移动边缘计算网络的通信–计算–存储资源进行了联合优化。本文主要贡献如下。

1) 通过 Lyapunov 优化理论有效地将长期迁移成本预算转换为实时优化，在不需要任何用户位置先验信息的条件下，根据用户的实时位置信息进行快速、高效的在线服务迁移，克服了由于用户运动轨迹的随机性而导致的无法建立马尔可夫状态转移模型的问题。

2) 从服务提供商的角度出发，考虑到用户服务的频繁迁移会造成系统长期迁移成本超出迁移阈值，运用 Lyapunov 优化方法在保证系统长期迁移成本稳定的基础上，最小化用户服务请求的平均感知时延。

3) 为降低混合整数非线性规划问题的求解复杂度，设计了一种以用户为中心的异步最佳响应方案来寻找近似最优的边缘服务迁移策略，在显著降低算法运行时间的同时，进一步降低了用户服务请求的平均感知时延。

2 相关工作

针对移动边缘计算网络中动态服务迁移问题，国内外学者均开展了相关的研究。文献[5]根据用户的地理位置及其服务请求类型，考虑服务运行过程中表现出非对称带宽的需求，提出了一种满足多维（存储、计算与通信）约束的联合优化算法，有效解决了多维约束条件下 MEC 增强的多蜂窝网络中的服务放置和请求路由问题。文献[6]考虑到边缘服务器的异构性问题，以最大化服务放置奖励为优化目标，提出了一种灵活的服务放置算法。该算法将每个边缘节点看作某一类应用服务器，在每个时隙进行服务放置决策。上述 2 种算法无法为正在移动的用户提供无间断服务，严重降低了用户服务质量。文献[7]通过构造多参数马尔可夫决策过程（MDP, Markov decision process）收益函数解决了由于边缘服务器服务范围有限而造成对移动车辆服务中断的问题，改进了单纯基于距离进行服务迁移方案的不足，但在迁移决策中没有对服务提供方的成本开销进行合理地建模与评估。在用户接入节点固定的非重叠覆盖场景下，Ouyang 等[8]考虑到用户的移动性与服务请求到达的不确定性问题，提出了具有概率分布模型的马尔可夫近似算法，通过构建不可约马尔可夫链得到了不同时间段的最优服务迁移策略。但对于多节点覆盖场景，该算法无法做出用户服务请求的节点选择决策。文献[9]提出了云–边–端三层架构模型，在该模型中，每个基站覆盖下的用户首先向本地 MEC 服务器请求服务，当本地 MEC 服务器没有响应用户的服务需求时，该请求可通过基站路由到远端云服务器。然而，该方法将导致边缘服务器负载不均衡，部分用户服务质量严重下降等问题，此外，该方案忽略了用户的移动性。文献[10]提出了一种基于边缘认知计算（ECC, edge cognitive computing）的动态服务迁移机制，该机制可以根据用户的行为认知快速地进行服务迁移，解决了用户在不同行为下服务质量大幅度波动的问题。但该机制以满足用户个性化服务为目标，忽视了多用户场景中边缘服务器资源有限的情况。文献[11]结合远程加载与重定向技术提出了基于虚拟机的快速服务迁移方法（FSMM, fast service migration method），该方法基于最短距离为用户实现用户服务迁移，具体地，在边缘服务器覆盖范围内的用户均就近接受服务，不在任何边缘服务器覆盖范围内的用户均由远端云服务器提供服务。文献[12]从理论的角度来量化间歇性连接对移动边缘计算的影响，所提间歇性服务迁移方法（ISMM, intermittent service migration method）根据当前网络环境自适应调整用户服务迁移响应间隔，有效降低了用户的平均感知时延。文献[13]将服务迁移问题描述为马尔

可夫决策过程，通过用户与服务位置之间的相对距离来获取近似的底层状态空间，然而，该决策模型仅适用于一维服务迁移场景。

综上所述，已有的相关研究主要从 3 个方面实现用户的服务迁移。1) 假设用户移动路径已知的情况下，对系统资源与服务迁移策略进行联合优化，使系统某一性能（如用户感知时延、系统吞吐量等）达到最优。2) 在服务迁移决策过程中，忽略服务迁移所产生的成本开销与用户被多个微基站重叠覆盖的情况，以提高用户服务质量为优化目标。3) 假设边缘服务器拥有充足的计算与存储资源，把用户的移动性建模为一维或二维马尔可夫状态转移模型，根据用户在每个时刻所处的环境状态进行边缘服务迁移决策。然而，在实际应用场景中，用户的地理位置具有部分可预测性，对于运动轨迹无法预测的用户，其移动性无法通过马尔可夫状态转移模型进行刻画。其次，已有研究没有从服务提供商的角度考虑如何在保证用户服务质量的前提下，使边缘服务迁移成本开销保持长期稳定。最后，由于移动边缘计算网络的通信资源、计算能力以及存储容量是有限的，忽略对这类资源的合理调度将导致网络无法应对数量呈爆发式增长且服务质量需求日益增高的各类应用请求。

3　系统模型

本文的系统模型如图 1 所示，在宏基站（MBS, macro base station）覆盖范围内存在多个小基站（SBS, small base station），其中 MEC 增强的 SBS 又称为边缘节点（EP, edge point）。考虑到每个小基站覆盖范围内服务请求量的差异性，为节约部署成本，服务请求量相对较小的地区不需要部署边缘服务器，但对于热点区域情况则相反。实际上，非热点区域的服务请求可以通过小基站转发至宏基站，并由宏基站转发到远端云服务器或者周边的边缘服务器。例如：小基站 SBS_1 没有部署边缘服务器，所以用户 user_1 只能通过宏基站向云服务器或者邻近的边缘服务器请求服务。对于没有被小基站覆盖的用户，其服务请求只能通过宏基站转发到云服务器或者边缘服务器。由于用户的移动性或新用户的加入使网络的拓扑结构发生变化，导致用户的服务进程从原来的服务器迁移到另一个服务器。开始时刻，小基站 SBS_2 为用户 user_2 提供服务。一段时间后，用户 user_2 移动到小基站 SBS_3 所覆盖的区域。由于边缘服务器存储资源、计算资源以及通信资源有限，在每个边缘服务器上仅能运行有限的应用服务，小基站 SBS_3 没有可供用户 user_2 服务需求使用的资源，因此用户 user_2 在小基站 SBS_2 的服务配置文件只能通过宏基站迁移到邻近的边缘服务器 SBS_4，从而实现边缘服务迁移。

在宏基站覆盖范围内部署 M 个 SBS 与 I 个边缘服务器，用户数量为 $N+1$。令 $\mathcal{T}=\{0,1,2,\cdots,T\}$ 表示时间离散化序列集合，$y_k^j(t)$ 表示用户 k 在时隙 t 内关于 BS_j（j=0 时为 MBS，$j\neq 0$ 时为 SBS）的接入参数，用户 k 接入 BS_j，则 $y_k^j(t)=1$，否则 $y_k^j(t)=0$。通常情况下，用户 k 只能接入到一个 BS。因此，对任意的用户 k 来说，其接入约束条件如式(1)所示。

$$\sum_{j=0}^{M} y_k^j(t)=1 \tag{1}$$

由于边缘服务器处理任务的输出结果远小于任务大小[14-17]，本文只考虑上行信道资源分配。假设系统的通信资源块（RB, resource block）总数为 F，每个通信资源块的带宽为 W MHz，宏基站的资源块总数为 C_0，剩余的 $F-C_0$ 通信资源块由 SBS

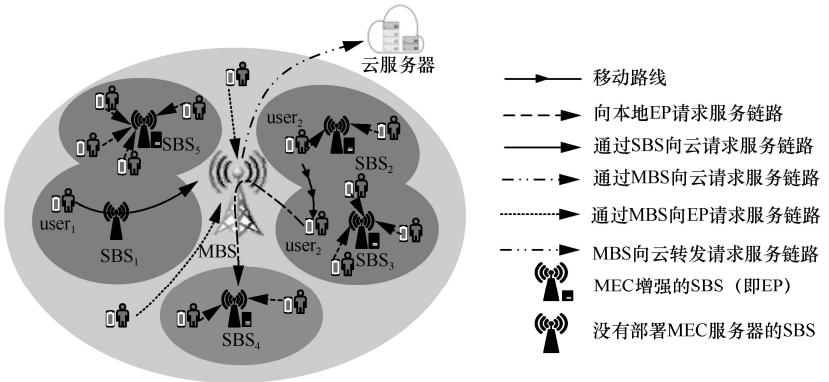

图 1　系统模型

复用。由于不同 SBS 所管理的无线资源存在差异性，所以对于任意的 SBS_j，其占用的无线资源块数应满足 $C_j \leqslant F - C_0$，$1 \leqslant j \leqslant M$。因此，对任意的 BS_j，为保证用户所占用的通信资源在其承受范围内，其通信资源约束条件需满足

$$\sum_{k=0}^{N} \lambda_k^s(t) y_k^j(t) \leqslant C_j \tag{2}$$

其中，$\lambda_k^s(t) = \left\lceil \frac{R_k^s(t)}{r_k^j(t)} \right\rceil$ 表示时隙 t 内用户 k 请求服务 s 的资源块数，$R_k^s(t)$ 表示用户 k 请求服务 s 的上行数据传输速率，$r_k^j(t)$ 表示在用户 k 使用单个资源块上传数据到 BS_j 所获得的传输速率，$\lceil \cdot \rceil$ 表示向上取整。

由于服务器计算与存储能力有限，对于任意的时隙 t，服务器 i 服务用户所需的计算与存储不能超过其总的计算与存储资源，因此服务器 i 的计算资源约束条件如式(3)所示，存储资源约束条件如式(4)所示。

$$\sum_{k=0}^{N} \sum_{s=0}^{S} \phi_k^s(t) x_k^i(t) \leqslant \phi_i \tag{3}$$

$$\sum_{k=0}^{N} \sum_{s=0}^{S} B_k^s(t) x_k^i(t) \leqslant B_i \tag{4}$$

其中，ϕ_i 和 B_i 分别表示服务器 i 的总计算资源与总存储资源；$\phi_k^s(t)$ 和 $B_k^s(t)$ 分别表示用户 k 请求服务 s 所需要的计算资源与存储资源；$S+1$ 表示所有用户服务请求服务类别的总和；$x_k^i(t)$ 表示用户 k 被服务器 i 所托管，$i=0$ 时表示云服务器，否则表示边缘服务器。

3.1 服务体验模型

由于用户发起的服务请求被边缘服务器或者云服务器响应，因此用户的数据传输存在一定的传输时延。在本文模型中，传输时延可分为以下三类：1)用户的服务请求被本地边缘服务器响应；2)用户的服务请求通过 MBS 转发到邻近的边缘服务器；3)没有边缘服务器响应用户的服务请求时，该请求只能被远端云服务器响应。因此，系统中所有用户发送服务请求到 BS 的传输时延 $D_c(t)$ 为

$$D_c(t) = \sum_{k=0}^{N} \sum_{i=0}^{M} \sum_{s=0}^{S} \frac{y_k^i(t) q_k^s(t)}{R_k^s(t)} \tag{5}$$

MBS 转发服务请求到 SBS 或者云服务器的传输时延 $D_f(t)$ 为

$$D_f(t) = \sum_{k=0}^{N} \sum_{i=0}^{M} \sum_{s=0}^{S} \frac{y_k^0(t) q_k^s(t) x_k^i(t)}{r_i(t)} \tag{6}$$

所有用户的服务请求被服务器响应的总时延 $D_{\text{total}}(t)$ 为

$$D_{\text{total}}(t) = D_c(t) + D_f(t) \tag{7}$$

其中，$q_k^s(t)$ 表示用户 k 请求服务 s 的数据大小，单位为 bit；$r_i(t)$ $(i \neq 0)$ 表示 MBS 转发服务请求到 SBS_i 的传输速率；$r_0(t)$ 表示 MBS 转发服务请求到云服务器的传输速率。根据文献[7]，令 $D_{ij}^s(t)$ 和 $E_{ij}^s(t)$ 分别表示服务 s 从 SBS_i 迁移到 SBS_j 的迁移时延与迁移成本，当 $i=j$ 时，$D_{ij}^s(t) = E_{ij}^s(t) = 0$。令 $D_{i0}^s(t)$ 和 $E_{i0}^s(t)$ 分别表示服务 s 从 SBS_i 迁移到云服务器的时延与成本，$D_{0i}^s(t)$ 和 $E_{0i}^s(t)$ 分别表示从云服务器迁移到 SBS_i 的时延与成本。系统中所有用户的迁移时延总和 $D_m(t)$ 及迁移成本总和 $E(t)$ 分别为

$$D_m(t) = \sum_{k=0}^{N} \sum_{s=0}^{S} \sum_{i=0}^{M} \sum_{j=0}^{M} x_k^i(t-1) x_k^j(t) D_{ij}^s(t) \tag{8}$$

$$E(t) = \sum_{k=0}^{N} \sum_{s=0}^{S} \sum_{i=0}^{M} \sum_{j=0}^{M} x_k^i(t-1) x_k^j(t) E_{ij}^s(t) \tag{9}$$

3.2 系统性能与成本

首先，为缓解数据上传到远端云服务器对核心网造成的流量拥塞问题，部署在边缘服务器上的服务应最大化满足用户的需求。其次，为降低网络拓扑结构变化所带来的服务迁移成本，系统中所有用户在时隙 t 内的服务迁移成本 $E(t)$ 应保持在预期的范围之内，令 E_{avg} 表示长期预算迁移成本，可得成本约束条件如式(10)所示。

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E(t) \leqslant E_{\text{avg}} \tag{10}$$

在保证系统长期平均迁移成本小于预算成本的约束条件下，根据用户服务请求制定最优的服务迁移策略，使所有用户的平均感知时延与服务迁移时延之和最小，可得如下优化问题。

$$\text{P1}: \min_{\boldsymbol{X}(t)\boldsymbol{Y}(t)\boldsymbol{\lambda}(t)} \frac{1}{T} \lim_{T \to \infty} \sum_{t=0}^{T} \left(D_{\text{total}}(t) + D_m(t) \right)$$

$$\text{s.t. 式(1)~式(4),式(7),式(8),式(10)}$$

其中，$\boldsymbol{Y}(t) = \{y_k = (\arg_i y_k^i(t) \neq 0)\}$、$\boldsymbol{X}(t) = \{x_k = (\arg_i x_k^i(t) \neq 0)\}$、$\boldsymbol{\lambda}(t) = \{\lambda_k^s(t)\}$（$k \in \{1, \cdots, N\}, i \in \{0, 1, \cdots, M\}, s \in \{0, 1, \cdots, S\}, t \in \mathcal{T}$）分别表示无线接入节点选择向量、边缘服务

节点选择向量及通信资源分配矩阵。由于优化问题P1不仅考虑了用户的移动性，还与用户边缘节点选择、服务迁移策略及通信资源的分配相关，因此该优化问题是高度耦合了时间与空间相关性的混合整数非线性规划问题。

4　基于 Lyapunov 优化的在线服务迁移

用户的移动容易造成用户服务在服务器之间频繁迁移，为服务提供商带来高昂的迁移费用。从服务提供商角度考虑，如何在保证服务迁移成本不超出其所能承受阈值的基础上，最大化降低用户所需服务的感知时延是动态服务迁移策略必须满足的关键目标之一。为实现该目标，本文运用 Lyapunov 优化方法保证系统的长期迁移成本小于迁移阈值的同时，满足用户对感知时延的要求。Lyapunov 优化的关键思想是在当前用户的感知时延和迁移成本之间取得相对的平衡，通过为长期成本预算引入虚拟队列来保持迁移成本的长期稳定。首先将虚拟队列定义为超出预算成本的历史度量，假设初始队列长度为 0 bit（即 $Q(0)=0$ bit），队列更新方式如式(11)所示。

$$Q(t+1)=\max\left[Q(t)+E(t)-E_{\text{avg}},0\right] \tag{11}$$

直观上，$Q(t)$ 可以作为当前服务迁移成本状况的评价标准。$Q(t)$ 越大，表示自实施在线服务迁移策略开始，迁移成本超过长期预算迁移成本 E_{avg} 越多。为了保证长期迁移成本低于 E_{avg}，$Q(t)$ 必须满足 $\lim\limits_{T\to\infty}\dfrac{\mathbb{E}\left[Q(t)\right]}{T}=0$，其中 $\mathbb{E}[\cdot]$ 表示数学期望。

对式(11)进行化简，得到 $Q(t+1)\geqslant Q(t)+E(t)-E_{\text{avg}}$，进一步求和并重新排列，得到式(12)。

$$\frac{Q(T)-Q(0)}{T}+E_{\text{avg}}\geqslant\frac{1}{T}\sum_{t=0}^{T-1}E(t) \tag{12}$$

由于 $Q(0)=0$，可进一步得到

$$\lim_{T\to\infty}\frac{\mathbb{E}\left[Q(T)\right]}{T}+E_{\text{avg}}\geqslant\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[E(t)\right] \tag{13}$$

式(13)表明了长期迁移成本不能超过预算成本。

为了保持虚拟队列稳定，首先分别定义二次 Lyapunov 函数与 Lyapunov 漂移函数[18-20]如式(14)与式(15)所示。

$$L(Q(t))=\frac{1}{2}Q(t)^2 \tag{14}$$

$$\Delta(Q(t))\triangleq\mathbb{E}\left[L(Q(t+1))-L(Q(t))\,|\,Q(t)\right] \tag{15}$$

其中，$\Delta(Q(t))$ 表示在二次 Lyapunov 函数中，相邻时隙间迁移成本的变化量，该变化量有助于调整服务迁移策略以适应系统的动态变化。

为了得到当前时隙下最优的服务迁移策略，可通过 Lyapunov 优化方法把原始问题分解为实时性优化问题。为保持系虚拟队列稳定，通过定义 Lyapunov 漂移加惩罚函数构建实时服务迁移问题，其中 Lyapunov 漂移加惩罚函数如式(16)所示。

$$\Delta(Q(t))+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right] \tag{16}$$

进一步可以得到

$$\begin{aligned}&\Delta(Q(t))+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]\leqslant\\&B+Q(t)\mathbb{E}\left[(E(t)-E_{\text{avg}})\,|\,Q(t)\right]+\\&V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]\end{aligned} \tag{17}$$

其证明如下。

$$\begin{aligned}&\Delta(Q(t))+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]=\\&\mathbb{E}\left[\frac{1}{2}\left(\max\left[Q(t)+E(t)-E_{\text{avg}},0\right]\right)^2-\frac{1}{2}(Q(t))^2\,|\,Q(t)\right]+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]\leqslant\\&\mathbb{E}\left[\frac{1}{2}\left(Q(t)+E(t)-E_{\text{avg}}\right)^2-\frac{1}{2}(Q(t))^2\,|\,Q(t)\right]+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]=\\&\mathbb{E}\left[\frac{1}{2}(E(t)-E_{\text{avg}})^2+Q(t)(E(t)-E_{\text{avg}})\,|\,Q(t)\right]+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]\leqslant\\&\mathbb{E}\left[\frac{1}{2}(E(t))^2+\frac{1}{2}(E_{\text{avg}})^2+Q(t)(E(t)-E_{\text{avg}})\,|\,Q(t)\right]+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]\leqslant\\&\frac{1}{2}(E_{\max})^2+\frac{1}{2}(E_{\text{avg}})^2+Q(t)\mathbb{E}\left[(E(t)-E_{\text{avg}})\,|\,Q(t)\right]+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]=\\&B+Q(t)\mathbb{E}\left[(E(t)-E_{\text{avg}})\,|\,Q(t)\right]+V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]\end{aligned}$$

其中，$B=\frac{1}{2}(E_{\text{avg}}^2+E_{\max}^2)$，$E_{\max}=\max_{t_0\in t}E(t_0)$。式(17)证明式(16)在每一个时隙都存在唯一的上界。通过化简式(17)右边部分可进一步得到式(18)。

$$V\mathbb{E}\left[D_{\text{total}}(t)+D_m(t)\,|\,Q(t)\right]+Q(t)\mathbb{E}\left[E(t)-E_{\text{avg}}\,|\,Q(t)\right]\leqslant\sum_{k=0}^{N}\sum_{j=0}^{M}\left(Vy_k^j(t)D_k^s(t)+x_k^j(t)\varsigma_k^j(t)\right)\tag{18}$$

其中，$\varsigma_k^j(t)=\sum_{s=0}^{S}\left(\sum_{i=0}^{M}x_k^i(t-1)(Q(t)E_{ij}^s(t)+D_{ij}^s(t))+\frac{Vy_k^0(t)q_k^s(t)}{r_j(t)}\right)$，$D_k^s(t)=\frac{\sum_{s=0}^{S}q_k^s(t)}{\lambda_k^s(t)r_k^j(t)}$。最小化式(18)右边部分得到最小化实时上确界，可表示为问题P2。

$$\text{P2}:\min_{\boldsymbol{Y}(t)\boldsymbol{X}(t)\boldsymbol{\lambda}(t)}\sum_{k=0}^{N}\sum_{j=0}^{M}\left(Vy_k^j(t)D_k^s(t)+x_k^j(t)\varsigma_k^j(t)\right)$$
$$\text{s.t. 式(1)}\sim\text{式(4)},\text{式(7)}\sim\text{式(9)},\text{式(11)}$$

运用 Lyapunov 优化将问题P1转化为用户接入节点选择与服务迁移决策的问题P2。对于问题P2的求解，可使用穷举法列举所有可能的服务迁移策略，然后基于拉格朗日乘子法或者内点法求出当前服务迁移策略下的最优通信资源分配$\boldsymbol{\lambda}(t)$。然而，穷举法算法复杂度为$O(FNM)$，复杂度过高，无法应用于大型移动边缘计算网络。观察问题P2可知，其优化目标的第一项与用户无线接入节点选择$\boldsymbol{Y}(t)$和通信资源分配$\boldsymbol{\lambda}(t)$相关，第二项表示MBS转发所有用户的服务请求到边缘服务器或云服务器所需的总时延。

5 自适应最佳响应策略

5.1 快速边缘决策算法

本文求解问题P2的总体思路如下。首先，假设在t时隙内，用户最优的无线接入节点选择参数为$\boldsymbol{Y}^*(t)$。其次，将问题P2分解为2个子问题：通信资源分配问题与边缘节点选择问题。分别求解在已知无线接入策略$\boldsymbol{Y}^*(t)$的情况下最优的$\boldsymbol{\lambda}^*(t)$和$\boldsymbol{X}^*(t)$。最后，将$\boldsymbol{\lambda}^*(t)$与$\boldsymbol{X}^*(t)$代入问题P2，求解出最优的$\boldsymbol{Y}^*(t)$。具体地，通信资源分配问题P2_1表述如下。

$$\text{P2\_1}:\min_{\boldsymbol{\lambda}(t)}\sum_{k=0}^{N}\sum_{j=0}^{M}Vy_k^{j^*}(t)D_k^s(t)$$
$$\text{s.t.式 (2)}$$

运用拉格朗日乘子法消去约束条件(2)，得到如下优化算子。

$$f(\boldsymbol{\lambda}(t),\boldsymbol{\beta}(t))=\sum_{k=0}^{N}\sum_{j=0}^{M}\frac{Vy_k^{j^*}(t)\sum_{s=0}^{S}q_k^s(t)}{\lambda_k^s(t)r_k^j(t)}+\sum_{j=0}^{M}\beta_j(t)\left(\sum_{k=0}^{N}\lambda_k^s(t)y_k^{j^*}(t)-C_j\right)$$

其中，$\boldsymbol{\beta}(t)=\left\{\beta_j(t)\,|\,j\in\{0,1,\cdots,M\},t\in\mathcal{T}\right\}$，分别对$\boldsymbol{\lambda}(t)$和$\boldsymbol{\beta}(t)$中的$\lambda_k^s(t)$与$\beta_j(t)$元素求偏导，得

$$\frac{\partial f\left(\boldsymbol{\lambda}(t),\boldsymbol{\beta}(t)\right)}{\partial\lambda_k^s(t)}=-\frac{V\sum_{s=0}^{S}q_k^s(t)\sum_{j=0}^{M}\frac{y_k^{j^*}(t)}{r_k^j(t)}}{\left(\lambda_k^s(t)\right)^2}+$$

$$\sum_{j=0}^{M}\beta_j(t)y_k^{j^*}(t)=0\Rightarrow\begin{cases}\lambda_k^{s^*}(t)=\sqrt{\frac{V\sum_{s=0}^{S}q_k^s(t)}{\beta_j^*(t)r_k^j(t)}},\ y_k^{j^*}=1\\ \lambda_k^{s^*}(t)=0,\ y_k^{j^*}=0\end{cases}\tag{19}$$

$$\frac{\partial f\left(\boldsymbol{\lambda}(t),\boldsymbol{\beta}(t)\right)}{\partial\beta_j(t)}=\sum_{k=0}^{N}\lambda_k^{s^*}(t)y_k^{j^*}(t)-C_j=0\Rightarrow$$

$$\sum_{k=0}^{N}\lambda_k^{s^*}(t)y_k^{j^*}(t)=C_j\tag{20}$$

进一步联立式(19)和式(20)，可得

$$\lambda_k^{s^*}(t)=\frac{C_j\sqrt{\frac{V\sum_{s=0}^{S}q_k^s(t)}{r_k^j(t)}y_k^{j^*}(t)}}{\sum_{k=0}^{N}\sqrt{\frac{V\sum_{s=0}^{S}q_k^s(t)}{r_k^j(t)}y_k^{j^*}(t)}}\tag{21}$$

通过式(21)可以求得所有用户的最佳通信资源分配策略$\boldsymbol{\lambda}^*(t)$。

同理可得，在已知全局最优的无线接入节点选择策略$\boldsymbol{Y}^*(t)$后，服务迁移问题P2_2表述如下。

$$\text{P2\_2}:\ \min_{\boldsymbol{X}(t)}\sum_{k=0}^{N}\sum_{j=0}^{M}x_k^j(t)\varsigma_k^j(t)$$
$$\text{s.t. 式(3)和式(4)}$$

由于$\sum_{i=0}^{M}x_k^j(t-1)\left(Q(t)E_{ij}^s(t)+D_{ij}(t)\right)$与$\frac{Vy_k^0(t)q_k^s(t)}{r_j(t)}$在$\boldsymbol{Y}^*(t)$和$\boldsymbol{\lambda}^*(t)$已知的情况下均为常数，因此$\varsigma_k^j(t)$为常数。其中$\boldsymbol{\Psi}(t)=\{\varsigma_k^i(t)\,|\,i\in\{0,1,\cdots,M\},\ k\in\{0,1,\cdots,N\},t\in\mathcal{T}\}$，即

$$\boldsymbol{\Psi}(t)=\begin{pmatrix}\varsigma_0^0(t) & \cdots & \varsigma_0^M(t)\\ \vdots & \ddots & \vdots\\ \varsigma_N^0(t) & \cdots & \varsigma_N^M(t)\end{pmatrix}$$

化简问题 P2_2，可得

$$\min_{\boldsymbol{X}(t)} \operatorname{sum}\left(\boldsymbol{X}(t).*\boldsymbol{\Psi}(t)\right)$$
$$\text{s.t. 式(3)和式(4)}$$

其中，符号 .* 表示矩阵点乘，符号 sum(·) 表示对矩阵所有元素求和。为减少计算复杂度，本文提出快速边缘决策算法求解最优服务迁移决策参数 $\boldsymbol{X}(t)$，如算法 1 所示。

算法 1　快速边缘决策算法

输入　初始化矩阵 $\boldsymbol{X}(t)=\boldsymbol{0}$

输出　最优服务迁移策略 $\boldsymbol{X}^*(t)$

1) for $k=0,1,2,\cdots,N$ do
2) $\quad\boldsymbol{\Psi}_c(t)=\boldsymbol{\Psi}(t)$
3) $\quad$while true
4) $\quad\quad\varsigma_k^j(t)=\arg\min_{k,j}\boldsymbol{\Psi}_c(t)$
5) $\quad\quad x_k^{j*}(t)=1$
6) $\quad\quad$if $\sum_{k=0}^{N}\sum_{s=0}^{S}\phi_k^s(t)x_k^j(t)>\phi_j$ 或者 $\sum_{k=0}^{N}\sum_{s=0}^{S}B_k^s(t)x_k^j(t)>B_j$ do
7) $\quad\quad\quad x_k^{j*}(t)=0$
8) $\quad\quad\quad\varsigma_k^j(t)=+\infty$
9) $\quad\quad$else 转到 11)
10) $\quad\quad$end if
11) $\quad$end while
12) end for

通过式(21)及算法 1 求解 $\boldsymbol{X}^*(t)$ 并化简问题 P2，使优化目标只包含决策参数 $\boldsymbol{Y}^*(t)$。由于化简后的问题仍是整数非线性规划问题。为在计算复杂度与用户的服务质量之间取得折中，本文进一步提出异步最佳响应算法，以用户为中心，可以为用户快速地选择无线接入节点。

5.2　异步最佳响应更新策略

令 $U(\boldsymbol{Y}(t))$ 表示问题 P2 的效用函数，用 $\boldsymbol{y}_{-k}=\{y_0,\cdots,y_{k-1},\cdots,y_N\}$ 表示除用户 k 之外，所有用户的边缘节点选择向量。在已知其他用户的边缘节点选择策略情况下，用户 k 通过选取某一个覆盖该用户的 SBS 或者 MBS 作为无线接入节点，使 $U(\boldsymbol{Y}(t))$ 最小化，如式(22)所示。

$$y_k^*=\arg\min U(y_k,\boldsymbol{y}_{-k},t) \tag{22}$$

由于上述问题为非合作性的无线接入节点选择问题，存在一个基于博弈论的无线接入策略，使所有用户的接入节点选择达到纳什均衡。因此，当任何用户都无法通过单方面更新其无线接入策略来进一步降低其成本时，所选择的无线接入策略为最优 $\boldsymbol{Y}^*$，即对于任意的用户 m 来说，都有式(23)所示的不等式成立。

$$U(y_m^*,\boldsymbol{y}_{-m}^*,t)\leqslant U(y_m,\boldsymbol{y}_{-m}^*,t) \tag{23}$$

本文根据式(22)与式(23)提出了异步最佳响应算法，其主要流程如下。用户将服务请求发送到初始连接的 BS 后，MBS 为其服务请求分配一个唯一编号 ID；然后根据分配的 ID 顺序更新所有用户的边缘节点选择策略，具体地，对当前还没有更新的最小 ID 的用户的服务请求分配一个更好的边缘服务器或云服务器。因此，任意用户 k 的无线接入节点选择更新流程如式(24)所示。

$$y_k(r+1)=\arg\min U(y_k,\{y_1(r+1),\cdots,y_{k-1}(r+1),y_{k+1}(r),\cdots,y_N(r)\},t) \tag{24}$$

其中，r 表示策略更新轮数。在已知当前所有用户的接入策略情况下，比用户 k 的 ID 小的用户根据不等式(24)进行更新，比用户 k 的 ID 大的用户接入策略保持不变。异步最佳响应算法如算法 2 所示。

算法 2　异步最佳响应算法

输入　初始化 $\boldsymbol{Y}=(y_1(1),y_2(0),\cdots,y_N(0))$，随机为每一个用户服务请求分配一个接入节点，并设置初始迭代轮数 $r=0$

输出　$\boldsymbol{Y}^*$

1) while $\boldsymbol{Y}$ 没有达到纳什均衡循环 do
2) $\quad$for $k\leqslant N$ do
3) $\quad\quad$根据式(24)更新用户 k 的边缘节点选择策略
4) $\quad\quad k=k+1$
5) $\quad$end for
6) $\boldsymbol{Y}$ 达到纳什均衡，转到 7)；否则，转到 8)
7) 更新迭代轮数 $r=r+1$，转到 2)
8) end while

由于用户的随机移动，使用户在每一个时隙之间的地理位置可能不同。通过快速边缘决策算法及异步最佳响应算法求解出每个时隙 t 的最优的边缘节点与无线接入选择策略 $(\boldsymbol{X}^*(t),\boldsymbol{Y}^*(t))$，然后不断更新每个时隙的虚拟队列长度，最后得到系统长期服务迁移决策。其流程如算法 3 所示。

算法 3　自适应最佳响应算法（AORAM, adaptive optimal response algorithm）

输入　初始化虚拟系统成本队列 $Q(0)=0$

1) for　$t=0,1,2,\cdots$　do

2) 运用算法 1 与算法 2 解决问题 P2，得到最优 $\boldsymbol{Y}^*(t),\boldsymbol{X}^*(t),\boldsymbol{\lambda}^*(t)$

3) 运用式(11)及 $\boldsymbol{Y}^*(t),\boldsymbol{X}^*(t),\boldsymbol{\lambda}^*(t)$，更新虚拟队列

4) end for

6　数值分析

本文采用 Python IDE PyCharm 作为实验的仿真软件，版本为 PyCharm Community Edition 2019.2.5。为缩短仿真时间，使用戴尔易安信 PowerEdge R730 机架式服务器（Xeon E5-2603 V3/8 GB/1.2 TB）搭建仿真平台。仿真场景设置如下：在 900 m×900 m 区域内部署 4 个 SBS 和一个 MBS，初始用户数 $M=50$。本文采用 3 种对比算法，分别为当前最优服务器迁移（COSM, current optimal server migration）、ISMM [12]与 FSMM [11]。COSM 为一个基准算法，为防止用户频繁进行边缘节点选择，在每个时隙开始阶段，只对位置发生变化的用户进行边缘节点选择，其他用户接入参数保持不变。ISMM 中，每隔一段时间 τ_0（实验设置为 4 s），BS 根据当前时刻 T_1 的用户位置信息进行服务迁移，并在下一个时间段 T_1–$(T_1+\tau_0)$ 内不进行任何服务迁移[12]。FSMM 基本思想为每个边缘服务器只服务其覆盖范围内的用户，无边缘服务器覆盖的用户均由远端云服务器提供服务[11]。

考虑到现实情况中，用户的运动轨迹可以分为确定性运动轨迹与非确定性运动轨迹，例如，用户每天的上下班路径及其所处地理位置可以根据大量的历史数据预测出来，而用户出差或者旅游的运动路径由于数据稀疏性难于预测。本文把用户的运动分为沿着确定轨迹运动与每个时隙的地理位置随机 2 种情况。具体仿真参数设置如表 1 所示。

6.1　确定性用户轨迹仿真模型

确定性用户轨迹模型如图 2 所示，其中，用户 user_1 与 user_2 分别沿着 2 个确定的路径轨道运动，其他用户的地理位置保持不变。用户 user_1 与 user_2 完成固定轨迹运动的时间均为 $T=800$ s，在每一个时隙 $\tau=1$ s 内运动长度在 x 轴上的分量为 1.125 m，其他用户的坐标位置在 $t=0$ 时刻随机分布。

表 1　　**仿真参数**

参数	取值
子信道带宽 W / kHz	180
小基站计算能力 ϕ_i / GHz	2
小基站存储能力 B_i /Gbit	2
小基站通信能力 C_i / MHz	18
服务请求总类别数 S	2
服务请求分组大小 q_k^s /kbit	(2, 3)
宏基站到小基站的下行传输速率 r_j /(Mbit·s^{-1})	20
宏基站到云服务器的上行传输速率 r_0 /(Mbit·s^{-1})	10
长期迁移预算成本 E_{avg}/Mbit	15
初始化惩罚因子 V	10 000
初始化队列长度 $Q(t)$/Mbit	0

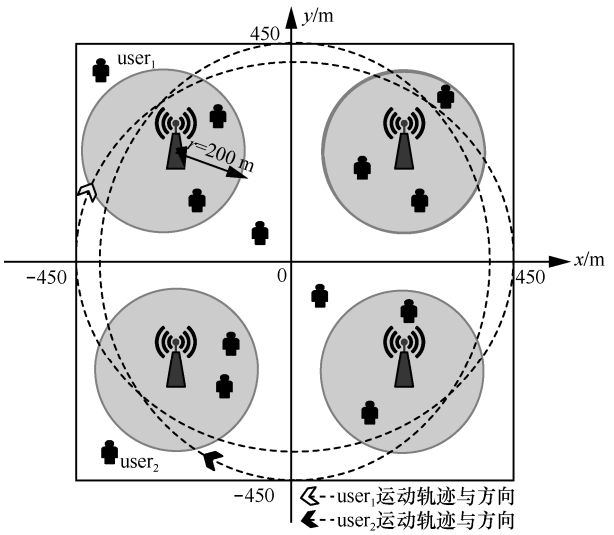

图 2　确定性用户轨迹模型

图 3 表示系统用户的平均感知时延。可以看出，本文所提 AORAM（即算法 3）得到的用户感知时延低于其他 3 种对比算法。与 AORAM 相比，其他算法均存在用户数量阈值 M_0，当用户数量 $M<M_0$ 时，对比算法与所提算法在降低用户感知时延性能方面表现大致相同；但当 $M\geqslant M_0$ 时，随着用户数的增加，AORAM 与对比算法的时延性能差异也随之增大，即所提算法明显优于对比算法。造成上述现象的主要原因在于：当用户数量相对较小时，本地边缘服务器的计算、存储与通信资源充足，能够最快地响应本地用户的服务请求，而当用户数大于某一个阈值时，容易造成某些边缘服务器负载过高，而某些边缘服务器则相反。AORAM 通过转发负载过重的边缘服务器的用户服务到空闲且离用户相对较近的边缘服务器，实现边缘服务器

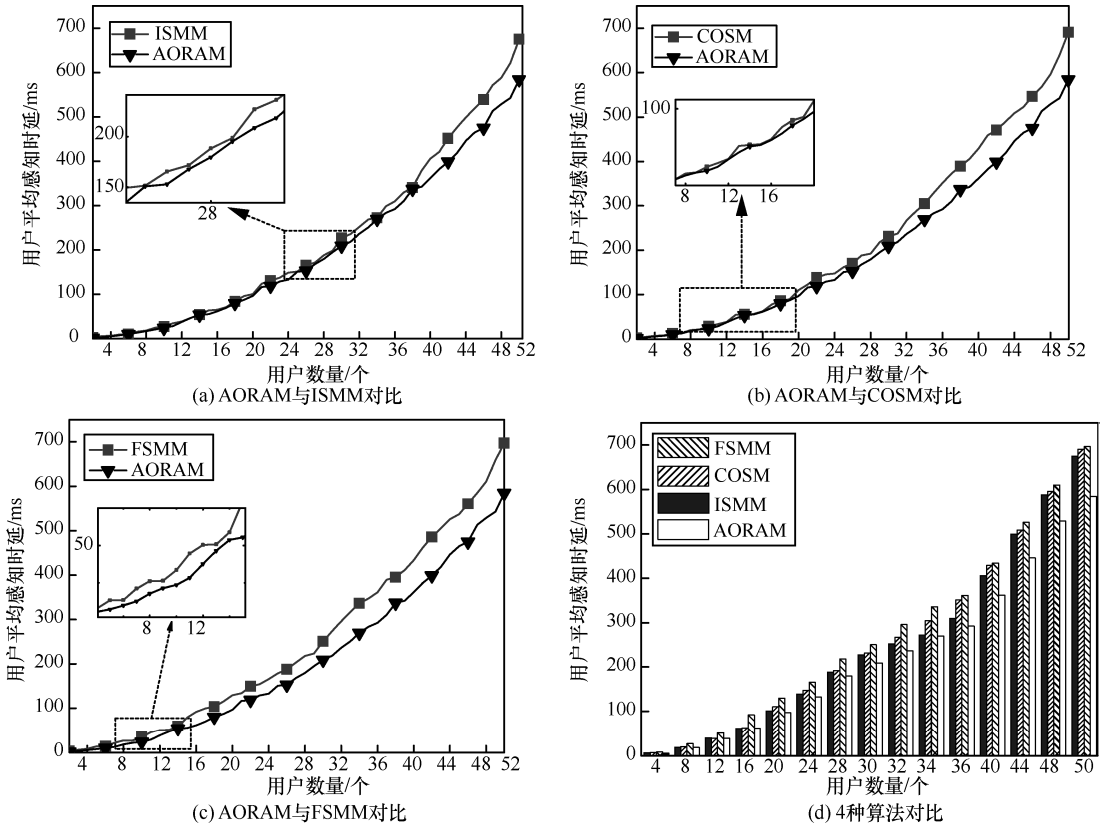

图 3　用户的平均感知时延

之间的负载均衡，从而降低用户的平均感知时延。此外，分别统计不同用户数量情况下，所提 AORAM 相对于其他算法的性能增益，并计算平均值。从图 3(a)中可以看出，与 ISMM 相比，AORAM 能够使用户的平均感知时延降低 8.746%，当用户数 $M \leqslant 38$ 时，2 种算法最大差值为 36.59 ms，即 2 种算法在用户服务质量性能方面表现大致相同，但用户数 $M > 38$ 时，时延差值随着用户数的增加而不断增大。同理可知，在图 3(b)中，与 COSM 相比，AORAM 能够使用户的平均感知时延降低 11.57%，当用户数 $M \leqslant 20$ 时，2 种算法时延差值百分比为 10.53%，最大差值为 24.68 ms。在图 3(c)中，与 FSMM 相比，AORAM 能够使用户的平均感知时延降低 21.59%，当用户数 $M \leqslant 12$ 时，2 种算法最大差值为 40.70 ms。从图 3(d)中可以看出，与其他算法相比，所提 AORAM 的用户平均感知时延分别降低了 8.74%、11.57%、21.59%。显然，AORAM 能有效地降低感知时延，提高用户服务质量。

图 4 为 4 种算法在不同惩罚因子 V 情况下虚拟队列长度 $Q(t)$ 随时间的变化曲线。图 4(a)为 AORAM 在不同惩罚因子 V 下，虚拟队列长度 $Q(t)$ 的变化情况，可以看出，随着时间推移，虚拟队列长度 $Q(t)$ 趋于某一个固定值（如当 $V = 500$ 时，虚拟队列长度 $Q(t)$ 的稳定值为 127）。此外，当增大惩罚因子 V 时，虚拟队列长度 $Q(t)$ 也随之增大，同时系统达到迁移成本稳定的收敛时间也变长。其原因在于，当增大惩罚因子 V 时，意味着系统优化目标更偏重于用户的感知时延，导致系统为获取更低的用户感知时延，需要进行更加频繁的服务迁移，从而增加了系统虚拟队列长度 $Q(t)$ 与达到稳定的时间。也就是说，为提高用户的服务质量，系统需要投入更多的服务迁移成本来换取更高的虚拟队列的积压容忍度。图 4(b)为 FSMM 在不同惩罚因子 V 下虚拟队列长度 $Q(t)$ 的变化情况。当用户 user_1 与 user_2 在如图 2 所示的单个 SBS 内移动时，其服务迁移决策保持不变，因此虚拟队列长度保持不变；当 user_1 和 user_2 跨 SBS 移动时，触发服务迁移策略，其虚拟队长随时间的增长而下降。图 4(c)表示 COSM 在不同惩罚因子 V 下虚拟队列长度 $Q(t)$ 的变化情况。可以看出，随着时间的推移，$Q(t)$ 逐渐下降到某一稳

定值。图 4(d)为 ISMM 在不同惩罚因子V下虚拟队列长度$Q(t)$的变化情况。当V值较小时，每个时刻迁移成本均小于长期迁移阈值E_{avg}，因此$Q(t)$值不断变小并最后达到稳定；当V较大时，频繁的服务迁移造成迁移成本大于迁移阈值，使其$Q(t)$不断增大直到稳定。

图 5 为不同惩罚因子V下用户的平均感知时延。可以看出，在平稳状态时，AORAM 得到的用户平均感知时延相对于其他 3 种对比算法分别减少了 90.84 ms、106.45 ms、112.55 ms。由于 FSMM 中服务迁移决策取决于用户与 SBS 的相对位置，所以V值不影响服务迁移决策，但容易造成边缘服务器负载失衡，导致服务质量降低。当$V<100$时，随着V的增加，ISMM、COSM 与 AORAM 的感知时延快速下降。但当$200 \leqslant V < 2000$时，AORAM 感知时延能够达到更低值，表明所提的 AORAM 在保

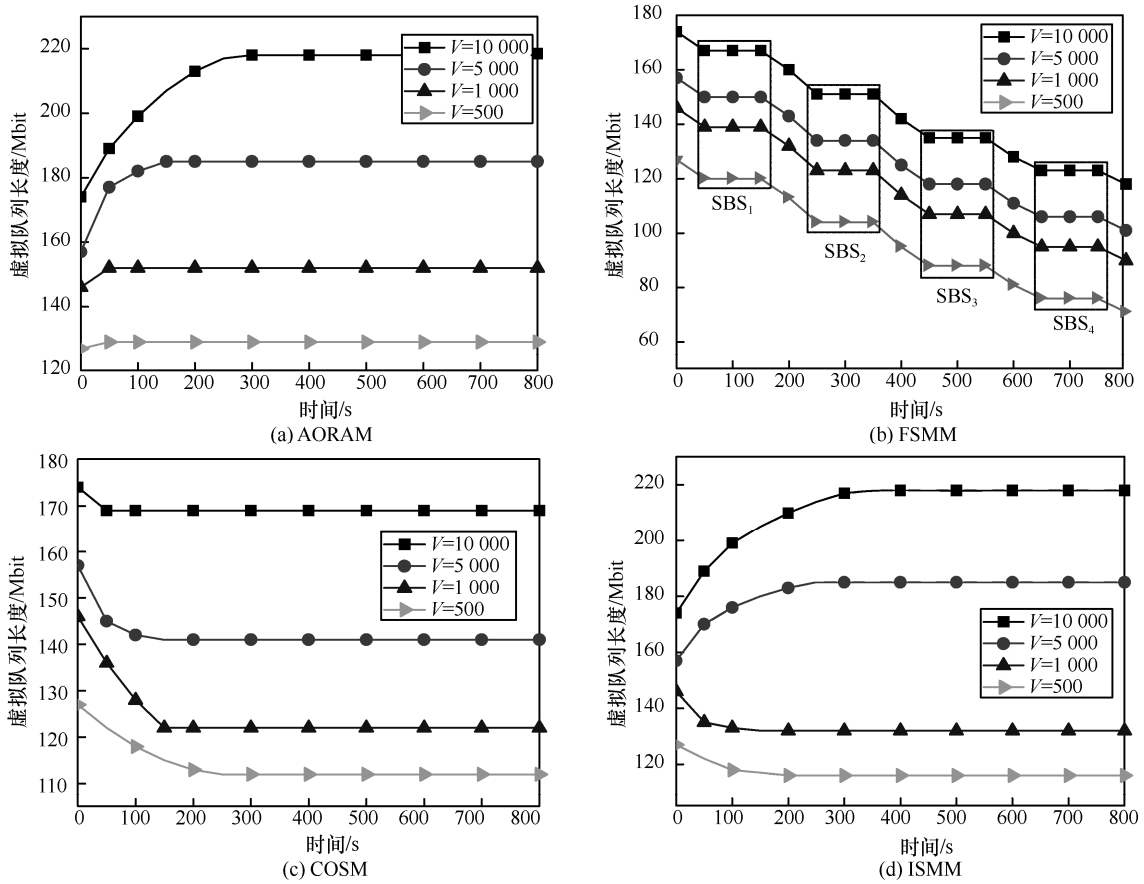

图 4　不同惩罚因子V下的虚拟队列长度

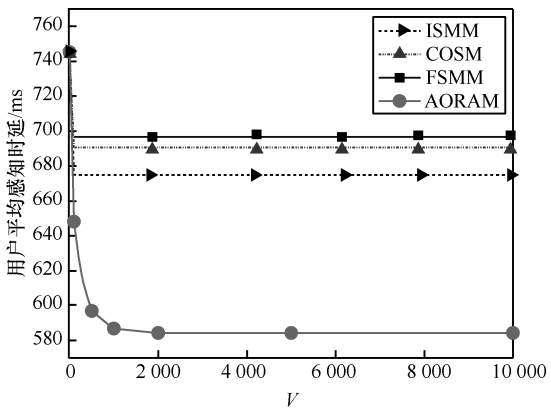

图 5　用户感知时延

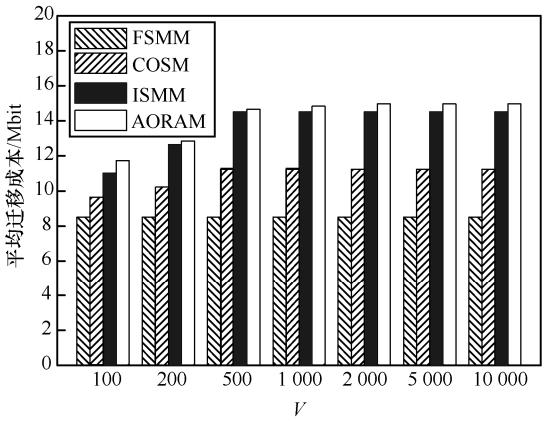

图 6　平均迁移成本

障更高服务质量的同时增大了长期服务迁移阈值的可调范围。图 6 表示不同惩罚因子 V 下系统的平均迁移成本。虽然 AORAM 相比于其他算法具有较高的迁移成本，但其迁移成本随着 V 值增大逐渐收敛，收敛值仍低于预期阈值 E_{avg}，结合图 5 可知，AORAM 在保持系统迁移成本相对稳定的情况下能够显著降低用户平均感知时延。

图 7 展示了不同长期迁移预算成本 E_{avg} 下虚拟队列的长度变化情况，其中 $M=20$。当 E_{avg}<15 时，虚拟队列长度随着时间的推移不断增大，无法收敛。根据式(11)可知，过小的 E_{avg} 将导致每个时隙迁移成本超出预期，使虚拟队列长度不断增大，出现虚拟队列长度随时间推移而无法收敛的情况。同理，当 E_{avg} 为 15~25 时，虚拟队列长度最终都随着时间的推移而收敛，且 E_{avg} 越大，收敛速度越快，收敛值越小。但是，过大的 E_{avg} 导致每个时隙的服务迁移成本都小于长期预算成本，如 E_{avg}=50，根据式(11)可知，虚拟队列长度急剧下降为 0。

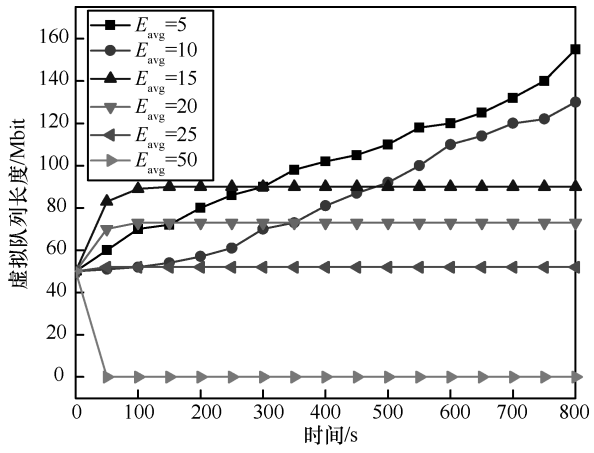

图 7　不同长期预算成本下的虚拟队列长度

6.2　非确定性用户轨迹的仿真模型

由于用户的运动路径具有不可预测性，因此在每个时隙 t 的开始时刻，用户的地理位置表现出随机性。为仿真该情形，在每一个时隙对系统中 15 个用户的地理位置进行随机化处理来代表用户的位置变化。对 3 000 个时隙进行了仿真，其中在每个时隙的开始时刻对用户的地理位置进行随机初始化处理，惩罚因子设置为 $V=10\,000$，迁移预算成本设置为 E_{avg}=35。图 8 展示了该场景下虚拟队列长度随时间的变化情况。当所有用户进行随机运动时，前一时刻所得到的边缘节点参数需全部更新为此时刻初始边缘节点参数，即 COSM 与 AORAM 算法相同。

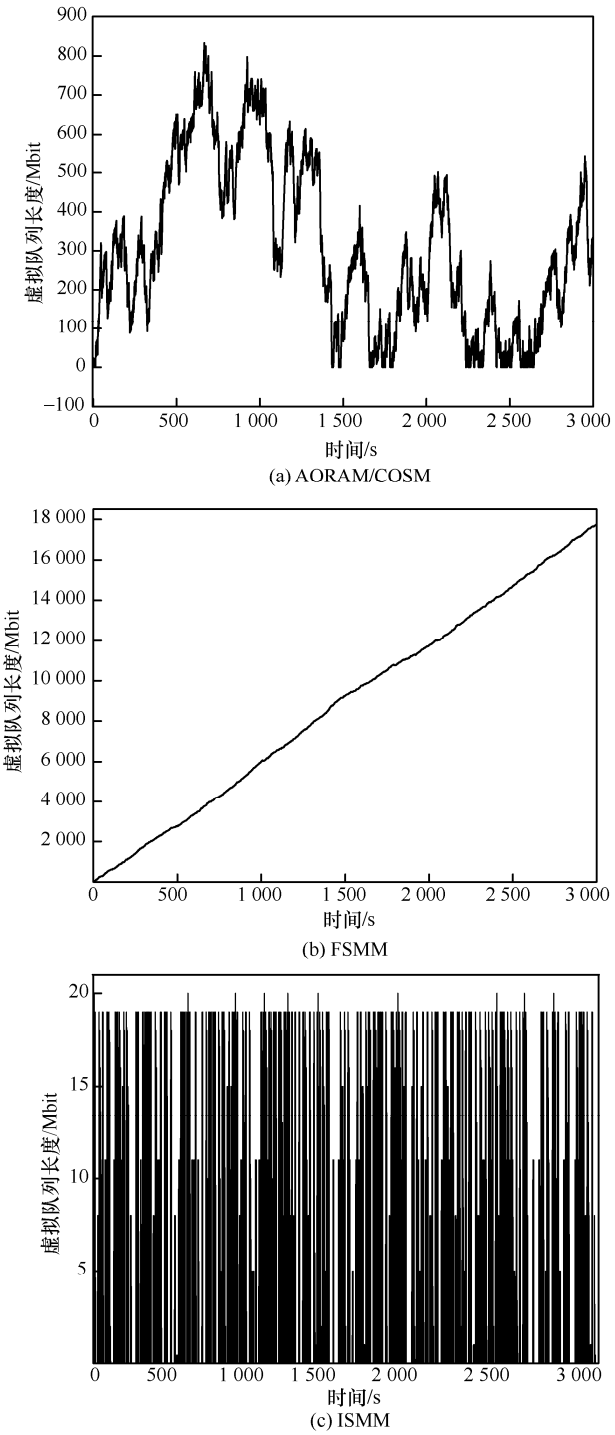

图 8　虚拟队列长度随时间的变化

从图 8(a)可以看出，当系统中所有用户进行无规则运动时，AORAM 和 COSM 得到的虚拟队列长度在 0~850 Mbit 之间波动。由于用户位置的随机性可能导致一段时间内大部分用户的位置没有大的变化，出现服务迁移成本低于迁移阈值的情况，根据式(11)可知，虚拟队列长度将会一直处于下降状态直到为 0。同理可得，造成虚拟队列长度接近 850 Mbit

的原因在于，在一段时间内用户在每个时隙内的服务迁移成本一直大于迁移阈值，导致虚拟队列长度一直处于上升阶段。由图 8(b)可以看出，当用户进行随机移动时，FSMM 的虚拟队列长度随着时间的推移不断增大，这表明长期迁移成本大于迁移阈值。从图 8(c)可以看出，由于 ISMM 要求系统每隔一段时间后进行服务迁移，导致虚拟队列长度出现为 0 的时间段占总时间的比例达 83.97%。通过分析可知，当用户进行无规则随机运动时，系统存在最小的迁移成本阈值，如果迁移预算成本 E_{avg} 大于该阈值，随着时间不断推移，虚拟队列长度将在某一范围内不断波动，相反，虚拟队列将不断增大，无法收敛。在这种情况下，用户的服务质量将严重失衡，例如，距离 SBS 更近的用户的服务质量更好，但距离 SBS 远的用户服务质量较差。过大的 E_{avg} 虽然能大幅度降低虚拟队列长度，提高用户的服务质量，但增加了服务提供商的服务迁移费用。

用户平均感知时延如表 2 所示。从表 2 可以看出，相比于 FSMM 与 ISMM，所提算法能够分别降低用户感知时延 18.45%、5.28%。

表 2　　用户平均感知时延

算法	用户平均感知时延/ms
AORAM	54.20
COSM	54.20
FSMM	66.46
ISMM	57.22

图 9 表示所提快速边缘决策算法（算法 1）与穷举法在不同惩罚因子 V 下的虚拟队列长度值。可以观察到，当惩罚因子 $V<3\,000$ 时，基于 2 种算法的虚拟队列长度曲线重合，表明了所提算法与穷举法所做出的服务迁移策略 $\boldsymbol{X}(t)$ 一致。当惩罚因子 $V>3\,000$ 时，不同的服务迁移算法所得到的系统性能出现差异，其主要原因在于，随着 V 值不断增大，系统通过放宽虚拟队列长度的容忍度，进行更好的服务迁移来进一步提高用户服务质量。另一方面，增大 V 值意味着 P2 优化目标更加偏重于用户的感知时延。通过计算得到所提算法与穷举法的平均虚拟队列误差值 $\sigma\approx3.75\,\%$。为进一步验证所提出的算法与穷举法在用户服务质量上的增益差异，仿真了不同用户数下 2 种算法的运行时间，如图 10 所示。可以看出，随着用户数的增加，2 种算法所需的平均运行时间也随之变大，所提算法在运行时间上远低于穷举法。

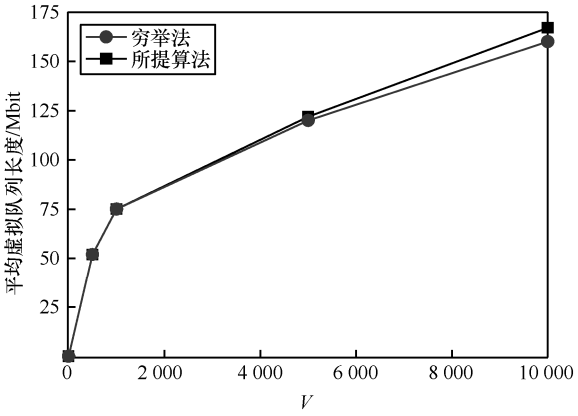

图 9　平均虚拟队列长度

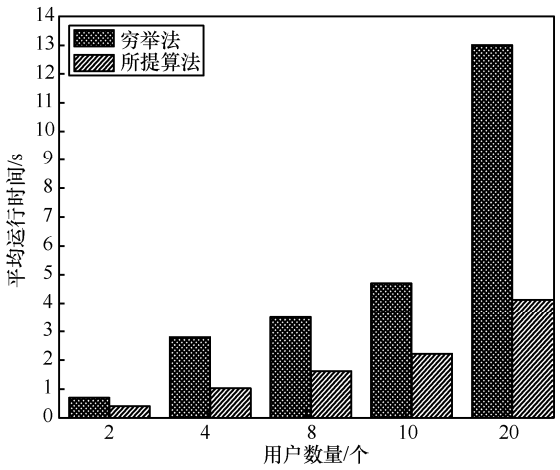

图 10　平均运行时间

7　结束语

本文研究了移动边缘计算网络中用户的移动性及其位置的动态性对用户服务请求的平均感知时延的影响，提出了一种移动性感知的在线边缘服务迁移算法。由于优化问题属于高度耦合时间与空间相关性的混合整数非线性规划问题，首先基于 Lyapunov 方法将优化问题解耦为边缘服务迁移与节点选择问题，其次提出快速边缘决策算法求解给定边缘节点选择策略下最优的资源分配与服务迁移策略，最后利用所提的异步最佳响应算法求解最优边缘节点选择策略。仿真结果表明，所提边缘服务迁移策略能够有效地降低用户服务请求的平均感知时延。本文方案为移动场景下在线服务迁移的研究提供了新的思路，然而在实际场景中，还需要考虑诸如服务类型、服务请求的最大容忍时延等因

素，未来工作将探索上述因素对在线服务迁移性能的影响，为实际网络提供更加准确的决策。

参考文献：

[1] DIAO X B, ZHANG J C, WU Y, et al. Joint computing resource, power, and channel allocations for D2D-assisted and NOMA-based mobile edge computing[J]. IEEE Access, 2019(7): 9243-9257.

[2] 田辉, 范绍帅, 吕昕晨, 等. 面向 5G 需求的移动边缘计算[J]. 北京邮电大学学报, 2017, 40(2): 1-10.
TIAN H, FAN S S, LYU X C, et al. Mobile edge computing for 5G demand [J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40 (2): 1-10.

[3] WANG S, ZHANG X, ZHANG Y, et al. A survey on mobile edge networks: convergence of computing, caching and communications[J]. IEEE Access, 2017(5): 6757-6779.

[4] 谢人超,廉晓飞,贾庆民,等.移动边缘计算卸载技术综述[J]. 通信学报, 2018, 39(11): 138-155.
XIE R C, LIAN X F, JIA Q M, et al. Overview of mobile edge computing offloading technology [J]. Journal on Communications, 2018, 39(11): 138-155.

[5] POULARAKIS K, LIORCA J, TULINO M A, et al. Joint service and request routing in multi-cell mobile edge computing networks[C]// IEEE INFOCOM 2019 – IEEE Conference on Computer. Piscataway: IEEE Press, 2019: 10-18.

[6] PASTERIS S, WANG S Q, HERBSTER M, et al. Service placement with provable guarantees in heterogeneous edge computing systems[C]//IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2019: 514-522.

[7] 郭辉,芮兰兰,高志鹏.车辆边缘网络中基于多参数 MDP 模型的动态服务迁移策略[J].通信学报, 2020, 41(1): 1-14.
GUO H, RUI L L, GAO Z P. Dynamic service migration strategy based on multi-parameter MDP model in vehicle edge network [J] . Journal on Communications, 2020,41 (1): 1-14.

[8] OUYANG T, ZHOU Z, CHEN X. 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.

[9] YU N, XIE Q Y, WANG Q Y, et al. Collaborative service placement for mobile edge computing applications[C]//2018 IEEE Global Communications Conference. Piscataway: IEEE Press, 2018: 1-6.

[10] CHEN M, LI W, FORTINO G, et al. A dynamic service-migration mechanism in edge cognitive computing[J]. ACM Transactions on Internet Technology, 2019, 19(2): 30.

[11] LU W, MENG X Y, GUO G F. Fast service migration method based on virtual machine technology for MEC[J]. IEEE Internet of Things Journal, 2019, 6(3): 4344-4354.

[12] HU M, WU D, WU W G, et al. Quantifying the influence of intermittent connectivity on mobile edge computing[J]. IEEE Transactions on Cloud Computing, 2019, doi: 10.1109/TCC.2019.2926702.

[13] WANG S Q, URGAONKAR R, ZAFER M, et al. Dynamic service migration in mobile edge computing based on Markov decision process[J]. IEEE/ACM Transactions on Networking, 2019, 27(3): 1272- 1288.

[14] WANG F, XU J, WANG X, et al. Joint offloading and computing optimization in wireless powered mobile-edge computing systems[J]. IEEE Transactions on Wireless Communications, 2018, 17(3): 1784-1797.

[15] CAO X W, WANG F, XU J, et al. Joint Computation and communication cooperation for energy-efficient mobile edge computing[J]. IEEE Internet of Things Journal, 2019, 6(3): 4188-4200.

[16] XING H, LIU L, XU J, et al. Joint task assignment and resource allocation for D2D-enabled mobile-edge computing[J]. IEEE Transactions on Communications, 2019, 67(6): 4193-4207.

[17] CHEN X, PU L J, GAO L, et al. Exploiting massive D2D collaboration for energy-efficient mobile edge computing[J]. IEEE Wireless Communications, 2017, 24(4): 64-71.

[18] HU Y, QIU C R, CHEN Y. Lyapunov-optimized two-way relay networks with stochastic energy harvesting[J]. IEEE Transactions on Wireless Communications, 2018, 17(9): 6280-6290.

[19] QIU C R, HU Y, CHEN Y. Lyapunov optimized cooperative communications with stochastic energy harvesting relay[J]. IEEE Internet of Things Journal, 2018, 5(2): 1323-1333.

[20] NEELY M. Stochastic Network optimization with application to communication and queueing systems[M]. San Rafael: Morgan & Claypool Publishers, 2010.

[作者简介]

吴大鹏（1979– ），男，黑龙江大庆人，博士，重庆邮电大学教授、博士生导师，主要研究方向为泛在网络、互联网服务质量控制等。

吕吉（1995– ），男，四川广安人，重庆邮电大学硕士生，主要研究方向为边缘计算。

李职杜（1990– ），男，海南澄迈人，博士，重庆邮电大学讲师，主要研究方向为边缘计算、网络演算等。

王汝言（1969– ），男，湖北浠水人，博士，重庆邮电大学教授、博士生导师，主要研究方向为泛在网络、多媒体信息处理等。