分 类 号_____ 学 号 D201477744

学校代码 10487 密 级_____

華中科技大學

博士学位论文

5G 网络移动边缘缓存与计算研究

学位申请人 : 郝义学

学科专业 : 计算机系统结构

指导教师 : 陈 敏 教授

答辩日期 : 2017 年 5 月 25 日

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Science

Mobile Edge Caching and Computing in 5G Network

Ph.D. Candidate : Hao Yixue

Major : Computer Architecture

Supervisor : Prof. Chen Min

Huazhong University of Science & Technology

Wuhan 430074, P. R. China

May, 2017

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：郝文学

日期：2017 年 5 月 25 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于 保密□，在_____年解密后适用本授权书。

不保密☑

（请在以上方框内打“√”）

学位论文作者签名：郝文学　　指导教师签名：陈孟

日期：2017 年 5 月 25 日　　日期：2017 年 5 月 25 日

摘 要

为满足大规模的移动设备接入和快速增长的通信容量的需求，small cell在下一代移动通信系统（5G）中将实现超密集部署，而且small cell的存储和计算资源能为移动应用（如增强现实游戏）提供无处不在的计算支持。但是该方案却会加重系统回程链路的负载，并且会带来巨大的能量消耗的问题。为解决上述问题，许多研究者提出了移动边缘缓存与计算的方案。

然而，现有的移动边缘缓存与计算方案存在以下问题：首先，现有的边缘缓存方案大多基于固定网络拓扑结构，忽略了用户移动性；其次，为解决5G网络高能耗的问题，采用可再生能量供电是一个可行方案，但是，可再生能量到达的随机性导致了边缘云服务器计算能力的动态性，使得现有基于电网供电的计算卸载策略难以适用；最后，由于用户移动性导致基于D2D (Device-to-Device) 的边缘计算(如移动微云)具有动态特征，可能会造成计算任务卸载的失败。面对上述问题和挑战，本文从以下四个方面展开研究：

(1) 针对边缘缓存中用户移动性问题进行研究。通过分析移动性对small cell和用户设备缓存的影响，提出了移动性缓存策略优化问题，并证明其是NP难问题。基于子模态优化，利用贪婪算法给出问题的解。实验结果显示，相较于传统的缓存策略，此策略在缓存命中率上有了明显提高。

(2) 针对边缘缓存中用户之间及用户与small cell之间接触时间的随机性进行研究。基于编码缓存建立了缓存命中率最大化的安置模型和能耗最小化的传输模型，通过对模型求解，提出绿色移动编码缓存策略。实验结果显示，与其他缓存策略相比，该策略具有最高缓存命中率和最低传输能耗。

(3) 针对可再生能量供电下移动边缘云计算进行研究。基于对可再生能量的分析，建立了用户计算任务时延和电网供电能耗最小化模型。利用交替优化将其分解为计算资源分配和任务安置两个子问题，通过求解子问题得出可再生能量供电下的计算任务卸载策略。实验结果表明，与随机计算卸载和均匀计算卸载策略相比，该策略能够至少缩短20%的任务延迟，节省30%的能耗。

(4) 针对移动边缘计算中连接不可靠的问题进行研究。本文突破传统的移动微云对D2D连接的依赖，提出了移动自组微云模式。同时分析了此模式的任务时延和能耗，得到最优卸载策略。最后给出了计算任务在远端云、移动微云和此模式下的选

择算法。实验结果证明，当任务处理前后比例小于1、用户接触频率大于0.0014时，此模式在延时和能耗方面均优于其他两种模式。

综上所述，本文所提出的移动边缘缓存与计算策略能充分利用网络边缘的存储计算资源、用户的移动性和动态的可再生能量供给，为用户提供缓存和计算的服务，提高用户的体验质量。

关键词：5G网络，移动缓存，边缘云计算，绿色通信，D2D通信，能效优化

Abstract

In order to meet the needs of large-scale mobile devices accesses and rapid growth of communication capacity, small cell will achieve ultra dense deployment in the next generation mobile communication system (5G), and storage and computing resources of small cell can be utilized to provide ubiquitous computing support for mobile applications, such as augmented reality game. However, this scheme will increase the load of the system backhaul link, and will bring huge energy consumption. In order to solve these problems, many researchers put forward the scheme of mobile edge caching and computing.

However, the existing schemes about mobile edge caching and computing have the following problems: First of al, most of the edge caching schemes are based on fixed network topology, while ignore the user mobility. Secondly, in order to solve the problem of high energy consumption of 5G network, it is feasible to use renewable energy. However, the randomness of renewable energy leads to the dynamic computing ability of the edge cloud server. Thus, the existing computing offloading strategies powered by main grid are difficult to be applied in this scenario. Finally, user mobility results D2D (Device-to-Device) based edge computing (mobile cloudlets) is dynamic, which can lead to the failure of the computational task offloading. Facing with the above problems and challenges, this thesis is conducted from the following four aspects:

1. User mobility in edge caching is studied. Through analyzing the effect of user mobility on caching of the small cell and the user device, this thesis puts forward the optimization mobility-aware caching strategy, which is proved that the problem is NP-hard. Based on the submodular optimization, the greedy algorithm is used to solve the problem. The experimental results show that, compared with traditional cache strategy, this strategy has significantly improved cache hit ratio.

2. Randomness of contact time between the user, and the user and small cell is studied. Based on the code cache, this thesis sets up the content placement model with maximization cache hit ratio and minimization transmission energy consumption. Through the solution of the model, this thesis proposes a green mobility-aware code cache strategy. The experimental results show that the proposed scheme has the highest cache hit ratio and the lowest

transmission energy consumption compared with other cache strategies.

3. Mobile edge cloud computing powered by renewable energy is studied. Based on the analysis of the renewable energy, a model for minimization the delay of user computing task and energy consumption of main grid is built. The model is decomposed into two sub-problems of computational resource allocation and task placement by using the alternating optimization method. By solving the two sub-problems, the computing task offloading strategy under the renewable energy supply is obtained. The experimental results show that the proposed algorithm can reduce at least 20% task delay and 30% energy consumption compared with the random computing offloading scheme and the uniform computing offloading scheme.

4. Connection unreliability of mobile edge computing is studied. This thesis breaks the traditional reliance on mobile cloudlet connection of D2D, and propose the opportunistic mobile ad hoc cloudlet service mode (OCS). At the same time, the task delay and energy consumption are analyzed, which result in optimal offloading strategy. Finally, the selection algorithm of computing tasks in the remote cloud, mobile cloudlet and this mode is given. The experimental results show that this model is superior to the other two models in terms of delay and energy consumption when the ratio of data size after task execution over original data size associated with the task is smaller than 1 (*i.e.* $r < 1$) and the contact rate of two mobile devices λ is larger than 0.00014.

In summary, mobile edge caching and calculation proposed by this thesis can make full use of computing resources of edge cloud, users mobility and the dynamic renewable energy, which can provide cache and computing services for users, and improve the quality of the user experience.

Key words: 5G network; Mobile caching; Edge cloud computing; Green communication; D2D communication; Energy efficiency

目　录

1 绪论

本章介绍了5G移动边缘缓存和计算的研究背景和意义，概述了国内外研究现状，从而引出本论文的研究问题与主要贡献，最后给出论文组织结构。

1.1 研究背景与意义

在过去的几十年里，无线通信技术和云计算技术取得了巨大的进步，智能手机和平板电脑等移动设备已经渗入到人们的日常生活中，为用户提供了便捷的服务，比如通过移动手机观看实时视频，利用移动设备进行实时导航定位，基于手机的增强现实游戏等。然而，这些移动应用给人们带来方便的同时，也给现在的网络带来了巨大的挑战。首先，大规模移动设备接入的挑战。预计到2020 年，全世界的无线设备总量将达到750亿，而且根据思科报告[1]，到2021年全球的移动流量将会超过35艾字节/月(exabytes/month)，如图1.1所示。所以未来的通信网络需要支持大规模的设备接入，满足快速增长的通信容量需求。其次，为移动设备提供无处不在的计算支持的挑战。随着移动设备的日益智能化，移动应用需要覆盖范围更广泛的计算和持续性的数据处理，同时这些应用具有较高的时延需求，但是移动设备通常受电池的容量和功耗的约束，所以这些对计算能力极度渴望的新兴应用只能依靠先进的计算卸载方式和新型网络架构。因此，如何设计既能满足大量终端的通信接入，又能为终端提供强大计算支持的未来的通信网络是一个重要的研究课题。

为满足随时随地的接入和计算支持，下一代移动通信网络系统（5G）将实现small cell(也称作小基站，一般相对于5G超密蜂窝网中的macrocell，宏基站)的超密集部署[2–4]。利用small cell的超密集部署可以增加空间的复用，进而提高频谱效率，也一定程度上增强了网络的吞吐量[5–7]。与此同时，超密集部署的small cell的存储和计算资源可为用户设备提供边缘云计算（edge cloud computing）服务[8]，将用户设备计算密集型或延时敏感型的任务卸载到传输延时小的边缘云上完成，能够提高任务的延时保障。然而5G网络加剧了回程链路（backhaul link，即基站到核心网的链接）的负载[9]，而且超密集的small cell 部署也将带来巨大的能量消耗，这显然违背了5G 网络绿色性的需求[10]。

为了克服5G网络中回程链路容量的瓶颈和能量消耗巨大的问题，研究者们通过对移动流量的观察发现[11,12]，不仅大多数的流量请求来源于高质量的多媒体视

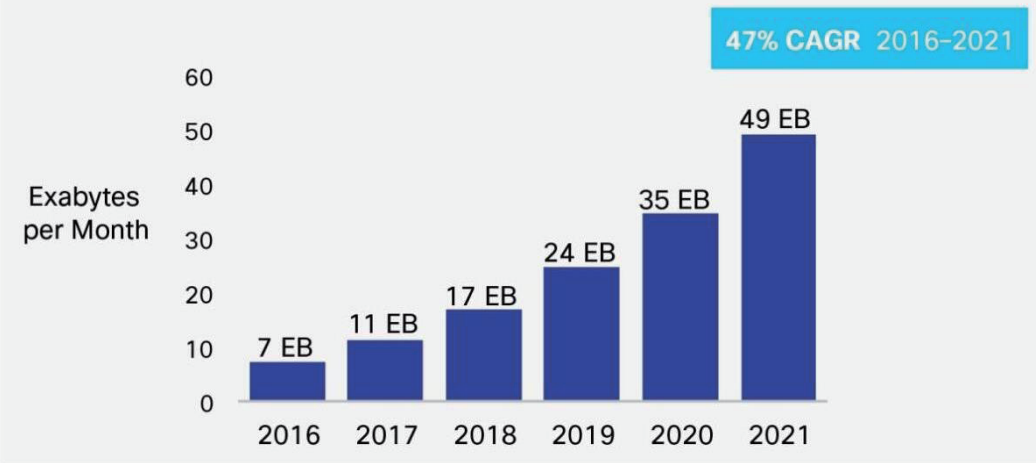


图 1.1 思科预测移动数据流量增长示意图

频流应用，而且是由少数内容产生的，即人们对内容的请求存在很大的重复性，比如流行的视频内容经常被重复的请求。进一步发现，不同于一般的数据应用请求，这些以内容为中心的请求可以进行缓存，比如在small cell 或移动设备上进行内容缓存，可以使用户从small cell 或其他移动设备上获得请求的内容，从而减少回程链路负载[13]，而且有利于减少任务请求延时和通信能量的消耗[14]。针对能量消耗巨大的问题，采用可再生能源（如太阳能、风能等）对small cell 进行供电已被证实是一个可靠的方案[15–17]，并且引起了学术界和工业界的广泛关注，如Marsan 等人[18]和Dhillon 等人[19]基于随机几何理论研究了可再生能源单独供电的蜂窝网络的可行性及其性能；通信设备商（如华为[20]）已生产和部署能够使用可再生能源进行供电的基站。值得注意的是，大量移动设备的剧增也带来了机遇[21]，随着移动设备日益增强的存储和计算能力[22]，研究者提出了基于Device-to-Device(D2D)[23,24] 的边缘计算（edge computing）[25]，此种计算方式能够减少核心网络的流量消耗，减少计算延时，满足移动用户的通信计算需求[26]。然而现有的边缘缓存和计算卸载策略在现实应用中普遍存在着局限性：

（1）现有的边缘缓存策略大多是基于固定的网络拓扑结构[27,28]，却忽略了用户移动性对边缘缓存的影响。而用户的移动性是5G无线网络的重要特征，且用户的移动性能够增加用户与基站，以及用户之间的通信机会[29]。此外，需要考虑到5G 网络的绿色性，在缓存内容传输时尽可能的减少small cell 和用户设备的能量消耗。因此考虑到用户的移动性，如何在边缘部署能效优化的缓存策略是一个亟待解决的研

究热点。

（2）现有的关于移动边缘云计算的研究大多是考虑电网供电[30,31]，然而由于可再生能量到达具有随机性以及small cell 的电池容量有限，导致了边缘云服务器计算能力的动态性，为计算卸载策略的优化增加了能量随机性和计算能力动态性的约束，所以现有的计算卸载方案并不适用。因此，设计出既能保证用户任务低时延性要求，又能充分利用可再生能源的卸载策略是一个值得研究的问题。

（3）现有的基于移动设备边缘计算的相关研究大多是利用用户周围闲置的移动设备，通过D2D通信的方式来完成用户计算任务的卸载[26,32]，但是没有考虑到由于用户移动性和网络动态性导致的移动边缘计算不可靠性。因此设计出既能利用用户设备的计算资源，又能增强任务完成度的计算策略是一个重要的研究课题。

综上所述，移动边缘缓存与计算作为5G网络发展的两个重要的挑战和机遇，能否充分利用异构且多样化的存储计算资源，用户的移动性，动态的可再生能量供给来提高缓存和计算的服务，降低网络能耗，具有兼备理论和实用的双重意义。

1.2 国内外研究现状

本节将介绍5G网络移动边缘缓存与计算的国内外研究现状。移动边缘缓存与计算利用了5G网络边缘侧的存储和计算能力，使得资源（包括无线访问、存储和计算）更接近用户请求的位置，从而减少了内容请求和任务卸载的延时，降低了传输的能耗，提高了用户的体验度和实现了5G的绿色性[33,34]。其主要针对下面两类服务，第一类为基于网络推送的服务，此类服务主要是由内容提供商发起，包括基于增强现实的移动场景，情境感知计算等。第二类基于用户请求的服务，此类包括了用户主动启用的服务，如工作和娱乐多媒体处理服务等[35]，这类服务的共同点是由用户主动从周边环境中获取他们所需的服务。

1.2.1 移动边缘缓存

移动边缘缓存能使用户从small cell或其他设备处获得请求的内容，实现了内容的本地可用，而不需要通过移动核心网和有线网络从内容服务提供商获取内容[14]。从而减少无线需求容量和可用容量之间的不均衡[36]，缓解了5G 网络的回传瓶颈，提高延时保障，降低网络能耗。边缘缓存一般包括两个步骤，内容的放置和传递，内容放置包括确定缓存的内容，缓存的位置以及如何将内容下载到缓存节点；内容的传递指的是如何将内容传递给请求的用户。一般来说，在网络流量较低、网络资

源廉价而丰富时（例如清晨），执行内容的放置。当网络流量较高时、网络资源稀缺和昂贵时（例如晚上），执行内容的传递。现有的工作研究主要集中在缓存的位置，缓存的形式以及缓存的内容三个方面[37,38]。与此同时，考虑到随着small cell 基站的致密化部署以及移动设备的激增，导致用户在小区之间的切换越来越频繁以及用户之间的D2D通信的机会越来越多，从而对缓存的影响度也越来越大。因此，本章将从缓存的形式、缓存的位置、缓存的内容和用户移动性四个方面来介绍。

缓存形式：缓存的形式一般分为编码缓存和非编码缓存，其中编码缓存是由[36]第一次提出的，编码缓存可以将每个文件分成几个互不重叠的编码段，每个基站或移动设备可以缓存不同的编码段，通过这些编码段可以将源文件恢复[39]。而非编码缓存一般假设文件完全缓存在基站或用户设备上，或者不缓存在基站或用户设备上。而对于编码缓存，一般假设基站或移动设备只存储编码文件的一部分，整个文件可以通过收集该文件的编码信息获取[40][36]。

缓存内容：研究者发现，流行的内容经常被请求（如流行的视频内容）。所以针对缓存内容，首先需要关注的就是缓存文件的流行度。缓存文件的流行度指的是一定区域内文件库中每个文件被所有用户请求的概率。根据参考文献可得[41]，内容的流行度服从Zipf分布，此分布可以通过文件库的大小和流行度偏置参数来表示。一般来说，内容流行度分布的变化速度比蜂窝网络的流量变化慢的多，通常在长时间内近似为常数（如电影的流行度通常为一周，消息的流行度通常为2 或3个小时[42]）。然而一个大区域（如一个城市甚至一个国家）和一个小区域（如校园）流行的内容往往是不同的[43]。此外，一些研究者给出了如何获得内容的流行度的方法，比如基于内容随时间的累积统计[44]。

另一个与缓存内容相关的因素就是用户对内容的喜好程度。这是因为用户通常对特定类别的内容有强烈偏好[43]，通过缓存此类内容，可以提高缓存命中率（请求的内容恰好在缓存服务器上），不同于文件流行度的定义，用户喜好指的是特定用户在一定时间内请求文件的概率。用户对内容的喜好可通过用户请求的历史数据，通过推荐算法（如协同过滤）来预测[45]。

缓存位置：5G网络中边缘缓存的位置主要有基站和用户移动设备[13]。与内容中心网络的缓存不同，边缘缓存需要考虑其特殊性（比如基站小区的干扰，用户设备的移动等）。（1）对于基站的缓存，可以在非高峰其将缓存内容提前部署在宏基站或small cell。small cell可以分为两种：一种是有回传链接的，一种是没有回传链接的（一般称为helper）[46]。对于small cell 的缓存，已有大量的研究，如Shanmugam

等人[40] 提出了femto缓存，通过部署分布式的femto-cell基站缓存，能够明显减少宏基站流量的传输。Golrezaei等人[42] 研究了helper灵活且高效的部署方案，通过缓存在helper上流行的内容，能够降低网络流量的传输，提高用户延时保障。（2）用户移动设备缓存，即请求内容的移动设备可以通过D2D 通信从缓存该内容的用户处获得，而不需要通过基站等获取。Ji等人[47–49]提出了如何在移动设备上部署缓存策略能提高D2D网络的容量，并且分析了缓存部署方案的一些局限性。Malak等人[50]提出了一种高效的单小区D2D缓存网络。

用户移动性：用户的移动性是边缘缓存的一个重要特征[51,52]。下面主要从空间和时间两个角度介绍现有工作对用户移动性的描述。空间角度指的是与用户移动模型相关的物理位置信息，时间角度指的是与用户移动模型相关的时间信息。

空间角度：用户的移动轨迹（即用户的移动路线）可以对用户移动性进行细粒度的描述。通过用户的移动轨迹，可以得到用户与small cell、宏基站之间的距离，如Bettstetter等人[53]利用了随机航点移动模型对用户轨迹进行了描述。用户在服务小区之间的切换，即用户从一个小区移动到另一个小区，可以得到用户在服务基站的信息，因此也可以描述用户的移动性[54]。这种描述和用户轨迹相比，由于不能具体到用户在每一个小区的移动轨迹，因此服务小区切换含有较少的细粒度信息。但在服务小区的切换过程可用马尔可夫链模型来描述[55]，其中，马尔可夫链模型中的状态数目等于基站数目，转移概率等于一个用户从一个基站移动到另一个基站的概率。此外研究者还发现，用户移动性模型很大程度上取决于用户之间的社会关系[56,57]。比如，Wang等人[58]提出了一种双级移动模型。首先，建立一个社交图，其中节点表示移动用户，加权边表示移动用户之间的社会关联程度。之后建立社会团体，研究每个团体中用户的移动性。空间角度主要描述了用户在基站小区内的移动信息。现在也有一些工作研究了用户移动性对基站缓存的影响，如Wang等人[59]讨论了文件非编码情况下基于用户移动性的femto缓存策略部署方案。Poularakis 等人[55,60] 利用马尔科夫模型建立了用户在small cell之间的移动性模型，并基于此模型给出了在small cell 上的缓存部署，能够最大化减少宏基站上的流量消耗。

时间角度：两个移动用户通信的频率和持续时间可以描述用户的移动性[61,62]。根据已有的工作[63]，任意一对用户的通信频率和通信时间可以使用接触时间（contact time）和接触间隔时间（inter-contact time）来表示。其中接触时间定义为一对移动用户在彼此的传输范围内的持续时间，接触间隔时间定义为两次连续接触时间之间的间隔时间。此移动模型已经广泛的应用到了无线网络中[28,32,64,65]。根据研

究表明[32,65]，用户的接触时间和接触间隔时间服从指数分布。用户在小区内的停留时间也可以描述用户的移动性。小区停留时间是指用户在特定基站的服务时间，这可能会影响用户从此基站接收到的数据量。此外，用户的返回时间也可以描述用户的移动性，用户的返回时间指的是任意用户返回到先前访问区域的时间间隔，它反映了用户移动的周期特性和用户重新访问特定区域的频率。Gonzales等人[66]研究了返回时间的分布，并计算出返回时间的峰值，从时间的角度可以刻画用户在small cell的移动情况或D2D网络中的用户移动性。现在也有一些工作研究了移动性对移动设备缓存的影响，如Lan 等人[67] 首次提出了在D2D 网络中的缓存策略，Wang 等人[28,64] 进一步讨论了D2D网络的编码缓存和非编码缓存策略

综上所述，目前相关的研究主要集中在如何在small cell和移动设备上进行缓存内容的部署，对移动性的缓存做了初步的研究，仅仅考虑到移动性对small cel缓存策略的影响或对移动设备缓存策略的影响，鲜有综合考虑移动性对small cell和移动设备的影响，目前也没有研究考虑到接触时间的有限性对缓存策略的影响。于是用户移动性对缓存的影响仍是一个值得研究的问题。

1.2.2 移动边缘计算

近年来，云计算（cloud computing）成为被广泛认可的最先进的计算基础设施，在虚拟化的基础上，云计算实现了在数据中心上同时运行多个操作系统和应用[68]，并且保证了多个操作系统和应用的隔离，从而能够保护在云端运行的程序和数据[69]。因此，可以将终端计算密集型任务卸载到云端，利用云端丰富的资源和计算能力，来提高终端计算的速度。而移动云计算（mobile cloud computing）指的是通过移动网络的云计算。当然，移动云计算也需要克服很多相关的实际挑战，如性能[70]、环境和安全等。然而，由于计算任务在端到端的传输过程中存在大量的时延以及频谱资源有限导致了无线接入网络的吞吐量不足，所以移动云计算服务在部署和维护方面变得越发困难[30]。于是，研究者提出将一些计算资源部署在离用户较近的位置，如基站附近，此时用户可以直接通过无线信道进行计算任务的卸载（称为移动边缘云计算mobile edge-cloud computing）[71]，或直接利用移动终端日益增强的计算能力，基于设备到设备通信（device-to-device communication）[72]来完成计算任务的卸载（称为移动边缘计算mobile edge computing 或移动微云mobile cloudlets），此模式能够显著的减少计算的延时和能耗。本小节从基于远端云的计算任务卸载，基于边缘云的任务任务卸载和基于移动微云的任务卸载三个方面来介绍国内外研究

现状。

基于远端云的任务卸载：基于传统的云计算，Clone-cloud[73]，ThinkAir[74]等为移动云计算的实现提出了可行方案。移动用户可以将计算密集型任务卸载到云端，在云端完成计算任务后，将结果返回给移动用户[75,76]。移动用户可以通过两种方式将计算任务卸载到云端，一种是通过WiFi的方式，另一种是在WiFi不可用时，通过蜂窝网络（比如3G/4G/5G网络）。因此一个主要的问题就是：移动用户应该根据实时的通信状态决策通过何种方式将计算任务卸载到云端。Barbera等人[77]通过真实的实验给出了如何根据实时带宽，给出终端能耗最小的计算卸载策略。Flores 等人[78]给出了环境感知的计算任务卸载策略，即根据实时通信状态来决定是否卸载。对于移动设备上的平行任务卸载，Li等人[79] 设计了一类启发式卸载机制，能够使得用户任务的延时最短。针对蜂窝网下的云计算，Lei等人[80] 首次探讨了移动云计算的卸载决策和无线异构网络中无线资源管理的相互作用，在此架构下，移动用户可以享受高质量的云服务，而不考虑频谱资源的缺乏。

基于边缘云的任务卸载。移动边缘云计算首先由[81]提出，边缘云可以分为以下两个部分：small cell云和宏小区云，前者由部署在small cell 上的计算资源构成，后者由部署在宏基站上的计算资源构成。small cell云由于受其硬件条件的制约，导致计算资源有限，所以能够提供的计算服务有限。但由于small cell 云离移动终端较近，用户可以直接通过无线信道与small cell云相连，所以其延时较短。宏小区云也提供一定的计算服务和保证较短的计算时延，但其计算资源仍然是有限的。文献[81]给出了移动边缘计算的系统介绍，从移动边缘云的部署、移动边缘云的移动管理、移动边缘的能效优化等方面系统地介绍了移动边缘计算。对于具体的用户计算卸载策略，Chen等人[82] 研究了多用户在边缘云上的计算任务卸载策略，利用纳什均衡给出了一种有效的多用户在边缘云上卸载方案，能够使得计算的延时和能耗较小。Hao 等人[83] 给出了在用户移动性环境下，如何在宏基站云和微基站云上进行计算任务的卸载，能够使得延时最短。

基于移动微云的任务卸载。微云（Cloudlet）的概念由Satyanarayanan等人[84]首次提出，Miettinen 等人[85] 做了进一步研究，指出微云是一种将服务器放在网络边缘的全新架构。微云一般放置在人群密集的公共或商业场所（比如机场，火车站和咖啡馆等），能够为移动设备提供较为丰富的计算资源[30]。微云计算也称作边缘计算[86]。但是部署和维护微云仍然是十分昂贵的，而且微云也不能解决用户移动性的问题。不过，随着移动设备存储和计算能力的发展，Li 等人[32] 提出了移动微云的概

念，给出了移动微云大小和寿命的定义，并解决了移动微云在什么条件下能够提供移动应用服务的问题。基于移动微云，Wang 等人[87] 提出了机会主义微云卸载机制，给出了基于机会主义的移动微云卸载方案，能够在规定时间内完成计算任务，同时系统消耗能效最优。

总而言之，移动云计算和移动边缘计算系统主要关注三点：减少延时，降低终端能耗以及提高系统能耗的效率。针对上面的三个目标，研究者设计了各种不同的计算任务卸载方案去保证任务的延时和能耗。然而现有的移动边缘云计算方案主要考虑电网供电，很少有研究关注可再生能源，而且由于可再生能量到达的随机性导致了计算能力的动态性，为计算任务的卸载带来了新的约束，所以上面提到的方法并不适用。然而针对基于D2D的移动边缘计算，值得注意到是，只有当有任务的用户（任务节点）和处理任务的用户（服务节点）在通信范围内时，计算任务才能够进行任务的卸载。但是由于D2D连接的随机性以及用户的移动性导致网络具有动态性，可能导致任务节点和服务节点之间的连接断开，从而导致任务卸载的失败。

1.3 论文的研究问题与主要贡献

本论文针对5G网络中存在的回程链路容量瓶颈和巨大的能耗问题，从移动边缘缓存与计算两个角度展开研究，提出了5G 网络移动性缓存策略和绿色移动缓存策略。这两种缓存策略不仅能够提高缓存命中率，减少回程链路负载，而且有利于减少网络能量消耗。进一步，本论文设计了可再生能源供电下的计算任务卸载策略和基于机会主义的移动自组微云计算任务卸载策略，这两种计算卸载策略不仅能够满足任务时延的需求，而且能够减少能量的消耗。具体来讲，本文的主要贡献和创新点如下：

提出5G网络移动性缓存策略：移动设备数量的急剧增加，导致出现了越来越多的流量和重复的内容请求。研究发现基于small cell和移动设备的缓存能够有效地减少高峰期回程链路的移动流量。然而，目前大多数的缓存策略研究都是假设固定的网络拓扑，很少有研究考虑到用户移动性对缓存策略的影响，而用户移动性是缓存网络的一个重要特征。针对此问题，本文提出了移动性缓存策略的优化问题，并证明其是NP难问题。通过子模态优化对问题进行转化，利用贪婪算法给出了问题的解。实验结果表明，本文提出的移动性缓存策略比其他现有的缓存策略更有效，同时得出当用户的移动性较低时，small cell 和用户设备应该缓存流行度较高的文件；当用户的移动性较高时，small cell 和用户设备应该考虑文件的多样性进行缓存。

提出5G网络绿色移动编码缓存策略：考虑到用户之间及用户与small cell之间接触时间的随机性，可能导致请求文件传输的失败。然而，现有的研究大多假设每次接触都能够传输固定的文件，这显然是与实际不符的。针对此问题，本论文从编码缓存的角度出发，通过对接触时间的动态性进行分析，提出了缓存命中率最大并且传输能量消耗最小的优化问题。针对此问题，本章通过子模态优化给出了缓存文件的安置策略，进一步给出了small cell 基站和移动设备基于最优发射功率的传输策略。实验结果显示，与其他缓存策略相比较，本文提出的缓存策略具有最高的缓存的命中率，最低的传输能耗。同时得出当用户移动性较高时，缓存内容的small cell和设备消耗的传输能量较少；当用户移动性较低时，其消耗的传输能量较多。

提出可再生能量供电下5G移动边缘云计算框架和卸载策略：由于现有的移动边缘计算卸载方案都是基于电网供电的，为满足5G网络绿色性需求，采用可再生能量进行供电是一个可行的方案。但是由于可再生能量到达的随机性导致了移动边缘云服务能力的动态性，为计算卸载策略的优化带来了计算能力动态性的约束。因此现有的方案均不适用。针对此问题，本论文首次提出了可再生能量供电下的5G移动边缘云计算框架。基于可再生能源的分析，建立了任务时延和电网能耗最小的优化模型，采用交替优化将其分解为两个子问题：计算任务的安置（计算任务是卸载到部署在small cell 的边缘云上，还是卸载到部署在宏基站的边缘云上）和计算资源的分配（对卸载到边缘云上的计算任务，边缘云如何对计算任务进行计算资源的分配）。通过求解子问题得到可再生能量供电下的计算任务卸载策略。实验结果表明，对比于其他计算任务卸载方案，此策略至少能够缩短20%的任务延时，降低30% 的能量消耗。

提出一种新颖的计算任务卸载模式并给出其卸载策略分析：随着移动设备的剧增和D2D通信的发展，基于D2D的移动边缘计算(如移动微云)引起了研究者的广泛关注，然而由于用户移动性导致了这种计算模式具有动态特性，从而造成计算任务卸载的失败。针对此问题，本文提出基于机会主义的移动自组微云计算模式。基于用户移动性模型给出了此模式的时延和能耗分析，提出了任务时延和系统能耗最小化的任务卸载策略，继而给出了计算任务在远端云、移动微云和移动自组微云的选择算法。实验结果表明，本文提出的卸载模式在任务前后处理比例小于1、用户接触频率大于0.0014时，优于其他两种模式，同时得出给具有较高移动性和较大计算能力的服务节点分配的工作量越多，越能够降低网络中的能耗，从而提高系统的性能。

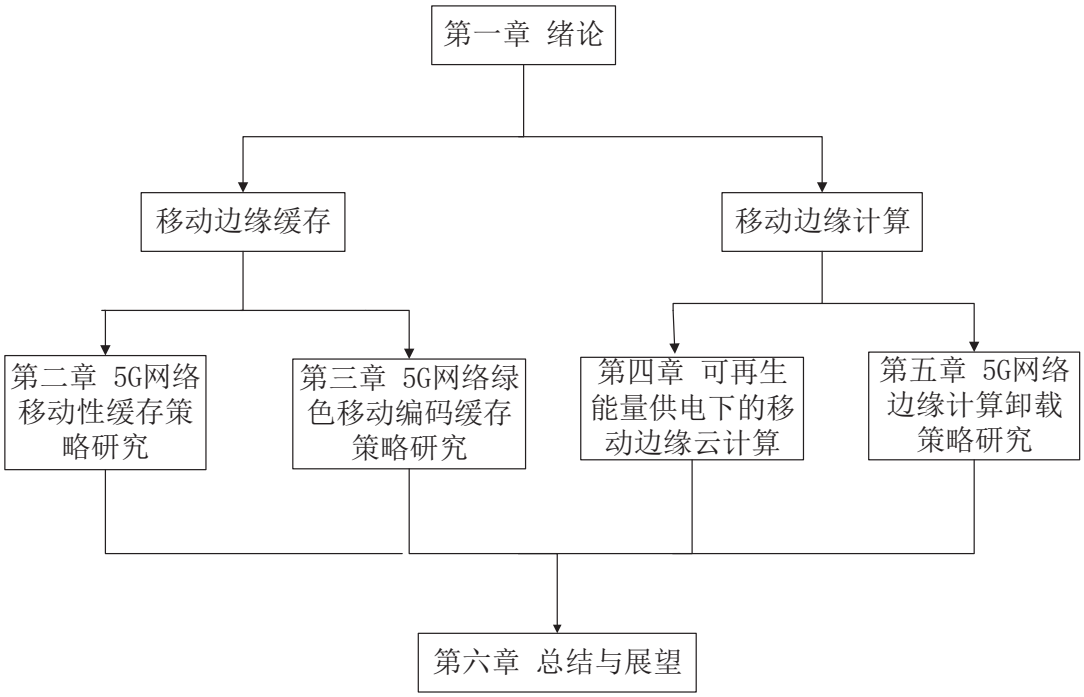


图 1.2 本文章节结构安排

1.4 论文的组织结构.

如图 1.2所示，本文的后续章节组织如下：

第二章研究5G网络移动性缓存策略。考虑到用户的移动性，设计了在small cell和用户设备上的的缓存策略，相比于传统的流行缓存策略和随机缓存策略，本章提出的移动性缓存策略能够使得缓存命中率最大。

第三章研究5G网络绿色移动编码缓存策略。考虑到编码缓存和接触时间间隔的随机性，设计了编码缓存的安置和传输策略，相比于现有的缓存策略，本章提出的缓存策略具有最高的缓存命中率和最低的传输能耗。

第四章研究由可再生能量供电下移动边缘云计算卸载策略。考虑到可再生能量到达的随机性，本章提出了可再生能量供电下的计算任务卸载策略模型，与其他卸载策略相比较，该卸载策略能够保证用户延时的基础上，最大化利用可再生能量。

第五章研究5G网络边缘计算卸载策略。考虑到基于D2D的边缘计算的不可靠性，本章提出一种新的计算任务卸载模式，并给出其计算卸载策略。实验证明了该策略的有效性。

第六章主要总结了全文的研究内容，并展望了未来可能的研究方向。

2 5G网络移动性缓存策略研究

本章研究了用户移动性对small cell和用户设备缓存的影响，建立了移动性缓存策略模型，针对此模型提出基于子模态优化的解决方案，并给出了此策略的仿真实验。

2.1 引言

5G网络的超密集部署small cell虽然能够增加空间的重用，进而提高频谱效率，也一定程度上增大了网络的吞吐量，但是此网络主要受限于回程链路(backhaul link)（即基站到核心网）的容量有限，一种解决方案就是在small cell和用户设备上进行缓存。通过缓存，用户可以直接在本地或基站上获取内容，而不需要通过回程链路，进而减少延时。

然而，目前的缓存研究策略大多是假设在固定网络结构的前提下，没有考虑用户的移动性对用户设备和small cell上缓存策略产生的影响，但是用户移动是缓存网络的重要特征。于是当考虑到用户的移动性时，如何在用户设备和small cell上设计缓存策略，使得缓存的命中率最大是一个重要的问题。为了解决这个问题，本章提出了基于缓存命中率最大的移动性缓存策略的优化问题。经证明，该优化问题是NP难的，并通过子模态优化算法给出了优化问题的近似最优解。综上所述，本章的主要贡献如下：

- 本章在考虑用户移动性的前提下，解决了如何在small cell和用户设备上进行内容的缓存，能够使得缓存的命中率最大的问题。具体来说，本章将此问题建模为0-1非线性规划问题，并证明该问题是一个NP难问题。进一步，将此问题转化为子模态优化问题，并利用贪婪算法给出问题的解。
- 实验结果显示，当用户的移动性比较低时，small cell和用户设备应该缓存流行的文件；当用户的移动性比较高时，small cell和用户设备应该缓存多样性的文件，这样能够使得缓存的命中率最大。

本章节组织如下：第2.2节给出系统的描述，第2.3节给出移动性缓存策略分析与建模，第2.4节提出了基于子模态优化的移动性缓存策略求解，第2.5节给出实验结果与分析，第2.6节对本章进行了小节。

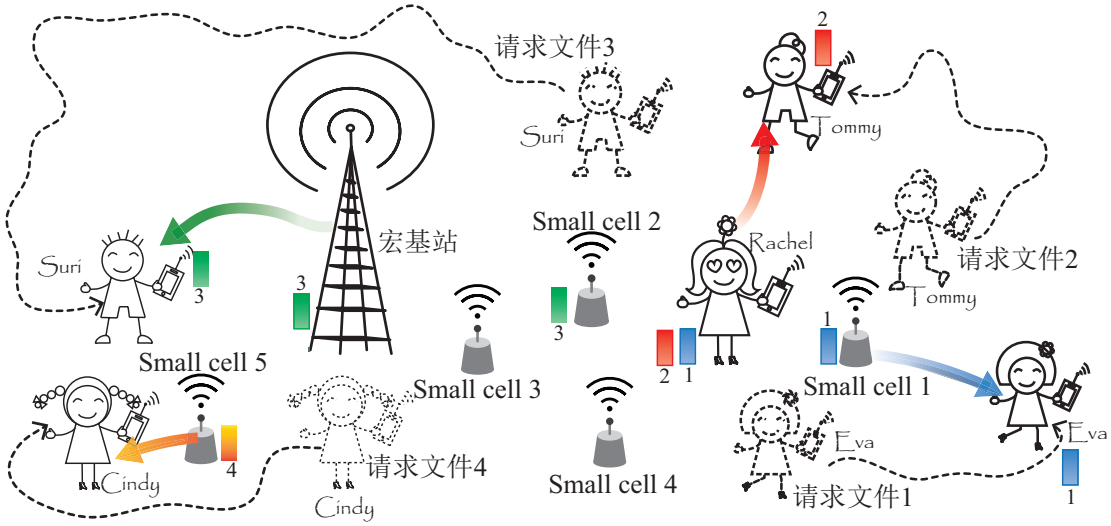


图 2.1 5G网络移动性缓存示例图

2.2 系统描述

本章考虑包括宏基站，small cell基站和移动用户的5G网络。图2.1给出了一个5G网络移动性缓存的例子。假定小区内1个宏基站，5个small cell。先考虑有5个用户，分别是*Rachel*，*Eva*，*Tommy*，*Suri*和*Cindy*，其中*Eva*，*Tommy*，*Suri*和*Cindy*分别请求文件1，文件2，文件3和文件4。对*Eva*来说，由于*Eva*的移动性导致了其不在*Rachel*的D2D 通信范围内，虽然*Rachel* 缓存了文件1，但是*Eva*无法通过D2D方式从*Rachel* 处得到文件1。但*Eva* 在缓存了文件1 的small cell 1 覆盖范围内，于是*Eva*可以通过small cell1获得文件1。对*Tommy*来说，虽然开始时*Tommy* 没有在*Rachel* 的D2D 通信范围内，但随着*Tommy*的移动，*Rachel* 和*Tommy*能够进行D2D 通信，于是*Rachel* 将缓存的文件2传递给了*Tommy*。对于*Suri*来说，由于*Suri*的移动性较高，导致*Suri*没有办法从缓存了文件3的small cell 2上获取文件，最终只能通过宏基站获取文件。对用户*Cindy* 来说，由于*Cindy*的移动性较低，一直在small cell5的覆盖范围内，于是*Cindy* 通过small cell 5获得内容。

通过上面的讨论，可以看出用户移动性对用户设备和small cell上的内容缓存会产生显著的影响。接下来，给出在移动性缓存策略模型。假设一个宏小区内有l个small cell，记为$\mathcal{S} = \{S_1, S_2, \cdots, S_l\}$；假定有$n$ 个用户设备，记为$\mathcal{D} = \{D_1, D_2, \cdots, D_n\}$。接下来给出用户移动和请求模型。

2.2.1 移动模型

在本章中使用的移动模型是成对碰面模型（pairwise connectivity model[63]，这个模型已经被广泛应用于无线网络[28,32,64,65]，与wang等人[28,64]，Lu等人[65]的工作假设一样，即一对节点碰面的过程是独立的泊松过程（independent poisson process）。

small cell网络中用户移动性模型: 用户设备D_i 能够和small cellS_k进行通信的条件是D_i 在S_k 的覆盖区域内，其中覆盖区域的半径为R_S。定义用户设备D_i不在small cell S_k的覆盖区域内的停留时间（接触间隔时间）$T_{i,k}$ 为：$T_{i,k} = \{(t-t_0) : ||\mathcal{L}_k^t - \mathcal{L}_i^t|| \leq R_S, t > t_0\}$，其中$t_0$ 表示最近一次用户设备D_i刚离开small cell S_k的覆盖区域的时刻，变量t 表示用户设备D_i刚进入small cell S_k 的覆盖区域的时刻，$\mathcal{L}_k^t$ 和$\mathcal{L}_i^t$ 表示small cell S_k和用户设备D_i 在t 时刻的位置。根据Poularakis等人[55]的研究表明，用户设备D_i 与small cell S_k的接触时间间隔$T_{i,k}$ 服从参数为$\mu_{i,k}$ 的指数分布，其中$\mu_{i,k}$ 为用户设备D_i和small cellS_k 的接触频率。由于small cell存储能力是有限的，不妨记small cell S_k 的缓存容量为c_k^S。

D2D网络中用户的移动性模型：只有当两个用户设备之间的距离不大于R_{D2D}时，两者才可以进行D2D通信。定义两个用户D_i和D_j 的接触间隔时间$T_{i,j}$为: $T_{i,j} = \{(t-t_0) : ||\mathcal{L}_i^t - \mathcal{L}_j^t|| \leq R_{D2D}, t > t_0\}$，类似地，$t_0$ 表示最近一次用户设备D_i离开D_j的通信范围R_{D2D}的时刻，变量t表示用户设备D_i进入用户设备D_j的时刻。$\mathcal{L}_i^t$ 和$\mathcal{L}_j^t$ 表示用户D_i和D_j 在t 时刻的位置。根据Lu等人[65]的工作表明，用户设备D_i和D_j 的接触时间间隔服从参数为$\lambda_{i,j}$ 的指数分布，其中称$\lambda_{i,j}$ 为用户设备D_i与D_j的接触频率。同样地，考虑到用户设备存储能力的有限性，记用户设备D_i的缓存容量为c_i^D。

2.2.2 用户请求内容模型

考虑到文件库有m个文件，定义这m个文件的集合为：$\mathcal{F} = \{F_1, F_2, \cdots, F_m\}$。其中文件是基于流行度排序的，也就是说，最流行的是文件F_1，最不流行的是文件F_m。假设每个文件的大小相同，且记F_f 的大小为$|F_f|$。每一个用户会随机且独立的从文件库中以概率p_f 请求文件F_f，并假设p_f 服从参数为γ 的Zipf 分布[41]。即：

$$p_f = \frac{f^{-\gamma}}{\sum_{i=1}^{m} i^{-\gamma}}, f = 1, 2, \cdots, m. \tag{2.1}$$

其中γ 表示这些内容流行度的参数。

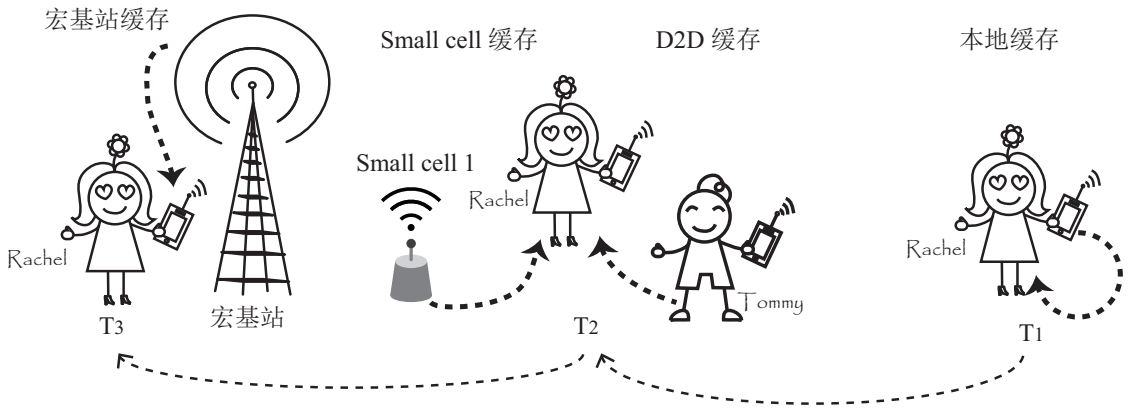


图 2.2 用户获取内容的方式：本地缓存，D2D或small cell缓存，宏基站缓存

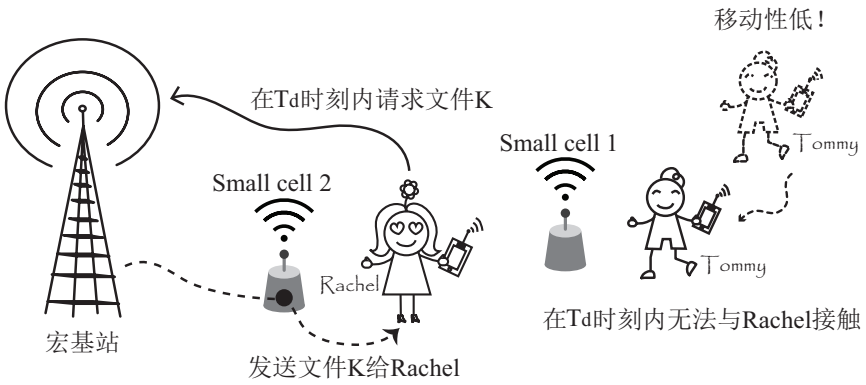
定义T_d为用户请求内容的最大延迟。如图2.2所示，当用户在T_d时间内请求内容时，主要能通过以下三种途径来获取内容：

- 本地缓存：用户设备请求文件时，首先会检查本地是否缓存了请求的文件，要是本地缓存了请求的内容，那么用户便从本地获取请求的内容。
- D2D缓存或small cell缓存：如果用户本地没有缓存其所请求的内容，那么用户设备可以在T_D时间内通过以下两种方式获取到请求的内容：（1）假设在其D2D通信范围内有其他用户设备缓存了请求的内容，那么他们之间可以通过建立D2D 通信，以此获得所请求的内容。(ii) 假设请求内容的用户设备在small cell的覆盖范围内，并且small cell缓存了所请求内容，于是small cell可以将缓存的内容传递给请求内容的用户。
- 宏基站缓存：如果用户在请求内容的最大延迟T_D内，上述方法都不能使用户获取到所请求的内容，那么宏基站会处理他的请求，从而获取到请求的内容。

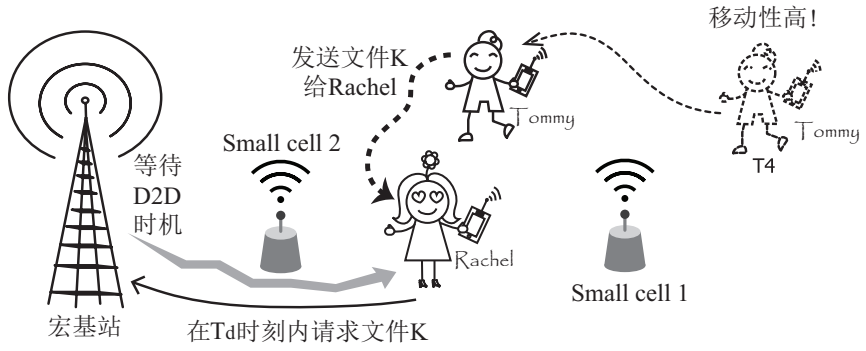
此外，参照Wang等人[59]和Poularakis等人[60]的工作，本章假设文件是非编码缓存的，也就是说能在一次接触时间内均能将用户设备请求的文件传递成功。

2.3 移动性缓存策略分析与建模

这一节提出移动性的缓存策略模型，即考虑到用户的移动性，small cell和用户设备应该缓存哪些内容, 才能使缓存的命中率最大，也就是说请求文件的用户能够从small cell和其他用户设备上获取的概率最大。



(a) 用户移动性较低时



(b) 用户移动性较高时

图 2.3 用户移动性对缓存策略的影响

2.3.1 研究动机

通过一个例子来说明当考虑到用户的移动性时，如何在small cell和用户设备上设计缓存策略，能够使得缓存命中率最大。图2.3体现了用户移动性对samll cell基站缓存和用户设备缓存的影响，具体来说，在图2.3a中，假定用户*Rachel*在T_d时间内请求内容K，此时网络中用户移动性较差，使得*Rachel* 在T_d时间内不能通过D2D通信方式与拥有内容的*Tommy* 碰面。但用户*Rachel* 一直在small cell2中，因此在这种情况下，应该在small cell2 进行内容的缓存（即在移动流量需求较少的时刻，将内容通过宏基站预先缓存在small cell 中），使*Rachel*能够获取到请求的内容。在图2.3b中，同样假定用户*Rachel* 在T_d 时间内请求内容K。此时网络中用户的移动性较强，使得*Rachel*能够在T_d时间内通过D2D 通信的方式与拥有内容的用户*Tommy* 碰面，因此在这种情况下，*Tommy* 应该缓存此内容，并通过D2D 的方式将内容传递给*Rachel*。从这个例子中可以看出，用户移动性的不同，导致缓存策略也不同。

2.3.2 移动性缓存策略分析

考虑到用户的移动性，本节给出small cell和用户设备的综合缓存策略，目标是最大化缓存命中率（cache hit ratio），从而减少宏基站上的流量消耗。

（1）small cell缓存命中率分析： 定义缓存矩阵$\mathbf{X}=(x_{k,f})_{l\times m}$为$l\times m$的$0-1$矩阵。即当内容$F_f$ 缓存在small cellS_k 时，则$x_{k,f}=1$，也就是说，对于任意的k和f，

$$x_{k,f}=\begin{cases}1, & \text{small cell } S_k\text{缓存了文件}F_f\\ 0, & \text{其余情况}\end{cases} \tag{2.2}$$

当用户设备D_i请求内容F_f时，如果用户设备在T_d时间内的某一个时刻进入到缓存了内容F_f 的small cell S_k 的通信范围内，则small cell S_k 可将内容F_f 传递给用户设备D_i。由于用户设备D_i 与small cell S_k 碰面时间间隔服从参数为$\mu_{i,k}$ 的指数分布，那么在T_d 时间内用户设备D_i 至少与一个含有内容F_f 的微基站S_k 碰面的概率为：

$$p_{i,f}^{S}=1-\exp\left(-\sum_{k=1}^{l}x_{k,f}\mu_{i,k}T_d\right). \tag{2.3}$$

因此，如果不考虑用户设备缓存，small cell的缓存命中率为：

$$P^{S}=\frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m}p_f p_{i,f}^{H}. \tag{2.4}$$

（2）用户设备缓存命中率分析 定义缓存矩阵$\mathbf{Y}=(y_{j,f})_{n\times m}$为$n\times m$ 的$0-1$矩阵，即当内容F_k 缓存在用户设备D_j 时，则$y_{j,k}=1$，也就是说，对于任意的j 和f，

$$y_{j,f}=\begin{cases}1, & \text{用户设备}D_j\text{缓存了内容}F_f\\ 0, & \text{其余情况}\end{cases} \tag{2.5}$$

当用户设备D_i请求内容F_f时，如果用户设备在T_d时间内的某一时刻能够与缓存了内容F_f的用户设备D_j碰面的话，那么D_j可以通过D2D 通信的方式将内容F_f传递给D_i。由于用户设备D_i与用户设备D_j 碰面的时间间隔服从参数为$\lambda_{i,j}$ 的指数分布，那么在T_d 时间内用户设备D_i 至少与一个含有内容F_f的用户设备D_j 碰面的概率为：

$$p_{i,f}^{D}=1-\exp\left(-\sum_{j\neq i}y_{j,f}\lambda_{i,j}T_d\right). \tag{2.6}$$

因此，如果不考虑small cell缓存，当用户设备D_i请求内容F_f时，D_i 会首先检查自身是否缓存了内容，如果自身没有缓存的话，则会通过D2D通信的方式从其他用户设备上获取。用户设备D_i通过D2D通信方式获得请求内容F_f的概率为：

$$p_i^f = \begin{cases} p_f, & \text{if } y_{i,f} = 1, \\ p_f p_{i,f}^D, & \text{otherwise}, \end{cases} \tag{2.7}$$

因此，在不考虑small cell缓存时，用户设备的缓存命中率为：

$$P^D = \frac{1}{n} \sum_{i=1}^{n} \sum_{f=1}^{m} p_i^f . \tag{2.8}$$

（3）small cell和用户设备缓存命中率分析： 当用户设备D_i请求内容F_f时，用户设备D_i 在T_d 时间内不通过宏基站获取内容的概率等价于用户设备D_i在T_d 时间内至少与一个含有内容F_f 的small cell S_k或用户设备D_j碰面的概率，即：

$$\begin{aligned} p_{i,f}^{SD} &= 1 - (1 - p_{i,f}^S)(1 - p_{i,f}^D), \\ &= 1 - \exp\left[- \left(\sum_{k=1}^{l} x_{k,f} \mu_{i,k} T_d + \sum_{j \neq i} y_{j,f} \lambda_{i,j} T_d \right) \right]. \end{aligned} \tag{2.9}$$

于是可以得到small cell和用户设备的缓存命中率为：

$$P^{SD} = \frac{1}{n} \sum_{i=1}^{n} \sum_{f=1}^{m} p_f \left\{ 1 - (1 - y_{i,f}) \exp\left[- \left(\mu_{i,k} T_d \sum_{k=1}^{l} x_{k,f} + \lambda_{i,j} T_d \sum_{j \neq i} y_{j,f} \right) \right] \right\}. \tag{2.10}$$

至此得到了small cell和用户设备缓存的命中率，P^{SD}越大，用户通过缓存获得内容的概率越大，宏基站的负载也越小。

2.3.3 移动性缓存策略问题的定义

定义矩阵$\mathbf{Z} = [\mathbf{X}; \mathbf{Y}]$表示small cell和用户设备上的缓存策略矩阵。根据2.3.2节的分析得到了small cell和用户设备的缓存命中率的表达式P^{SD}，本章的目标是最大化缓存命中率P^{SD}，约束条件为基站和移动设备缓存容量的有限性。于是移动性的缓存问题可以定义为下面的优化问题：

$$\underset{\mathbf{Z}}{\text{maximize}} \quad P^{SD} \tag{2.11a}$$

$$\text{subject to} \quad \sum_{f=1}^{m} |F_f| x_{k,f} \leq c_k^S, \quad \forall 1 \leq k \leq l, \tag{2.11b}$$

$$\sum_{f=1}^{m}|F_f|y_{j,f} \le c_j^D, \quad \forall 1 \le j \le n, \tag{2.11c}$$

$$x_{k,f} \in \{0,1\}, \tag{2.11d}$$

$$y_{j,f} \in \{0,1\}. \tag{2.11e}$$

其中目标函数(2.11a)表示最大化缓存命中率，约束条件(2.11b)表示small cellS_k的缓存容量不能超过其最大缓存容量c_k^S，约束条件(2.11c)表示用户设备D_j的缓存容量不能超过其最大缓存容量c_j^D。由于矩阵**Z** 为0-1矩阵，所以这个优化问题是混合整数非线性规划。在不考虑限制条件的情况下，一共有$2^{(l+m)\times n}$ 个选法，根据Lan等人[67]的工作，可以将此问题转化为0-1背包问题，所以该问题为NP难问题。

2.4 基于子模态优化的移动性缓存策略求解

这一节将2.3节的优化问题转化为子模态优化(submodular optimization)问题，并利用贪婪算法去求解。

2.4.1 子模态优化问题

定义集合$Z = \{z_{i,f}|i = 1,\cdots,l+n, f = 1,\cdots,m\} = Z_1 \cup Z_2$，其中

$$Z_1 = \{x_{k,f}|k = 1,\cdots,l, f = 1,\cdots,m\}$$
$$Z_2 = \{y_{j,f}|j = 1,\cdots,n, f = 1,\cdots,m\}$$

当$i = 1,\cdots,l$ 时，$z_{i,f} = x_{k,f}$；当$i = l+1,\cdots,l+n$ 时，$z_{i,f} = y_{j,f}$。定义移动性的缓存策略集合为$A = A_1 \cup A_2$，其中$A_1 \subseteq Z_1$，$A_2 \subseteq Z_2$。如果$x_{k,f} \in A_1$ 当且仅当$x_{k,f} = 1$，$y_{j,f} \in A_2$ 当且仅当$y_{j,f} = 1$。于是原优化问题目标函数可以写为：

$$g(A) = \frac{1}{n}\sum_{f=1}^{m} p_f \left[n - \sum_{y_{i,f}\in Z_2 \setminus A_2} \exp\left(-\left(\sum_{x_{k,f}\in A_1} \mu_{i,k}T_d + \sum_{y_{j,f}\in A_2} \lambda_{i,j}T_d \right) \right) \right] \tag{2.12}$$

定义$Z_1^k = \{x_{k,f}|f = 1,\cdots,m\}$，$Z_2^j = \{y_{j,f}|f = 1,\cdots,m\}$。记$|\cdot|$ 表示集合元素的个数，于是约束条件可以写为：$I = I_1 \cup I_2$，其中

$$I_1 = \{A_1||F_f||A_1 \cap Z_1^k| \le c_k^H, k = 1,\cdots,l\}$$
$$I_2 = \{A_2||F_f||A_2 \cap Z_2^j| \le c_j^D, j = 1,\cdots,n\}$$

因此原优化问题可以转换为如下优化问题：

$$\begin{aligned} \underset{A}{\text{maximize}} \quad & g(A) \\ \text{subject to:} \quad & A \subseteq I. \end{aligned} \tag{2.13}$$

定理 2.1 优化问题(2.13)中，g(A)是单调的子模态函数（monotone submodular function），约束条件(Z, I)为拟阵（matroid）。

证明. （1）证明函数g(A)是单调的子模态函数，根据单调子模态函数的性质，即证明对于$\forall A \subset B \subset Z$和$\forall z_{i,f} \in Z-B$，有$g(A \cup \{z_{i,f}\}) - g(A) \geq g(B \cup \{z_{i,f}\}) - g(B) \geq 0$，也就是证明下面两式均成立。

$$\begin{aligned} g(A \cup \{x_{k,f}\}) - g(A) \geq g(B \cup \{x_{k,f}\}) - g(B) \geq 0 \\ g(A \cup \{y_{j,f}\}) - g(A) \geq g(B \cup \{y_{j,f}\}) - g(B) \geq 0 \end{aligned}$$

首先证明$g(A \cup \{x_{k,f}\}) - g(A) \geq g(B \cup \{x_{k,f}\}) - g(B) \geq 0$。记$E_1 = Z_1 - A_1$，$E_1' = Z_1 - B_1$，$E_2 = Z_2 - A_2$，$E_2' = Z_2 - B_2$，于是可以得到：

$$\begin{aligned} & g(A + x_{k,f}) - g(A) \\ & = \frac{1}{n} \sum_{f=1}^{m} p_f \Bigg\{ \sum_{y_{i,f} \in E_2} \exp\left[-T_d \left(\sum_{x_{k,f} \in A_1} \mu_{i,k} + \sum_{y_{j,f} \in A_2} \lambda_{i,j} \right) \right] - \\ & \sum_{y_{i,f} \in E_2 - x_{k,f}} \exp\left[-T_d \left(\sum_{x_{k',f} \in A_1 + x_{k,f}} \mu_{i,k'} + \sum_{y_{j,f} \in A_2} \lambda_{i,j} \right) \right] \Bigg\} \\ & = \frac{1}{n} \sum_{f=1}^{m} p_f \left[\sum_{y_{i,f} \in E_2 - x_{k,f}} e^{-\left(\sum\limits_{x_{k',f} \in A_1} \mu_{i,k'} T_d + \sum\limits_{y_{j,f} \in A_2} \lambda_{i,j} T_d \right)} \left(1 - e^{-\mu_{ik} T_d}\right) \right]. \end{aligned}$$

由于$A \subset B$，且对任意的$x \in R$，指数函数$e^x > 0$，那么

$$\begin{aligned} & g(A + x_{k,f}) - g(A) \\ & \geq \frac{1}{n} \sum_{f=1}^{m} p_f \left[\sum_{y_{i,f} \in E_2' - x_{k,f}} e^{-\left(\sum\limits_{x_{k',f} \in B_1} \mu_{i,k'} T_d + \sum\limits_{y_{j,f} \in B_2} \lambda_{i,j} T_d \right)} \left(1 - e^{-\mu_{ik} T_d}\right) \right] \\ & = g(B + x_{k,f}) - g(B) \geq 0. \end{aligned}$$

因此证明了$g(A \cup \{x_{k,f}\}) - g(A) \geq g(B \cup \{x_{k,f}\}) - g(B) \geq 0$。

其次证明$g(A\cup\{y_{j,f}\})-g(A)\geq g(B\cup\{y_{j,f}\})-g(B)\geq 0$，和上面证明类似，可以得到：

$$
\begin{aligned}
&g(A+y_{j,f})-g(A)\\
&=\frac{1}{n}\sum_{f=1}^{m}p_f\Bigg\{\sum_{y_{i,f}\in E_2}\exp\left[-T_d\left(\sum_{x_{k,f}\in A_1}\mu_{i,k}+\sum_{y_{j,f}\in A_2}\lambda_{i,j}\right)\right]-\\
&\sum_{y_{i,f}\in E_2-y_{j,f}}\exp\left[-T_d\left(\sum_{x_{k,f}\in A_1}\mu_{i,k}+\sum_{y_{j',f}\in A_2+y_{j,f}}\lambda_{i,j'}\right)\right]\Bigg\}\\
&=\frac{1}{n}\sum_{f=1}^{m}p_f\Bigg\{\sum_{y_{i,f}\in E_2-y_{j,f}}\exp\left[-T_d\left(\sum_{x_{k,f}\in A_1}\mu_{i,k}+\sum_{y_{j',f}\in A_2}\lambda_{i,j'}\right)\right]\left[1-\exp\left(-\lambda_{ij}T_d\right)\right]\\
&+\exp\left[-T_d\left(\sum_{x_{k,f}\in A_1}\mu_{j,k}+\sum_{y_{j',f}\in A_2}\lambda_{j,j'}\right)\right]\Bigg\}
\end{aligned}
$$

同理由于$A\subset B$，对于任意的$x\in R$，指数函数$e^x>0$，那么，

$$
\begin{aligned}
&g(A+y_{j,f})-g(A)\\
&\geq\frac{1}{n}\sum_{f=1}^{m}p_f\Bigg\{\sum_{y_{i,f}\in E_2'-y_{j,f}}\exp\left[-T_d\left(\sum_{x_{k,f}\in B_1}\mu_{i,k}+\sum_{y_{j',f}\in B_2}\lambda_{i,j'}\right)\right](1-e^{-\lambda_{ij}T_d})\\
&+\exp\left[-T_d\left(\sum_{x_{k,f}\in B_1}\mu_{j,k}+\sum_{y_{j',f}\in B_2}\lambda_{j,j'}\right)\right]\Bigg\}\\
&=g(B+y_{j,f})-g(B)\geq 0.
\end{aligned}
$$

这就证明了$g(A\cup\{y_{j,f}\})-g(A)\geq g(B\cup\{y_{j,f}\})-g(B)\geq 0$，所以可以得到g(A)是单调的子模态函数。

（2）证明(Z,I)为拟阵，即证(Z,I_1)和(Z,I_2)均为拟阵。首先对于(Z,I_1)，满足

- $\emptyset\in I_1$;
- 如果$B\subseteq I$ 和$A\subseteq B$, 则$A\subseteq I$;
- 如果$A,B\in I_1$ 且$|A|<|B|$, 存在元素$j\in B-A$, 使得$A\cup j\in I$.

所以(Z,I_1)为拟阵，同理可证(Z,I_2)也为拟阵，于是可以得到约束条件(Z,I)为拟阵。□

2.4.2 移动性缓存策略算法

对于目标函数是单调递增的子模态函数，约束条件为拟阵的优化问题，根据Calinescu等人[88]的研究表明贪婪算法是一个有效的解决算法，能够非常接近最优解，那么结合定理2.1可知，本章的优化问题可以通过贪婪算法求解。在本章中，此算法称为移动性的缓存策略算法，具体的算法如下所示，其思想就是先设置一个空的缓存集合A，在每一次迭代的过程中，加入一个文件使得目标函数最大化，直到达到small cell和用户设备的最大缓存容量。

算法 2.1: 移动性缓存策略算法

输入: 所有的$z_{i,f}$集合，Z;

Z剩余集合，Z_r;

small cell和用户设备的总存储容量，C;

输出: 在用户设备和small cell上的缓存策略，A

1: $A \leftarrow \emptyset, Z_r \leftarrow Z$;

2: Repeat ;

3: $z_{i^*,f^*} = \text{argmax}_{z_{i,f} \in Z_r}[g(Z_r + z_{i,f}) - g(Z_r)]$;

4: $A \leftarrow A + z_{i^*,f^*}$;

5: $Z_r \leftarrow Z_r - z_{i^*,f^*}$;

6: 如果$|A \cap Z_{i^*}| = C$，则$Z_r \leftarrow Z_r \backslash Z_{i^*}$;

7: end if;

8: 直到$|A| > (\sum_{k=1}^{l} c_k^H + \sum_{i=1}^{l} c_i^D)$;

从步骤3中可以得出在small cell或用户设备上缓存文件F_{f^*}能够最大化缓存命中率，所以在步骤4 中将z_{i^*,f^*}加入到最优的缓存策略中。步骤6表示当达到small cell或用户设备i的最大缓存容量C时，不能够再缓存文件。从步骤8中可以得到当$|A| > (\sum_{k=1}^{l} c_k^H + \sum_{i=1}^{l} c_i^D)$时，停止迭代。在算法 2.1中，当所有的small cell和用户设备达到其最大的缓存容量，将会有$\sum_{k=1}^{l} c_k^H + \sum_{i=1}^{l} c_i^D$ 次迭代。对于每一次迭代，最多有$(l+n)m$元素没有加入到缓存集合A中。对于每一次计算，其时间复杂度为$\mathcal{O}(n)$, 所以此算法的总体时间复杂度为$\mathcal{O}\left((l+n)mn(\sum_{k=1}^{l} c_k^H + \sum_{i=1}^{l} c_i^D)\right)$。

表 2.1 移动性缓存策略参数的设置

参数	取值
small cell的数目，l	5
用户设备的数目，n	60
用户设备D_i与用户设备D_j接触时间的指数分布参数，$\lambda_{i,j}$	$\Gamma(4.43, 1/1088)$
用户设备D_i与small cellS_k接触时间的指数分布参数，$\mu_{i,k}$	$\Gamma(10, 1/100)$
截至时间T_d	600 s
文件库的数目，m	40
Zipf分布参数，γ	0.8

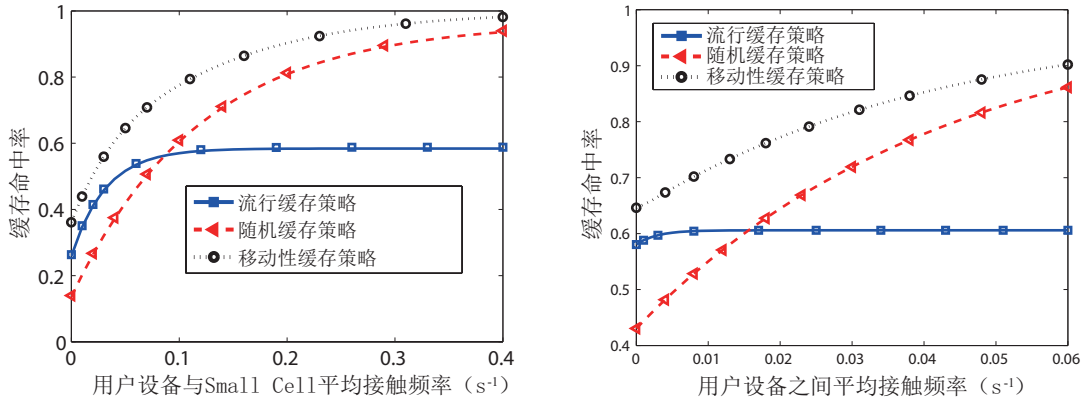
2.5 实验结果与分析

这一节对本章提出的移动性缓存策略进行评估。为了简单起见，假定所有用户设备具有相同的存储能力c^D，类似地，假定所有的small cell具有相同的存储能力c^B。在本章中，参照Poularakis等人[55]对超密蜂窝网络的设置，考虑一个较小的区域，即$250m \times 250m$的区域内含有5个small cell，60个用户。根据Passarella等人[89]和Wang等人[64]对用户移动性的设置，设置用户D_i 与用户D_j 的接触率$\lambda_{i,j}$ 服从Gamma 分布，用户D_i 与small cell S_k 的接触率也服从Gamma 分布。考虑到接触时间的有限性，在有限的接触时间内只能传输相对较少的文件，于是基于Liu等人[90]的工作，假设文件的大小为10 MB，文件的个数为10^4 个。基于Poularakis等人[55] 对缓存容量大小的设置，本章假设small cell最多能够缓存文件库的10%，用户设备最多能缓存文件库的5%，表 2.1给出了具体参数设置

2.5.1 对比方案

本节将本章提出的移动性缓存策略与流行缓存策略[43]和随机缓存策略[47]进行比较。流行缓存策略和随机缓存策略具体设置如下：

- 流行缓存策略：在small cell和用户设备上的具体的流行缓存策略如下。small cell上的缓存策略：每个small cell S_k上缓存c_k^S个最流行的文件。用户设备上的缓存策略：每一个用户用户设备D_j 上缓存c_j^D 个最流行的文件。
- 随机缓存策略：在small cell和用户设备上的具体的随机缓存策略如下。small cell上的缓存策略，每个small cell S_k上随机缓存c_k^S个文件。用户设备上缓存策



(a) 用户设备与small cell平均接触频率($\overline{\mu}_{i,k}$)对缓存命中率的影响

(b) 用户设备之间平均接触频率($\overline{\lambda}_{i,j}$)对缓存命中率的影响

图 2.4 用户移动性对缓存命中率的影响

略，每一个用户设备D_j 随机缓存c_j^D 个文件。

可以看出流行缓存策略注重的是文件流行度，而随机缓存注重的是文件的多样性。

2.5.2 缓存命中率分析

本节利用缓存命中率来评估缓存策略。具体从用户移动性，用户设备，small cell和文件四个方面来给出缓存策略的评估。

（1）用户移动性对缓存命中率的影响。

图2.4讨论了用户移动性对三种不同缓存策略命中率的影响，也就是$\mu_{i,k}$ 与$\lambda_{i,j}$ 对缓存命中率的影响。在本实验中，为了表示方便，图中横坐标分别取$\overline{\mu}_{i,k}$ 和$\overline{\lambda}_{i,j}$, 对应Gamma分布的期望，即$\overline{\mu}_{i,k}$表示单位时间内用户设备D_i 与small cellS_k平均碰面的次数，$\overline{\lambda}_{i,j}$表示单位时间内用户设备D_i与D_j 的平均碰面次数。从图2.4a中可以看出：移动性缓存策略要优于随机缓存和流行缓存策略。这是由于流行缓存策略中只缓存了最流行的文件，而随机缓存策略是随机地缓存文件，以上两种方案都没有考虑到用户的移动性，而移动性缓存策略考虑到了用户的移动性，能够利用用户的移动性来增加缓存命中率。

从图2.4a和图2.4b中还可以得出：当用户的移动性较低时，移动性缓存策略和流行性缓存策略的缓存命中率相差不大，于是可以得到small cell和用户设备应该缓存较为流行的内容，这是因为当用户的移动性较低时，用户设备不仅在不同small cell之间的切换较少，而用户和用户之间碰面的机会比较少，但考虑到用户对流行文

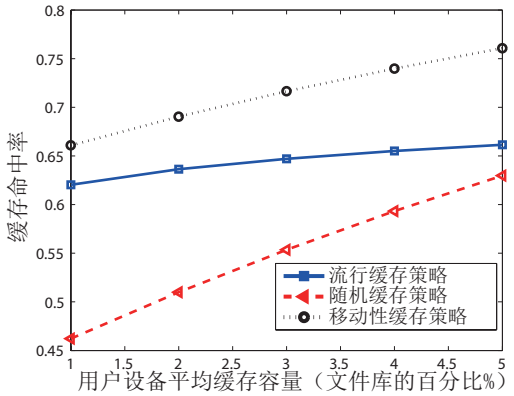


图 2.5 用户设备缓存容量大小对缓存命中率的影响

件的需求量比较大，因此small cell和用户设备应该缓存流行的文件，从而能提高缓存的命中率。当用户的移动性比较高时，可以看到移动性缓存策略和随机缓存策略的差距较小，这是因为此时网络比较活跃，用户在small cell之间切换和用户设备之间碰面的概率比较大，从而获取到需求量较大的流行文件的概率变大，此时可以通过考虑文件多样性进行缓存，来提高总体的缓存命中率。

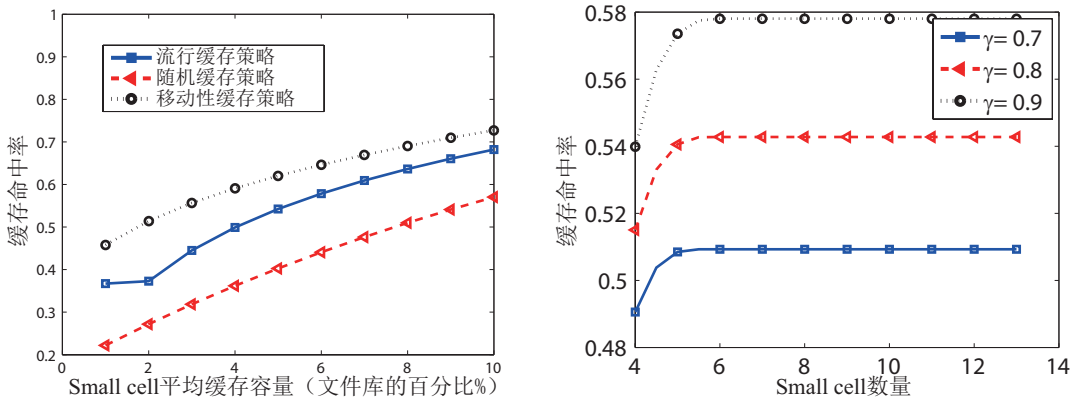
（2）用户设备缓存容量对缓存命中率的影响。

图2.5讨论了用户设备的缓存容量与缓存命中率的关系。横坐标是移动设备的缓存容量占文件库的比重，比如横坐标中的1 表示用户设备最多能够缓存文件库中1%的文件。图2.5给出，当所有用户设备的平均缓存容量增加时，内容缓存的命中率增加。这是因为当用户设备缓存容量变大时，能够存储更多的文件，即当用户设备存储能力变大时，能够提高缓存命中率。另外，从图2.5 中还可以得出，当用户设备的缓存容量变大时，对随机缓存策略命中率的影响要大于对流行缓存策略命中率的影响，这是因为当用户移动设备的缓存容量变大时，随机缓存策略缓存的文件多样性要多于流行缓存策略，从而提高了缓存的命中率。

（3）small cell对缓存命中率的影响。

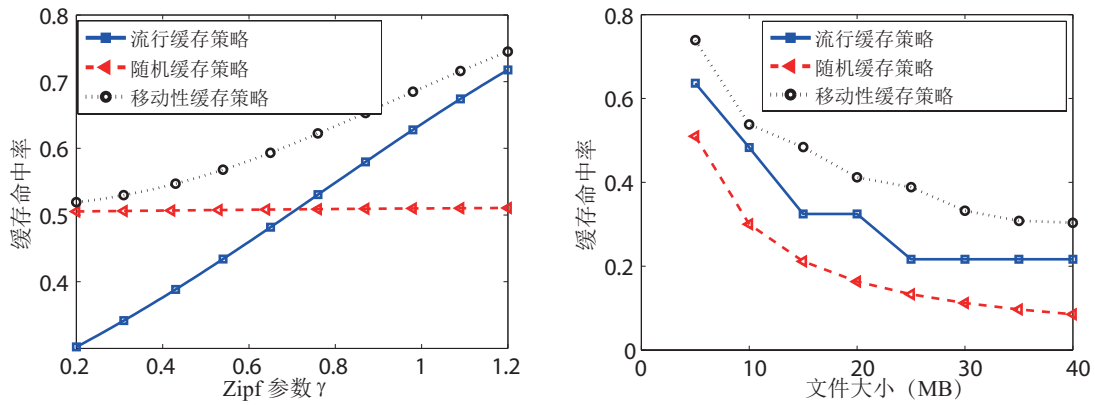
图2.6a讨论了small cell的缓存容量对缓存命中率的影响。图2.6a的横坐标表示small cell 缓存容量占文件库的比例。从图2.6a得出，small cell平均缓存容量变大时，内容缓存命中率变大。这是因为当small cell 的缓存容量变大时，可以缓存更多的内容。即当small cell的存储能力变大，能够提高缓存命中率。

图2.6b给出了小区small cell数量和缓存命中率的关系。从图中可以得出，当一个小区内的基站数目小于6 时，随着基站数目的增加缓存的命中率不断增加，当小区内



(a) small cell缓存容量对缓存命中率的影响　(b) small cell数量对缓存命中率的影响

图 2.6 small cell对缓存命中率的影响



(a) Zipf分布参数(γ)对缓存命中率的影响　(b) 文件大小(F_f)对缓存命中率的影响。

图 2.7 文件流行性分布参数与文件大小对缓存命中率的影响

的small cell数目多于6时，随着small cell数目的增加而变化不大。此外，从图2.6b我们还可以得出，当$\gamma = 0.7$时，small cell的个数达到5时就变化不大，而$\gamma = 0.9$时，small cell的个数达到6时才变化不大。于是可以得到，随着流行度γ的增大，small cell越应该进行分布式的部署（即每个small cell的存储量不大，但总的small cell个数比较多）。这是因为随着γ越高，用户会更多地请求流行的内容，采用分布式的部署，每个small cell都缓存较为流行的内容，能够满足大部分用户对内容的请求，进而提高内容的缓存命中率，

（4）文件对缓存命中率的影响。

如图2.7所示，讨论了文件大小和流行性分布对缓存命中率的影响。在图2.7a 中，讨论了Zipf分布参数γ 对缓存命中率的影响。从图2.7a中可以得出，随着流行度γ 增加，移动性缓存策略的命中率逐渐增加。此外，从图中还可以得出，随着γ的增加，流行缓存策略的命中率逐渐增加，而随机缓存的命中率变化不大，即γ对流行缓存策略命中率的影响比较大，而对随机缓存策略命中率的影响比较小。这是因为随着γ的增加，用户对流行文件的请求变多，流行缓存策略更能满足用户的请求，所以流行缓存策略的命中率随着γ 的增加而增加。而随机缓存策略一直随机缓存文件，所以γ 对其影响较小。

图2.7b讨论了文件库中每个文件大小对缓存命中率的影响。给定文件库的大小，于是可知，每个文件越大，则文件个数越少。图中的坐标仍然表示的为每个文件的大小。图2.7b给出，每个文件越大，系统的缓存命中率越小。这是因为当文件比较大时，由于small cell和用户设备的存储容量是一定的，那么small cell和用户设备能够缓存的文件数目变少，所以降低了缓存的命中率。

2.6 本章小节

本章首先分析了用户移动性对small cell和用户设备上缓存策略部署的影响，其次建立了最大化缓存命中率的优化问题，经证明，此问题为NP难问题，于是将此问题转化子模态优化问题进行了求解，基于此，给出了移动性缓存策略。基于仿真实验得出，对比其他现有的方案，本章提出的缓存策略最优。

3 5G网络绿色移动编码缓存策略研究

本章首先分析了由用户移动性导致的接触时间的随机性对small cell和用户设备上缓存策略的影响，其次，基于编码缓存建立了缓存命中率最大化的安置策略模型和能量消耗最小化的传输策略模型。通过两个模型的求解给出了5G网络绿色移动编码缓存策略，最后给出了实验验证。

3.1 引言

在5G网络中，用户的移动性导致了请求内容的用户与small cell或其他用户的接触时间具有动态的特性，在接触时间内可能只完成了部分内容的传输，导致文件传输失败。然而，目前所有的研究均假设请求内容的用户与缓存内容的small cell 或其他用户在接触时间内能完成整个文件[60,67] 或编码后的固定文件[28,55] 的传输。而Lu等人[65]研究表明，每次的传输量和接触时间是相关的。考虑到用户接触时间具有随机性，若假设每次都能够传输固定的值显然是与实际不符的。因此，考虑到接触时间的随机性，如何在small cell 和用户设备上进行内容安置，使得缓存的命中率最大，与此同时，为保证5G网络的绿色性，如何减少缓存传输过程中的能耗是一个挑战性问题。

为了解决上述问题，本章采用编码缓存，并且假设每次的传输量和接触时间相关，提出了绿色移动编码缓存策略。此策略包括编码缓存的安置和传输策略，其中编码缓存的安置策略保证了缓存命中率最大，编码缓存的传输策略能够保证内容的传输能耗最小。实验证明，相比于其他卸载策略，本章提出的策略在缓存命中率和能耗方面最优。

综上所述，本章的主要贡献如下：

- 针对用户与small cell之间或用户与用户之间接触时间的随机性，本章提出了5G网络绿色编码缓存策略。具体来说，本章基于编码缓存建立了缓存命中率最大化的安置策略模型和能量消耗最小化的传输策略模型，并通过子模态优化给出了缓存安置策略算法，进一步得到了small cell基站和移动设备的最优发射功率，从而保证了最少的传输能量消耗。
- 实验结果显示，本章提出的绿色移动编码缓存策略能够提高内容的缓存命中率，减少缓存内容传输的能量消耗。此外，实验结果显示，当用户移动性较

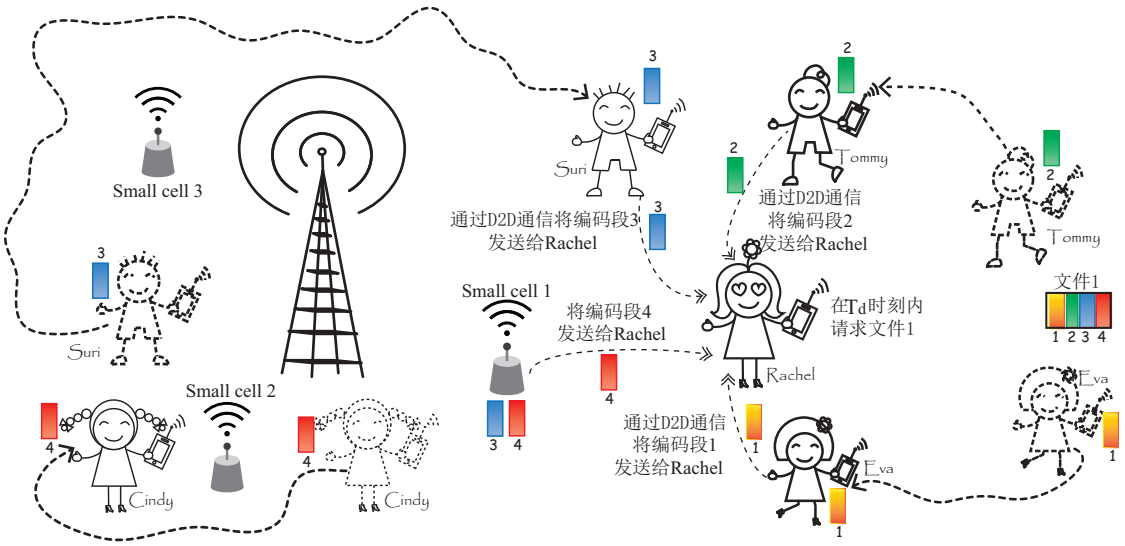


图 3.1 绿色移动编码缓存示例图

高时，缓存内容传输消耗的能量较少；当用户移动性较低时，缓存内容传输消耗的能量较多。

本章节组织如下。第3.2节给出了系统描述。第3.3节提出了绿色移动编码缓存策略模型。第3.4节给出了绿色移动编码缓存策略的求解。第3.5节给出了缓存策略的仿真实验，并分析了实验结果。第3.6 节对本章进行了总结。

3.2 问题的提出

本章考虑了基于5G网络中的绿色移动编码缓存中内容的安置与传输。图 3.1中给出了一个具体的场景。假设有5 个手机用户，分别为*Rachel*，*Eva*，*Tommy*，*Suri*和*Cindy*，其中*Rachel* 请求的文件1包含四个编码段s_1, s_2, s_3, s_4。考虑到用户存储能力的有限性，假设*Eva*, *Tommy*，*Suri* 和*Cindy* 分别存储了s_1，s_2，s_3 和s_4。考虑到用户接触时间是随机的，对于用户*Rachel*来说，她可以分别从*Eva*，*Tommy*，*Suri*获得编码段s_1, s_2, s_3，并且能从small cell 2中获取编码段s_4，从而减少宏基站的负载。当用户与small cell或其他用户在接触时间内，还需考虑small cell和其他用户应该以多大的发射功率传输缓存文件，能够使得其消耗的功率最少。通过上述讨论可知，接触时间对small cell 和用户设备的内容安置与传输有着显著的影响。系统的场景与第二章类似，假设小区内有l个small cell，记为$\mathcal{S}=\{S_1, S_2, \cdots, S_l\}$，$n$ 个用户设备，记为$\mathcal{D}=\{D_1, D_2, \cdots, D_n\}$。宏基站可以与small cel 以及用户设备进行通信。每个用户

表 3.1 本章常用符号

符号	含义
$\mathcal{S}$	small cell的集合
$\mathcal{D}$	用户设备的集合
$\mathcal{F}$	文件库集合
C_k^S	small cell S_k的缓存大小
C_i^D	用户设备D_i的缓存大小
s_f	编码后文件F_f的个数
p_f	用户请求文件F_f的概率
$\lambda_{i,k}^B$	用户D_i与微基站S_k接触时间的指数分布参数
$\lambda_{i,j}^D$	用户设备D_i与用户D_j接触时间的指数分布参数
$\mathcal{A}_{i,k}$	在一次接触时间间隔内，small cell S_k和用户设备D_i的最大数据传输量
$\mathcal{B}_{i,j}$	在一次接触时间间隔内，用户设备D_i与用户设备D_j的最大数据传输量
P_T^D	用户设备的发射功率
P_T^B	small cell的发射功率
W_D	D2D的通信带宽
W_B	small cell的下行传输链路带宽

请求的内容都是相互独立的。表 3.1 给出了本章使用的主要符号。

3.2.1 内容请求模型

假设文件库有m个文件，定义这m个文件的集合为：$\mathcal{F} = \{F_1, F_2, \cdots, F_m\}$。本章采用编码缓存的方案，即文件$F_f$可通过无速率喷泉（rateless fountain）编码[91]为s_f个编码段，并且假设文件F_f可以通过这s_f个编码段恢复[64,92]。定义文件F_f 的大小为$|F_f|$，假设文件的大小相同，则每个编码段的大小为$g_f = \frac{|F_f|}{s_f}$。在本章中，假设每个编码段的大小是相同的。

3.2.2 移动模型

本章沿用第二章的移动模型，假设用户设备D_i与small cell S_k的接触时间服从参数为$\lambda_{i,k}^B$ 的指数分布，用户设备D_i与D_j的接触时间服从参数为$\lambda_{i,j}^D$的指数分布。Wang[64] 等人通过实验得出，用户的移动速度和用户的接触比例呈正的线性相关，

即：用户的移动性越高，用户间的接触频率越高。所以用户移动速度可以近似看成用户的接触频率。换句话说，当用户的移动速度很快时，用户之间的接触频率也会相应的增加，从而导致了用户之间每次接触时间相应的减少，这就导致了在接触时间内数据传输量也相应地减少。因此假设每次接触能够传输整个缓存文件[60,67] 或固定的编码文件[28,55] 与实际不符。由于接触时间是服从指数分布的，因此可以假设在接触时间内内容的传输量也是服从指数分布。定义随机变量$\mathcal{A}_{i,k}$为第ω 次接触时间内small cell 基站S_k 传输给用户设备D_i 的最大数据量。定义随机变量$\mathcal{B}_{i,j}$ 为第ω次接触时间内用户设备D_j 传输给用户设备D_i 的最大数据量。

3.2.3 能量消耗模型

本小节首先给出用户设备D_j将编码缓存段传输给D_i时消耗的能量。为了简单起见，本章不考虑D2D 通信的干扰。定义P_T^D为用户设备的传输功率。基于chen等人[93]的工作，可以得到D2D 通信时数据的平均传输速率为：

$$R_D = \mathbb{E}\{W_D \log_2\left(1+\frac{P_T^D h_D^2 r_D^{-\alpha}}{\sigma_D^2}\right)\} \approx W_D \log_2\left(1+\frac{P_T^D r_D^{-\alpha}}{\sigma_D^2}\right). \tag{3.1}$$

其中W_D表示用户设备D_i到用户设备D_j的传输带宽，h表示信道增益且服从均值为0的高斯分布，r_D 表示D_i与D_j之间的距离，σ_D^2 表示高斯白噪声，α为路径损失指数。根据Liu等人[90]的工作，D_j消耗的功率为$\beta_D P_T^D + P_C^D + P_H^D$，其中$P_C^D$表示$D_j$的电路消耗的功率，$\beta_D$为功率放大因子的倒数，$P_H^D$ 表示缓存的硬件功率。本章不考虑缓存文件消耗的硬件功率。于是当用户设备D_j将文件$\mathcal{B}_{i,j}$ 传输给用户设备D_i 时，用户设备D_j对应的能量消耗为：

$$E_D = \frac{\mathcal{B}_{i,j}}{R_D}(\beta_D P_T^D + P_C^D), \tag{3.2}$$

其次给出small cellS_k将编码缓存段传输给请求用户D_i消耗small cellS_k 的能量。定义P_T^B为small cell的传输功率。同理可得，基站的下传速率为：

$$R_B = \mathbb{E}\{W_B \log_2\left(1+\frac{P_T^B h_B^2 r_B^{-\beta}}{\sigma_B^2}\right)\} \approx W_B \log_2\left(1+\frac{P_T^B r_B^{-\beta}}{\sigma_B^2}\right), \tag{3.3}$$

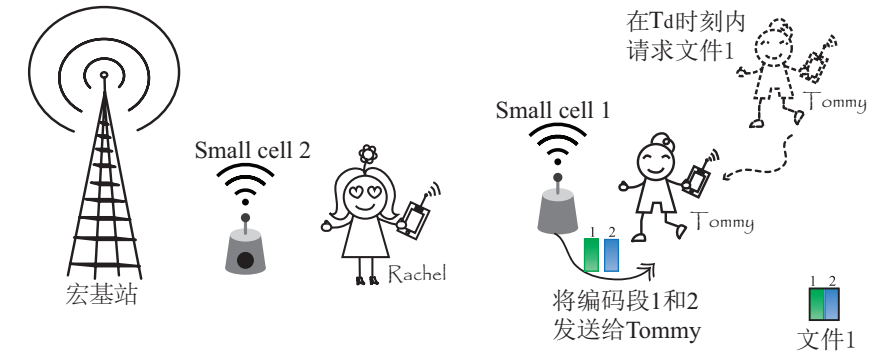
其中W_B表示small cell到用户设备的传输带宽，r_B 是用户设备D_i到small cellS_k的距离，h_B 表示信道的增益，其服从均值为0的高斯分布，β 为路径损失指数，σ_B^2为高斯白噪声。和前面类似，本章不考虑small cell缓存文件带来的能量消耗。基于Liu等人的工作[90] 可以得到当small cellS_k 将$\mathcal{A}_{i,k}$ 传输给用户设备D_i 时，small cell的能量

消耗为：

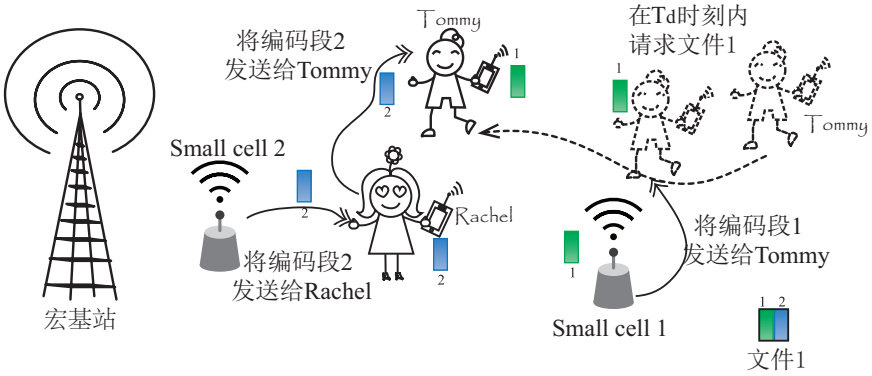
$$E_B = \frac{\mathcal{A}_{i,k}}{R_B}(\beta_B P_T^B + P_C^B), \tag{3.4}$$

其中P_C^B电路功率，β_B是功率放大因子的倒数。

3.3 绿色移动编码缓存策略模型



(a) 接触时间较长



(b) 接触时间较短

图 3.2 接触时间对缓存策略的影响。

本节介绍了绿色移动编码缓存策略模型，首先概述了研究动机，其次提出了缓存内容的安置模型，最后提出了缓存内容的传递模型。

3.3.1 研究动机

图3.2a描述了如何根据用户与small cell基站的接触时间来设计缓存内容的安置和传输策略。假定用户*Tommy* 在T_d 时间内请求文件1，文件1 包含两个编码段。

图3.2a描绘了用户移动性较低，接触时间较长的场景。此时由于用户*Tommy*移动性较低，*Tommy* 与基站和其他用户的接触次数较少，但每次接触时间较长，从图中可以看出，用户*Tommy* 可以从small cell 基站1 上获得需求的整个文件，因此此时的缓存内容安置策略为：将编码段安置在small cell1上。并且在接触时间间隔内，small cell1以最节省能耗的方式将缓存的内容传输给*Tommy*。图3.2b 描绘了用户移动性较高，接触时间较短时的场景。由于此时用户*Tommy* 的移动性较高，用户*Tommy*在与small cell1接触时只完成了一个编码段的传输，在T_d 时间内*Tommy* 与*Rachel* 碰面了，*Rachel*将缓存的编码段传递给*Tommy*，因此此时的缓存内容安置策略为：在small cell 基站1 和*Rachel*处分别缓存编码段。在传输阶段，Small cell1和用户*Rachel* 在接触时间内，分布以能耗最小的方式将缓存内容传输给*Tommy*。因此缓存的策略和接触时间等密切相关，下面分别给出编码缓存的安置和传输策略模型。

3.3.2 编码缓存的安置策略模型

定义矩阵$\mathbf{X}_{l\times m}$为small cell上编码段的安置方案，其中$x_{k,f} \in \mathbf{X}$为small cellS_k上缓存的编码段的个数。定义矩阵$\mathbf{Y}_{n\times m}$ 为用户设备上编码段的安置方案，其中$y_{j,f} \in \mathbf{y}$ 为用户设备D_j 缓存的编码段个数。定义矩阵$\mathbf{Z} = [\mathbf{X}; \mathbf{Y}]$ 为需要求解的编码段安置策略矩阵。定义$\mathcal{U}_i^f(\mathbf{Z})$ 为用户D_i在T_d时间内从small cell 和用户设备上获得的编码段的总量。因此最大化缓存命中率的安置策略优化问题如下：

$$\underset{\mathbf{Z}}{\text{maximize}} \quad \frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m} p_f Pr(\mathcal{U}_i^f(\mathbf{Z})) \tag{3.5a}$$

$$\text{subject to} \quad \sum_{f=1}^{m} x_{k,f} g_f \le C_k^S, \quad \forall k \in \{1,\cdots,l\}, \tag{3.5b}$$

$$\sum_{f=1}^{m} y_{j,f} g_f \le C_j^D, \quad \forall j \in \{1,\cdots,n\}, \tag{3.5c}$$

$$x_{k,f} \in \{0,1,\cdots,s_f\}, \forall k \in \{1,\cdots,l\}, \quad \forall f \in \{1,\cdots,m\} \tag{3.5d}$$

$$y_{j,f} \in \{0,1,\cdots,s_f\}, \forall j \in \{1,\cdots,n\}, \quad \forall f \in \{1,\cdots,m\}. \tag{3.5e}$$

其中目标函数(3.5a)表示如何在small cell和用户设备上进行缓存，使命中率最高。约束条件(3.5b) 表示在small cell上缓存的编码段不能超过其最大缓存量。约束条件(3.5c)表示在用户设备上缓存的编码段不能超过其最大缓存量。约束条件(3.5d)和(3.5e)分别表示在small cell 和用户设备上的缓存的编码段个数必须为整数。

接下来给出$Pr(\mathcal{U}_i^f(\mathbf{Z}))$的求解。根据第二章对用户移动性模型的讨论，可以得到用户设备与small cell 基站，用户设备之间的接触次数服从泊松过程。假设在T_d时间内，用户设备D_i和small cellS_k 接触的次数为$M_{i,k}$，用户设备D_i 与用户设备D_j接触的次数为$N_{i,j}$，于是$M_{i,k}$ 和$N_{i,j}$是服从泊松分布的随机变量。因此可以得到在T_d时间内，用户设备D_i 获取的文件F_f编码段的总量为：

$$V_{i,j,k}^f = \sum_{\omega=1}^{M_{i,k}} \mathcal{A}_{i,k}^{\omega} + \sum_{\omega=1}^{N_{i,j}} \mathcal{B}_{i,j}^{\omega} \tag{3.6}$$

其中$\mathcal{A}_{i,k}^{\omega}$和$\mathcal{B}_{i,j}^{\omega}$均服从指数分布。定义变量$U_{i,j,k}^f$为用户设备$i$ 从small cellS_k 和用户设备D_j上获得的缓存编码段个数，那么：

$$U_{i,j,k}^f = \min(\lfloor \frac{V_{i,j,k}^f}{g_f} \rfloor, x_{k,f} + y_{j,f}) \tag{3.7}$$

因此，可以得到$Pr(\mathcal{U}_i^f(\mathbf{Z})) = Pr(\sum_{k=1}^{l}\sum_{j=1}^{n} U_{i,j,k}^f \geq s_f) = 1 - Pr(\sum_{k=1}^{l}\sum_{j=1}^{n} U_{i,j,k}^f \leq s_f)$，下面我们给出$Pr(\sum_{k=1}^{l}\sum_{j=1}^{n} U_{i,j,k}^f \leq s_f)$的求解。

定义$Pr(l+n, s_f) = Pr(\sum_{k=1}^{l}\sum_{j=1}^{n} U_{i,j,k}^f \leq s_f)$。根据全概率公式可以得到：

$$Pr(l+n, s_f) = \sum_{a=0}^{s_f} Pr(U_{i,l,n}^f = a) Pr(l+n-1, s_f - a) \tag{3.8}$$

根据(3.8)上面公式，于是得到问题转化为求解$Pr(U_{i,j,k}^f = a)$的概率。下面我们给出$Pr(U_{i,j,k}^f = a)$ 的求解。根据(3.6)和(3.7)可以得到：

$$Pr(U_{i,j,k}^f = a) = \begin{cases} \int_{g_f a}^{g_f(a+1)} f_{V_{i,j,k}^f}(v)dv, & \text{if } 0 \leq a \leq (x_{k,f} + y_{j,f}) \\ \int_{g_f a}^{\infty} f_{V_{i,j,k}^f}(v)dv, & \text{if } a = x_{k,f} + y_{j,f} \\ 0, & \text{otherwise} \end{cases} \tag{3.9}$$

其中$f_{V_{i,j,k}^f}(v)$为随机变量$V_{i,j,k}^f$的概率密度函数（PDF），根据Ross等人[94] 的工作，可以得到，此PDF没有解析解，下面给出其近似解。

由于接触次数服从泊松分布，于是可以得到用户设备D_i与small cellS_k接触的平均次数为$\lambda_{ik}^B T_d$，用户设备D_i 与用户设备D_j 接触的平均次数为$\lambda_{ij}^D T_d$。用$\lambda_{ik}^B T_d$，$\lambda_{ij}^D T_d$分别代替$M_{i,k}$，$N_{i,j}$，带入(3.6)中可得：

$$V_{i,j,k}^f = \sum_{\omega=1}^{\lambda_{ik}^B T_d} \mathcal{A}_{i,k}^{\omega} + \sum_{\omega=1}^{\lambda_{ij}^D T_d} \mathcal{B}_{i,j}^{\omega} \tag{3.10}$$

由于$\mathcal{A}_{i,k}^{\omega}$服从参数为$B_{ij}^{S}$指数分布，即$\mathcal{A}_{i,k}^{\omega} \sim \mathrm{Exp}(B_{ij}^{S})$，同时$\mathcal{A}_{i,k}^{\omega}$是相互独立的随机变量，根据参考文献[94]，可以得到$\sum_{\omega=1}^{\lambda_{ik}^{B}T_d} \mathcal{A}_{i,k}^{\omega} \sim \mathrm{Gamma}(\lambda_{ij}^{B}T_d, B_{ij}^{S})$。同理可以得到$\sum_{\omega=1}^{\lambda_{ij}^{D}T_d} \mathcal{B}_{i,j}^{\omega} \sim \mathrm{Gamma}(\lambda_{ij}^{D}T_d, B_{ij}^{D})$。

接下来给出$f_{V_{i,j,k}^{f}}(v)$的概率分布函数的计算。记$f_{V_{i,k}^{f}}(v)$为变量$\sum_{\omega=1}^{\lambda_{ik}^{B}T} \mathcal{A}_{i,k}^{\omega}$的概率分布函数，记$f_{V_{i,j}^{f}}(v)$为变量$\sum_{\omega=1}^{\lambda_{ij}^{D}T} \mathcal{B}_{i,j}^{\omega}$的概率分布函数，于是可以得到，$f_{V_{i,j,k}^{f}}(v)$ 的概率分布函数为$f_{V_{i,k}^{f}}(v)$ 和$f_{V_{i,j}^{f}}(v)$的离散卷积和，即

$$f_{V_{i,j,k}^{f}}(v) = \frac{v^{\lambda_{ik}^{B}T_d-1}e^{-vB_{ik}^{S}}}{(B_{ik}^{S})^{-v}\Gamma(v)} \otimes \frac{v^{\lambda_{ij}^{D}T_d-1}e^{-vB_{ij}^{D}}}{(B_{ij}^{D})^{-v}\Gamma(v)} \tag{3.11}$$

由于(3.11)卷积仍然很难求解，基于lu等人的工作[65]，基于Welch-Satterthwaite[94]估计可以得到，

$$f_{V_{i,j,k}^{f}}(v) \approx \frac{v^{\gamma-1}e^{-t\delta}}{\delta^{-\gamma}\Gamma(\gamma)} \tag{3.12}$$

其中

$$\gamma = \frac{(\lambda_{ik}^{B}T_dB_{ij}^{S} + \lambda_{ij}^{D}TB_{ij}^{D})^2}{\lambda_{ij}^{B}T_d(B_{ij}^{S})^2 + \lambda_{ij}^{D}T_d(B_{ij}^{D})^2} \quad \delta = \frac{\lambda_{ik}^{B}T_d(B_{ij}^{S})^2 + \lambda_{ij}^{D}T_d(B_{ij}^{D})^2}{\lambda_{ik}^{B}T_dB_{ik}^{S} + \lambda_{ij}^{D}T_dB_{ij}^{D}}$$

因此$V_{i,j,k}^{f}$可以近似看作服从伽玛分布，即$V_{i,j,k}^{f} \sim \mathrm{Gamma}(\gamma, \delta)$, 于是可以得到$Pr(U_{i,j,k}^{f} = a)$的分布。

3.3.3 编码缓存的传输策略模型

当将small cell 或移动设备缓存的内容传输给请求内容的移动设备时，需要消耗能量。本小节研究的是缓存内容的small cell 或其他设备如何以最小的能量消耗将编码段传输给请求内容的用户，即在接触时间内的最优发射功率。根据3.2.3节的讨论，当移动设备之间是D2D 通信时，可以得到以下的优化问题：

$$\begin{aligned} &\underset{P_T^D}{\text{minimize}} \quad E_D \\ &\text{subject to:} \quad 0 < P_T^D \leq P_{\max}^D. \end{aligned} \tag{3.13}$$

其中目标函数是对于缓存了编码段$\mathcal{B}_{ij}$的用户设备D_j来说，应该以多大的功率将缓存编码段$\mathcal{B}_{ij}$发送给请求内容的用户D_i，能够最节省用户设备D_j的能量消耗。约束条件为移动设备的发射功率介于$(0, P_{\max}^D]$ 之间，其中$P_{\max}^D$ 是移动设备的最大发射功率。

同理，可以得到small cell 基站的最优发射功率可以通过如下的优化问题得到：

$$\begin{aligned} &\underset{P_T^B}{\text{minimize}} \quad E_B \\ &\text{subject to} \quad 0 < P_T^B \leq P_{\max}^B. \end{aligned} \tag{3.14}$$

其中目标函数是对于缓存了编码段$\mathcal{A}_{ik}$的small cell来说，应该以多大的功率将缓存的编码段$\mathcal{A}_{ik}$发送给请求内容的用户设备D_i，能够使得small cell消耗的功率最小。限制条件为small cell 的发射功率P_T^B 介于$(0, P_{\max}^B]$, 其中$P_{\max}^B$是small cell BSs 的最大发射功率。

3.4 绿色移动编码缓存策略求解

本节给出了编码缓存的安置策略模型和编码缓存的传输策略模型求解。基于子模态优化给出了编码缓存安置策略模型的求解，基于优化分析给出了编码缓存的传输策略模型的求解。

3.4.1 编码缓存安置策略模型的求解

对于编码缓存安置策略模型的优化问题(5)，涉及到small cell和用户设备上缓存编码段的个数安置，因此是一个混合整数规划(MIP)问题。如何在small cell网络中和D2D网络中缓存文件已经分别被证明为NP-难问题[55,67]，而此优化问题同时考虑到了在small cell基站和用户设备上如何安置编码段，因此该问题为NP-难问题。

对于缓存安置策略的优化问题，本章利用子模态优化来进行求解。在子模态优化问题中，基于krause 等人[88,95]的研究工作证明，如果目标函数为单调递增的子模态函数，限制条件为拟阵，则可以利用贪婪算法求解。而且如果OPT 表示原问题的最优解，$\mathbf{Z}^*$ 表示利用贪婪算法求出的最优解，则有$\mathbf{Z}^* \geq (1-\frac{1}{e})OPT$成立。即利用贪婪算法可以得到近似最优解。在接下来，首先将原问题转化为子模态优化问题，其次利用贪婪算法求解此问题。

编码缓存安置策略优化问题的转化：定义集合$Z = \{z_{i,f,\nu} | i = 1, \cdots, l+n, f = 1, \cdots, m, \nu = 1, \cdots, s_f\} = Z^1 \cup Z^2$，其中

$$Z^1 = \{z_{k,f,\nu} | k = 1, \cdots, l, f = 1, \cdots, m, \nu = 1, \cdots, s_f\}$$
$$Z^2 = \{z_{j,f,\nu} | j = 1, \cdots, n, f = 1, \cdots, m, \nu = 1, \cdots, s_f\}$$

也就是说，当$i=1,\cdots,l$，$z_{i,f,\nu}=z_{k,f,\nu}$，当$i=l+1,\cdots,l+n$，$z_{i,f,\nu}=z_{j,f,\nu}$。定义集合A_1是在small cell上的缓存方案，其中$A_1\subseteq Z^1$，如果元素$z_{k,f,\nu}\subseteq A_1$，则表示文件F_f的ν个编码段缓存在small cellS_k上。定义集合A_2是在用户设备的缓存方案，其中$A_2\subseteq Z^2$，如果元素$z_{k,f,\nu}\subseteq A_2$，则表示文件F_f的ν个编码段缓存在用户设备D_j。定义集合$Z^1_{k,f}=\{z_{k,f,\nu}|\nu=1,\cdots,s_f\}$表示在small cell$S_k$上缓存的关于文件$F_f$的所有编码段，同理定义集合$Z^2_{j,f}=\{z_{j,f,\nu}|\nu=1,\cdots,s_f\}$表示在用户设备$D_j$上缓存的关于文件$F_f$的所有编码段。于是可以得到将$x_{k,f}$与$y_{j,f}$分别为：

$$x_{k,f}=|A_1\cap Z^1_{k,f}|,\quad y_{j,f}=|A_2\cap Z^2_{j,f}|$$

其中$|\cdot|$表示集合的基数。定义$A=A_1\cup A_2$，于是原问题的目标函数可以表示为：

$$f(A)=\frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m}p_f Pr\left[\sum_{k=1}^{l}\sum_{j=1}^{n}\min\left(\lfloor\frac{V^f_{i,j,k}}{g_f}\rfloor,|A_1\cap Z^1_{k,f}|+|A_2\cap Z^2_{j,f}|\right)\right].$$

定义$Z^1_k=\{z_{k,f,\nu}|f=1,\cdots,m,\nu=1,\cdots,s_f\}$表示在small cell$S_k$上缓存的所有文件的编码段，定义$Z^2_j=\{z_{j,f,\nu}|f=1,\cdots,m,\nu=1,\cdots,s_f\}$表示在用户设备上$D_j$缓存的所有文件的编码段，那么原优化问题的限制条件可以改写为$I=I_1\cup I_2$。其中

$$I_1=\{A_1|g_f|A_1\cap Z^1_k|\le C^s_k,k=1,\cdots,l\}$$
$$I_2=\{A_2|g_f|A_2\cap Z^2_j|\le C^s_D,k=1,\cdots,l\}$$

因此，原优化问题可以表示为：

$$\begin{aligned}&\underset{A}{\text{maximize}}\quad f(A)\\&\text{subject to:}\quad A\subseteq I.\end{aligned}\tag{3.15}$$

定理 3.1 对于优化问题(3.15)，$f(A)$是单调的子模态函数，约束条件(Z,I)为拟阵。

证明.（1）证明$f(A)$是单调的子模态函数。要证$f(A)$为单调的子模态函数，即证$\forall A\subset B\subset Z$和$\forall z_{i,f,\nu}\in Z-B$，$f(A\cup z_{i,f,\nu})-f(A)\ge f(B\cup z_{i,f,\nu})-f(B)\ge 0$成立，也就是说，要证明下面两个式成立。

$$f(A\cup z_{k,f,\nu})-f(A)\ge f(B\cup z_{k,f,\nu})-f(B)\ge 0$$
$$f(A\cup z_{j,f,\nu})-f(A)\ge f(B\cup z_{j,f,\nu})-f(B)\ge 0$$

首先，证明$f(A\cup\{z_{k,f,\nu}\})-f(A)\geq f(B\cup\{z_{k,f,\nu}\})-f(B)\geq 0$成立。设$\forall A\subset B\subset Z$，$E=B-A$，$\forall z_{i,f,\nu}\in Z-B$，有如下等式成立：

$$
\begin{aligned}
&f(A\cup z_{k,f,v})-f(A)=\frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m}p_f\times\Bigg\{Pr\Bigg[\sum_{k'=1,k\neq k'}^{l}\sum_{j=1}^{n}\\
&\min\left(\lfloor\frac{V_{i,j,k}^{f}}{g_f}\rfloor,|A_1\cap Z_{k',f}^{1}|+|A_2\cap Z_{j,f}^{2}|\right)+\min\left(\lfloor\frac{V_{i,k}^{f}}{g_f}\rfloor,|A_1\cap Z_{k,f}^{1}|+1\right)\geq S_f\Bigg]\\
&-Pr\Bigg[\sum_{k'=1,k\neq k'}^{l}\sum_{j=1}^{n}\min\left(\lfloor\frac{V_{i,j,k}^{f}}{g_f}\rfloor,|A_1\cap Z_{k',f}^{1}|+|A_2\cap Z_{j,f}^{2}|\right)+\\
&\min\left(\lfloor\frac{V_{i,k}^{f}}{g_f}\rfloor,|A_1\cap Z_{k,f}^{1}|\right)\geq S_f\Bigg]\Bigg\}
\end{aligned}
$$

其中$V_{i,k}^{f}$表示用户设备D_i从small cell S_k获取文件F_f的大小。利用全概率公式，可以得到：

$$
\begin{aligned}
&f(A\cup z_{k,f,v})-f(A)=\frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m}p_f Pr\left[\lfloor\frac{V_{i,k}^{f}}{g_f}\rfloor\geq|A_1\cap Z_{k,f}^{1}|+1\right]\\
&\times Pr\left[\sum_{k'=1,k\neq k'}^{l}\sum_{j=1}^{n}\min\left(\lfloor\frac{V_{i,j,k}^{f}}{g_f}\rfloor,|A_1\cap Z_{k',f}^{1}|+|A_2\cap Z_{j,f}^{2}|\right)=s_f-|A_1\cap Z_{k,f}^{1}|-1\right]\\
&=\frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m}p_f Pr\left[\lfloor\frac{V_{i,k}^{f}}{g_f}\rfloor\geq|B_1\cap Z_{k,f}^{1}-E_1|+1\right]\\
&\times Pr\Bigg[\sum_{k'=1,k\neq k'}^{l}\sum_{j=1}^{n}\min\left(\lfloor\frac{V_{i,j,k}^{f}}{g_f}\rfloor,|B_1\cap Z_{k',f}^{1}-E_1|+|B_2\cap Z_{j,f}^{2}-E_2|\right)\\
&=s_f-|B_1\cap Z_{k,f}^{1}-E_1|-1\Bigg]
\end{aligned}
$$

对于上式，由于$\min\left(\lfloor\frac{V_{i,j,k}^{f}}{g_f}\rfloor,|B_1\cap Z_{k',f}^{1}-E_1|+|B_2\cap Z_{j,f}^{2}-E_2|\right)$是小于$\min\left(\lfloor\frac{V_{i,j,k}^{f}}{g_f}\rfloor,|B_1\cap Z_{k',f}^{1}|+|B_2\cap Z_{j,f}^{2}|\right)$，并且

$$
\begin{aligned}
|B_1\cap Z_{k,f}^{1}-E_1|&=|B_1\cap Z_{k,f}^{1}|-|E_1\cap Z_{k,f}^{1}|\\
|B_2\cap Z_{j,f}^{2}-E_2|&=|B_2\cap Z_{j,f}^{2}|-|E_2\cap Z_{j,f}^{2}|
\end{aligned}
$$

因此得到：

$$f(A\cup z_{k,f,v})-f(A)=\frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m}p_f Pr\left[\lfloor\frac{V_{i,k}^{f}}{g_f}\rfloor\geq|B_1\cap Z_{k,f}^{1}|-|E_1\cap Z_{k,f}^{1}|+1\right]$$
$$\times Pr\left[\sum_{k'=1,k\neq k'}^{l}\sum_{j=1}^{n}\min\left(\lfloor\frac{V_{i,j,k}^{f}}{g_f}\rfloor,|B_1\cap Z_{k',f}^{1}|-|E_1\cap Z_{k',f}|+|B_2\cap Z_{j,f}^{2}|-|E_2\cap Z_{j,f}^{2}|\right)\right.$$
$$\left.=s_f-|B_1\cap Z_{k,f}^{1}|-|E_1\cap Z_{k,f}^{1}|-1\right]$$
$$\geq\frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m}p_f Pr\left[\lfloor\frac{V_{i,k}^{f}}{g_f}\rfloor\geq|B_1\cap Z_{k,f}^{1}|+1\right]$$
$$\times Pr\left[\sum_{k'=1,k\neq k'}^{l}\sum_{j=1}^{n}\min\left(\lfloor\frac{V_{i,j,k}^{f}}{g_f}\rfloor,|B_1\cap Z_{k',f}^{1}|+|B_2\cap Z_{j,f}^{2}|\right)=s_f-|B_1\cap Z_{k,f}^{1}|-1\right]$$
$$=f(B\cup\{z_{k,f,\nu}\})-f(B)$$

这就证明了$f(A\cup\{z_{k,f,\nu}\})-f(A)\geq f(B\cup\{z_{k,f,\nu}\})-f(B)\geq 0$。同理，可以得到$f(A\cup\{z_{j,f,v}\})-f(A)\geq f(B\cup\{z_{k,f,\nu}\})-f(B)$。至此证明了$f(A)$是一个单调的子模函数。

（2）约束条件$\{Z,I\}$为拟阵。要证明$\{Z,I\}$为拟阵，即证明(Z^1,I_1)和(Z^2,I_2)均为拟阵。对于(Z^1,I_1)，有下面式子成立：

- $\emptyset\in I_1$;
- If $B_1\subseteq I$ and $A_1\subseteq B_1$, then $A_1\subseteq I$;
- If $A_1,B_1\in I_1$ and $|A_1|<|B_1|$, there exists an element $k\in B-A$ that makes $A\cup\{k\}\in I$.

因此，(Z^1,I_1)是拟阵。同理可得(Z^2,I_2)也是拟阵。因此可以得到(Z,I)为拟阵。 □

通过上面证明得到优化问题(3.15)的目标函数为单调递增的子模态函数，约束条件为拟阵。根据上面的分析，利用贪婪算法能够求解。接下来介绍出具体的算法.

编码缓存安置策略算法：具体的编码缓存安置策略算法如下，开始时设置一个缓存集合A为空，在每一次迭代的过程中，加入一个能够使得目标函数具有最大值的编码段，直到达到small cell和用户设备的最大缓存容量。根据本章3.3.2 提到的概率计算，可以计算$\text{argmax}_{z_{i,f,v}\in Z_r}[f(Z_r+z_{i,f,v})-f(Z_r)]$，于是可以得到具体的算法如算法3.1所示：

算法 3.1: 编码缓存安置策略算法

输入: 所有的$z_{i,f}$集合，Z;

Z剩余集合，Z_r;

微基站和用户设备的总存储容量，C;

输出: 在用户设备和微基站上的缓存策略，A

1: $A \leftarrow \emptyset, Z_r \leftarrow Z$;

2: Repeat ;

3: $z_{i^*,f^*,k^*} = \text{argmax}_{z_{i,f,v} \in Z_r}[f(Z_r + z_{i,f,v}) - f(Z_r)]$;

4: $A \leftarrow A + z_{i^*,f^*,v^*}$;

5: $Z_r \leftarrow Z_r - z_{i^*,f^*}$;

6: 如果If $|A \cap Z_{i^*}| = C$, Then $Z_r \leftarrow Z_r \backslash Z_{i^*}$;

7: end if;

8: 直到$|A| > (\sum_{k=1}^{l} c_k^H + \sum_{i=1}^{l} c_i^D)$;

步骤3可以得出在small cell或用户设备上缓存关于文件F_{f^*}的k^*个编码段能够最大化缓存命中率，所以在步骤4中将z_{i^*,f^*,k^*}加入到最优的缓存策略A中。步骤6表示当达到small cell或用户设备的最大缓存容量C 时，其不能够再缓存文件。从步骤8中可以得到当$|A| > (\sum_{k=1}^{l} c_k^H + \sum_{i=1}^{l} c_i^D)$ 时，停止迭代。通过这个算法，可以得到编码缓存安置策略。根据上面的讨论可知，此算法可近似达到最优解。在算法 3.1中，当所有的small cell和用户设备达到其最大的缓存容量，将会有$\sum_{k=1}^{l} c_k^H + \sum_{i=1}^{l} c_i^D$ 次迭代。对于每一次迭代，最多有$(l+n)ms_f$元素没有加入到缓存集合A中。对于每一次计算，其时间复杂度为$\mathcal{O}(n)$, 所以算法 3.1的总体时间复杂度为$\mathcal{O}\left((l+n)mns_f(\sum_{k=1}^{l} c_k^H + \sum_{i=1}^{l} c_i^D)\right)$。

3.4.2 编码缓存传输模型的求解

这一节给出了优化问题(3.13)和(3.14)的解，也就是给出了small cell基站和用户设备的最优发射功率。如定理3.2所示：

定理 3.2 Small cell基站和用户设备的最优发射功率P_T^{D*} 和P_T^{B*}如(3.19)和(3.19) 所示。

证明. 首先给出P_T^D的证明。对于优化问题(3.13)，定义

$$x = P_T^D, \quad \delta = \frac{r_D^2}{\sigma_D^2} \quad \theta = \eta_D P_C^D$$

于是可以得到，

$$f(x) = \frac{\mathcal{B}_{ij}(x+\theta)}{\eta_D W_D \log_2^{(1+\delta x)}} \tag{3.16}$$

对(3.16)关于x求导，于是可以得到：

$$f'(x) = \frac{\mathcal{B}_{ij}[(1+\delta x)\log_2^{(1+\delta x)} \ln 2 - (1+\delta x) - (\delta\theta - 1)]}{\eta_D W_D \log_2^{2(1+\delta x)}(1+\delta x)\ln 2} \tag{3.17}$$

在(3.17)中，$f'(x)$的分母恒为正，于是当$f'(x)$的分子大于0时，则$f'(x) > 0$，反之亦然。由于$x \in (0, P_{\max}^D]$，故$1+\delta x \in (1, 1+\delta P_{\max}^D]$。令$y = 1+\delta x$，$\xi = \delta\theta - 1$，那么(3.17)可以化为：

$$f'(y) = \frac{\mathcal{B}_{ij}[y\log_2^y \ln 2 - y - \xi]}{\eta_D W_D \log_2^{2y}(y)\ln 2} \tag{3.18}$$

令$g(y) = Fy\log_2^y \ln 2 - y - \xi$，对$g(y)$关于$y$求导，可以得到$g'(y) = F\log_2^y \ln 2$，从而可以推出，当$y \in (1, 1+\delta P_{max}]$时，$g(y)$是单调递增函数。于是，当$y \to 1$ 时，$g(y) < 0$。若$g(1+\delta P_{\max}^D) < 0$时，则$f'(y) < 0$，于是得到$f(x)$是单调递减函数，所以在$x = P_{\max}^D$ 时取得最小值。若$g(1+\delta P_{\max}^D) > 0$，则存在零点$y_0$使得$g(y_0) = 0$，于是可以得到函数$f(x)$ 先是单调递减，然后单调递增，所以在y_0点取得最小值。根据上面的讨论，可以得到最优的P_T^{D*} 为：

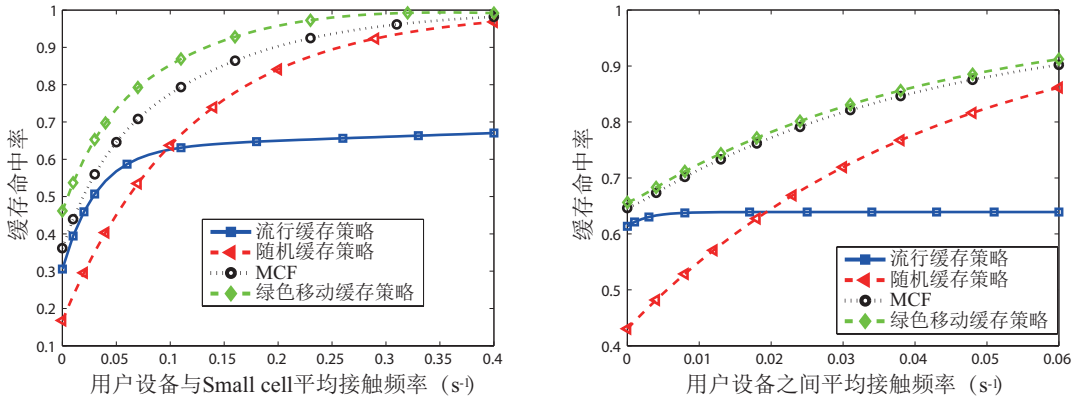
$$P_T^{D*} = \begin{cases} P_{\max}^D, & \text{if } g(1+\alpha P_{\max}^D) < 0, \\ P_{y_0}, & \text{if } g(1+\alpha P_{\max}^D) \geq 0 \end{cases} \tag{3.19}$$

同理可以得到

$$P_T^{B*} = \begin{cases} P_{\max}^B, & \text{if } g(1+\delta P_{\max}^B) < 0, \\ P_{y_0}, & \text{if } g(1+\delta P_{\max}^B) \geq 0 \end{cases} \tag{3.20}$$

至此定理3.2证明完毕。 □

通过定理3.2，得到了small cell和用户设备的最优发射功率分别为P_T^{D*}和P_T^{B*}。将其带入(3.2)和(3.4) 中，可以得到编码缓存传输消耗的能量为E_D^*和E_B^*。从而可以得



(a) 用户设备与small cell平均接触频率($\overline{\mu}_{i,k}$) 对缓存安置策略命中率的影响。

(b) 用户设备之间平均接触频率($\overline{\lambda}_{i,j}$) 对缓存安置策略命中率的影响。

图 3.3 移动性对缓存安置策略命中率影响分析。

到网络传输消耗的平均能量为:

$$\bar{E}^* = \frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m} p_f \left(\sum_{\omega=1}^{M_{i,k}} E_B^*(\mathcal{A}_{i,k}^{\omega}) + \sum_{\omega=1}^{N_{i,j}} E_D^*(\mathcal{B}_{i,j}^{\omega}) \right), \tag{3.21}$$

由于接触的次数服从泊松分布，利用用户设备D_i与small cellS_k接触的平均次数$\lambda_{i,k}^{B}T_d$，用户设备D_i与用户设备D_j的接触的平均次数$\lambda_{i,j}^{D}T_d$来代替$M_{i,k}$和$N_{i,j}$，于是可以得到

$$\bar{E}^* = \frac{1}{n}\sum_{i=1}^{n}\sum_{f=1}^{m} p_f \left(\sum_{\omega=1}^{\lambda_{i,k}^{B}T_d} E_B^*(\mathcal{A}_{i,k}^{\omega}) + \sum_{\omega=1}^{\lambda_{i,j}^{D}T_d} E_D^*(\mathcal{B}_{i,j}^{\omega}) \right) \tag{3.22}$$

即缓存内容传输时网络的平均能耗。

3.5 实验结果与分析

3.5.1 实验的设置

本节对绿色移动编码缓存策略进行评估。和第二章类似，考虑一个区域内含有5个small cell，60个用户。对于缓存的请求的时延，根据Wang等人[64]的工作，设置$T_d = 600s$。设置文件的流行性服从参数为$\gamma = 0.8$的Zipf 分布。对于请求的文件，根据文献[55]，设置文件的大小为$10MB$，其中包括10^4个文件，每个文件可以被编码为2 个编码段。对于small cell和用户设备的缓存容量来说，设置small cell最多能缓存

表 3.2 能量消耗的仿真参数

参数	取值
small cell的最大传输功率，P_{max}^{S}	6.3 W
small cell的固定功率，P_{C}^{S}	56 W
small cell的信道带宽，W_S	20 MHz
small cell的噪声功率，σ_B^2	-104dBm
small cell的功率放大器效率因子，$1/\beta_S$	0.38
D2D通信的带宽，W_S	$20MHz$
用户设备的最大传输功率P_{max}^{D}	0.2W
用户设备的固定功率，P_{C}^{D}	115.9mW
用户的噪声功率，σ^2	-95dBm
用户设备的功率放大器效率因子，$1/\beta_D$	0.2
路径损耗因子，α，β	4

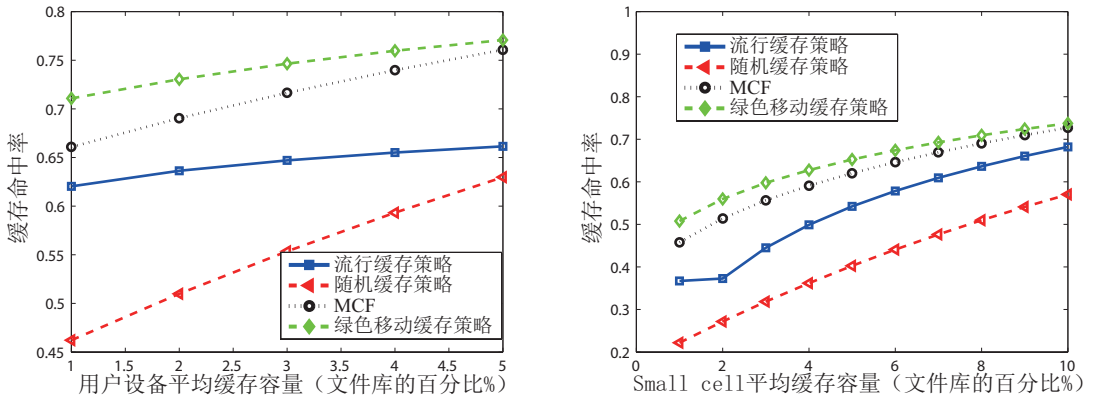
文件库的10%，移动设备最多能缓存文件库的5%。对于用户移动性的设置，和第二章一样，用$\Gamma(4.43, 1/1088)$ 表示用户设备D_i 与用户设备D_j 接触时间的指数分布参数，用$\Gamma(10, 1/100)$ 表示用户设备D_i与small cell 基站S_k 接触时间的指数分布参数。根据文献[55][64]，分别设置small cell和用户设备在一次接触时间内的传输量服从均值为$20MB$和$10MB$的指数分布。根据Zhang等人[96]和Chen等人[93]的工作分别给出了在small cell和用户设备上的能量消耗的参数设置。具体如表 3.2 所示，在本章中，主要从编码缓存安置策略命中率和编码缓存传输能耗两个角度来评估绿色移动编码缓存策略。

3.5.2 对比试验

本节将绿色移动编码缓存策略与三种不同的缓存策略（流行缓存策略[43]，随机缓存策略[47]和MCF缓存策略（移动性的缓存策略，每次传输量为固定值）[55,64,97]）进行比较，其中流行缓存策略和随机缓存策略具体设置和第二章类似，对于MCF的缓存策略，设置每次传输的量为随机变量$\mathcal{A}_{i,k}^{\omega}$和$\mathcal{D}_{i,j}^{\omega}$ 的平均值。

3.5.3 编码缓存安置策略命中率分析

图3.3讨论了用户移动性对四种不同缓存安置策略命中率的影响。和第二章类似，仍以平均接触次数$\overline{\mu}_{i,k}$和$\overline{\lambda}_{i,j}$来刻画横坐标。从图3.3 中可以看出，本章提出的缓



(a) 用户设备缓存容量大小对缓存命中率的影响　(b) small cell缓存容量大小对缓存命中率的影响

图 3.4 用户设备和small cell缓存容量大小对缓存命中率的影响

存方案要优于其他缓存方案，这是因为在流行缓存策略中只缓存了流行的内容，而随机缓存策略中缓存的内容是随机的，这两种方案都没有考虑到用户的移动性，虽然MCF缓存策略考虑了用户的移动性，但是没有考虑到在接触时间内能否将缓存内容全部传递，所以本章提出的缓存策略性能最优。

图3.4a和图3.4b讨论了用户设备和small cell的缓存容量大小与缓存命中率的关系。图3.4的横坐标表示用户设备或small cell的最大的缓存容量占文件库的比重，从图中可以得出，当用户设备或small cell 的存储容量变大时，缓存的命中率变大，这是因为缓存容量变大后，能够缓存更多的内容，从而使得缓存命中率变大。此外，对比其他缓存策略，本章提出的编码缓存安置策略比其他缓存策略都优。

3.5.4 编码缓存传输策略能量分析

图3.5和图3.6讨论了用户缓存命中率和传输能量消耗的关系。本文将能量消耗进行了归一化处理。从图中可以得到，随着缓存命中率的增加，能量消耗也在增加。这是因为随着缓存命中率的增加，请求文件的用户从small cell 基站和其他用户上获取文件的概率增加，从而增加了small cell和用户设备能量的消耗。此外，图3.5a表明，当用户与small cell接触频率$\bar{\mu}_{i,k}$（即进入small cell的覆盖范围内）增加时，在相同的缓存命中率下，$\bar{\mu}_{i,k}$越大，其消耗的传输能量就越少。这是因为随着$\bar{\mu}_{i,k}$的增加，用户的移动性增加，从而导致了缓存命中率的增加，所以在相同缓存命中率下，$\bar{\mu}_{i,k}$越大，消耗的能量越少。同样，从图3.5b中可以得出，在相同的缓存命中率下，用户设备之间的接触频率$\bar{\lambda}_{i,j}$越大，消耗的传输能量就越小。图3.6表明，在一定的

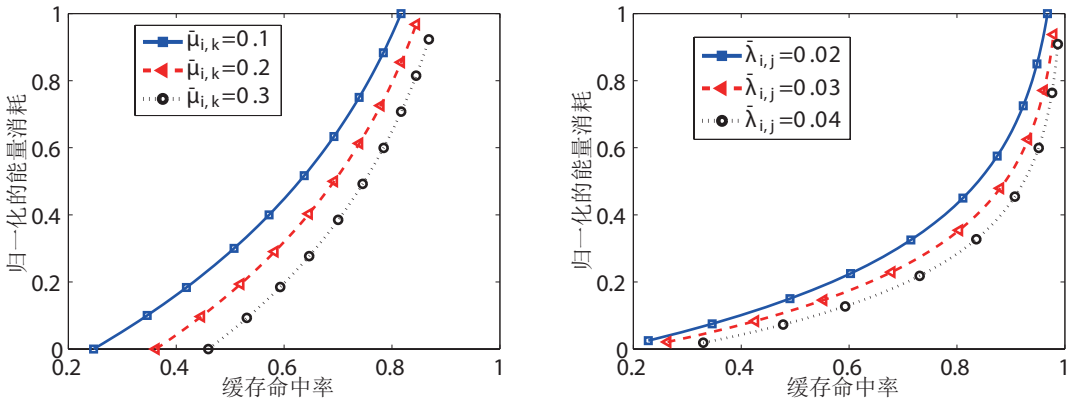


图 3.5 用户移动下能量消耗与缓存命中率的关系

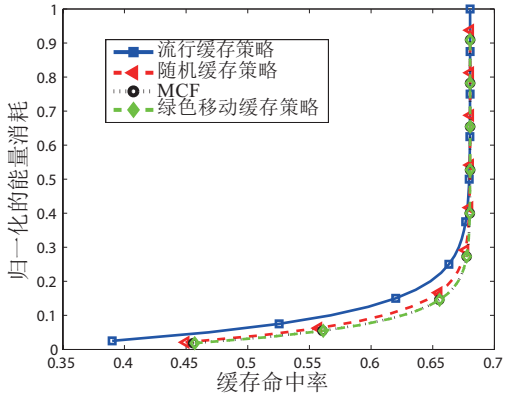


图 3.6 不同策略下能量消耗与缓存命中率的关系

缓存命中率下，本章提出的方案比流行性缓存策略，随机缓存策略和MCF缓存策略更能节省能量，但与MCF缓存策略的相差较小。这是因为两种策略都考虑到了用户的移动性，但本章同时考虑了接触时间的随机性。

3.6 本章小节

本章分析了用户移动性导致的接触时间随机性对small cell和用户设备上缓存内容安置和传输的影响，并基于编码缓存建立了最大缓存命中率的安置策略模型和最小能耗的传输策略模型。继而，基于子模态优化提出了缓存文件的安置策略，进一步得出了small cell 基站和用户设备的最优发射功率，从而得出网络的最小传输能

耗。最后，通过实验验证了本章提出的缓存策略的有效性，与其他缓存策略相比较，本章提出的缓存策略有效的提高了缓存命中率，减少了能量消耗。同时得出当用户移动性较高时，缓存内容传输所消耗的能量较少；当用户移动性较低时，缓存内容传输所消耗的能量较多。

4 可再生能量供电下的5G移动边缘云计算

本章首先提出了5G网络下可再生能源供电的移动边缘云计算框架，其次基于对可再生能量的分析，建立了用户任务时延和电网供电能耗的最小化模型，通过交替优化技术给出了可再生能源供电下的计算任务卸载策略，最后给出了实验验证。

4.1 引言

在5G网络中，部署在宏基站和small cell的边缘云具有一定的存储和计算资源，能够为移动用户提供计算任务卸载的服务[81,98]，即移动用户可以将计算密集型的任务卸载到宏基站云（部署在宏基站上的边缘云）或small cell 基站云（部署在small cell上的边缘云），这种计算卸载模式称为移动边缘云计算（mobile-edge cloud computing）。与现有的计算模式相比较，移动边缘云计算[82,99] 主要有两个优点：（1）与本地处理和移动边缘计算相比较[32,87,100]，移动边缘云计算克服了移动终端有限计算能力的限制；（2）与传统的移动云计算[73,77] 相比较，移动边缘云计算的传输延时小，避免了在网络较差情况下，将计算密集型任务卸载到云端产生的时延较长。因此，移动边缘云计算具有较强的计算能力而且距移动终端较近，从而可以满足延时敏感类计算任务的服务要求。许多研究者对移动边缘计算进行了初步的探讨，主要关注的是如何利用当时的通信环境及边缘云的计算资源进行任务的卸载。比如Chen等人[82] 研究了多用户在边缘云的计算任务卸载策略，利用纳什均衡给出了一种有效的多用户在边缘云上的卸载策略。Hao 等人[83] 给出了在用户移动环境性，如何在宏基站云和small cell 基站云上进行计算任务的卸载的策略。表 4.1从计算的资源池，是否面向5G 网络，是否由可再生能源供电，计算能力，计算任务的时延和计算任务的能耗等方面对几种不同的计算任务卸载策略进行比较。

然而，边缘云是部署在5G超密集蜂窝网络中的，超密集的部署small cell带来了巨大的能量消耗，给5G 的绿色性带来了新的挑战。根据Mao等人[15]，Zhang[96]等人的研究，采用可再生能源对small cell 基站供电是一个解决5G高能耗的可行方案。比如在Zhang等人[96] 中讨论可再生能源供电下的网络资源分配以及流量卸载。对于可再生能源供电下的移动边缘云计算，由于可再生能量到达具有随机性以及small cell基站的电池容量有限，导致了边缘云服务器计算能力的动态性，而现有的关于移动边缘云计算相关研究大多是考虑基于电网供电的，没有考虑到计算卸载时的能量随

表 4.1 几种不同的计算任务卸载策略比较

参考文献 提出的策略	计算 资源池	面向5G 网络	可再生 能量	计算 能力	计算任务 时延	计算任务 能耗
[73] [77]	远端云	否	否	高	高	高
[32] [87]	微云	是(D2D)	否	低	高	低
[100] [26]	微云	是(D2D)	否	低	高	低
[99]	边缘云	否	否	中	低	中
[82]	边缘云	是	否	中	低	中
TORE	边缘云	是	是	中	低	低

机性和计算能力动态性的约束，所以目前的计算卸载策略并不适用。因此，如何设计出既保证用户任务时延性要求，又能充分利用可再生能源的卸载策略是一个挑战性的问题。

为此，本章提出了可再生能源供电下的移动边缘云计算框架，通过对可再生能量的分析，建立了用户时延和网络能耗最小的优化模型。通过采用交替优化技术给出了此框架下的计算任务卸载策略（TORE），此策略既能满足用户任务的时延要求，又能最大程度的使用可再生能量。具体来说，此计算任务卸载策略包括两部分：（1）任务安置问题，即当small cell 基站是由可再生能源供电的条件下，每一个任务是安置在宏基站云还是small cell 基站云上；（2）计算资源分配问题，即边缘云给每一个任务分配多少计算资源。最后，通过实验验证了该策略至少能够减少20%的任务延时，降低30%的能量消耗。综上所述，本章的主要贡献分为以下两点：

- 基于可再生能量供电下的移动边缘云计算：本章首次提出了5G网络下由可再生能源供电的移动边缘云计算框架，并分析了边缘云计算中的任务安置问题和计算资源分配问题。
- 任务卸载策略分析：本章提出了一种有效的计算任务卸载策略（TORE），不仅能够充分利用随机到达的可再生能量，而且能够满足用户的延时需求。

本章节组织如下：第4.2节提出了系统框架，第4.3节对可再生能量进行了分析并建立了模型，第4.4节提出了可再生能量供电下的任务卸载策略，第4.5节给出了实验结果与分析，第4.6 节对本章进行了小节。

4.2 系统框架与描述

这一节描述了可再生能源供电下的5G超密蜂窝网移动边缘云计算框架，如图4.1所示，此框架包括5G 超密蜂窝网络，可再生能量的获取，可再生能量的存贮和主电网。具体来说，5G超密蜂窝网络包括宏基站，small cell 基站和用户终端。其中宏基站是由主电网供电，small cell是由可再生能源（比如，太阳能和风能) 供电。基于此框架，本章研究的计算任务处理场景如下：用户终端的计算任务通过蜂窝网络卸载到部署在宏基站或small cell 基站的边缘云上进行处理，当边缘云将计算任务处理完成后，它将计算结果反馈给用户终端。为了简单起见，本章仅考虑一个宏基站，一个small cell和n个需要处理的计算任务。而多个small cell的情形是单个情形的扩展，可以类似地进行分析，故本章不再赘述。假设部署在宏基站和small cell 基站边缘云的计算资源（CPU cycles per second）分别为$C = \{c_0, c_1\}$，其中c_0 表示的是部署在宏基站边缘云的总计算资源，c_1 表示的是部署在small cell 基站边缘云的总计算资源。用户终端有n个计算任务记为$Q = \{Q_1, Q_2, \cdots, Q_n\}$。

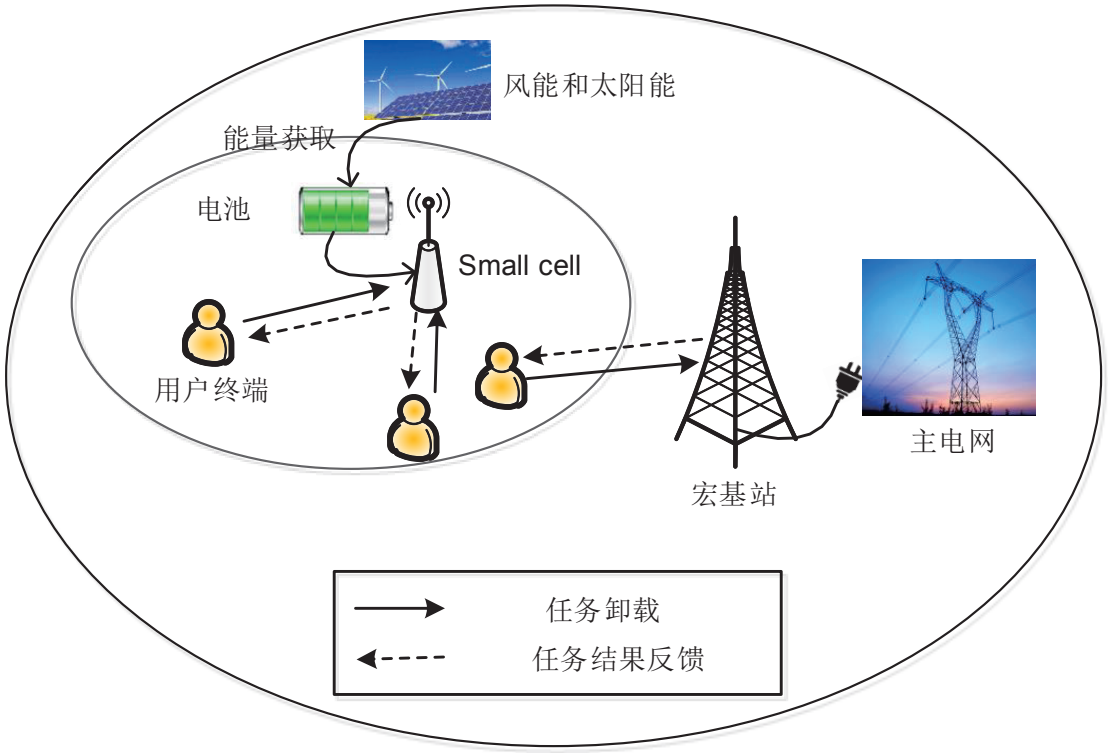


图 4.1 可再生能量供电下的5G网络。

4.2.1 通信模型

首先介绍通信模型。定义h_i^m和P_i^m分别表示用户设备将计算任务Q_i卸载到边缘云的信道增益和发射功率。根据香农定理可以得到用户终端将计算任务卸载到边缘云的上行链路速率r_i^u 为：

$$r_i^u = B_m \log_2 \left(1 + \frac{P_i^m (h_i^m)^2}{\sigma_m^2}\right) \tag{4.1}$$

其中σ_m^2表示复高斯白噪声，B_m表示上行链路信道传输带宽。类似地可以得到，宏基站或small cell 将计算任务结果反馈给用户终端的下行链路速率r_i^d为：

$$r_i^d = B_b \log_2 \left(1 + \frac{P_T^b (h_i^b)^2}{\sigma_b^2}\right) \tag{4.2}$$

其中B_b为下行链路的信道传输带宽，P_T^b表示基站（宏基站和small cell）的发射功率，h_i^b为基站的信道增益，σ_b^2为复高斯白噪声。

4.2.2 计算卸载模型

其次介绍计算任务卸载模型。基于Chen等人的工作[82]，将需要卸载的计算任务Q_i表示为：$Q_i = \{\omega_i, s_i\}$，其中ω_i 表示任务Q_i 的计算量，即完成任务Q_i总共需要的CPU周期/秒（CPU cycles per second），s_i 表示任务Q_i 的数据量，即需要上传到边缘云的数据文件（比如代码和参数）的大小。接下来讨论在5G超密蜂窝网下移动边缘云计算的任务时延和能量消耗。

（1）**任务时延**：计算任务处理的时延包括以下三方面：1) 用户终端将计算任务卸载到边缘云消耗的时间；2) 在边缘云处理计算任务消耗的时间；3) 宏基站或small cell 基站将计算结果反馈给用户消耗的时间。因此，第i个计算任务的总体时延为：

$$t_i = \frac{\omega_i}{\kappa_i^{\gamma_i} c_{\gamma_i}} + \frac{s_i}{r_i^u} + \frac{\beta s_i}{r_i^d} \tag{4.3}$$

其中γ_i 表示第i个任务是卸载宏基站边缘云上还是small cell上边缘云上。如果$\gamma_i = 0$，则表示将任务卸载到部署在宏基站的边缘云上；如果$\gamma_i = 1$，则表示将任务卸载到部署在small cell的边缘云上。c_{γ_i} 表示部署在宏基站或small cell 基站边缘云的总计算能力。$\kappa_i^{\gamma_i}$ 表示边缘云分配给任务i的计算能力所占边缘云总计算能力的比例。根据Chen等人[82]的研究表明，计算任务处理完后的数据量大小要远远小于计算任务卸载前数据量的大小，因此本章不考虑宏基站或small cell将计算结果反馈给用户消耗的时间。

表 4.2 符号表

变量名	解释
Q_i	用户的第i个计算任务
ω_i	任务Q_i 的计算量
s_i	任务Q_i 的计算量
$E_{\max}$	电池的总容量
E	单位能量
λ_E	可再生能量的到达速率
μ_E	可再生能量的消耗速率
$P_E^{\gamma_i}$	在边缘云上处理计算任务的功率消耗
$P_T^{\gamma_i}$	宏基站或small cell的传输功率
$P_C^{\gamma_i}$	宏基站或small cell的电路功率消耗

（2）**网络能量消耗：**由于用户终端将计算任务卸载到边缘云上不消耗网络的能耗，所以网络能耗包括以下两部分：1）边缘云上的处理计算任务消耗的能量；2）宏基站或small cell将计算结果反馈给终端用户时消耗的消耗。因此，第i 个计算任务消耗的网络能量为：

$$\varepsilon_i = \frac{\omega_i}{\kappa_i^{\gamma_i} c_{\gamma_i}} P_E^{\gamma_i} + \frac{\beta s_i}{r_i^d}\left(\frac{1}{\eta} P_T^{\gamma_i} + P_C^{\gamma_i}\right) \tag{4.4}$$

其中$P_E^{\gamma_i}$ 表示在边缘云上处理计算任务的功率消耗，$P_T^{\gamma_i}$ 表示宏基站或small cell 基站的发射功率，$P_C^{\gamma_i}$ 表示宏基站或small cell的固定功率消耗，η 表示功率放大器的效率因子。

4.2.3 可再生能量获取模型

这一小节给出可再生能量获取模型。虽然将可再生能量引入边缘云计算可以减少电网的功耗，同时增强无线设备工作的可持续性。然而，可再生能量（比如太阳能、风能等）可能受到天气等因素的影响，能量的到达呈现随机的特性。基于Zhang等人的工作[96]，将单位能量记为E，并且定义λ_E 为small cell 基站单位时间内到达单位能量的个数，即能量到达的速率。假设每一个时刻到达的能量都是相互独立的，由于small cell 基站在有能量供给时是可以持续工作的，所以不妨假设small cell在单

位时间消耗的能量是固定的。根据4.2.2节对small cell 的能量消耗的分析，可以得到small cell单位时间内消耗的单位能量，即：能量消耗速率为：

$$\mu_E = \frac{1}{E}\left(P_E^1 + \frac{1}{\eta}P_T^1 + P_C^1\right) \tag{4.5}$$

通过上面分析，可再生能量的供给和消耗均是时间随机的，所以可再生能量的到达模型可以用排队论来描述，而电池的容量可以表示为队列的长度。值得注意的是，small cell 基站的电池容量也是有限的，那么当可再生能源产生的能量多于small cell消耗的能量时，且small cell的电池容量没有到达最大容量时，此时是对small cell基站的电池充电。

4.3 可再生能量分析与模型的建立

这一节给出可再生能量分析和问题的定义，本节的目标是设计一个不仅能够满足任务时延又能最大程度上利用可再生能源的计算任务卸载策略。

4.3.1 可再生能量分析

根据4.2.3节的讨论可知，可再生能量的到达速率λ_E是时间随机的，和Zhang等人的工作[96]一样，本章假设λ_E 服从泊松分布。可再生能量的消耗速率μ_E 是一定的，如公式 (4.5) 所示。基于此，可以得到small cell的电池容量模型，可以$M/D/1$ 的队列模型来描述。于是，电池的总容量为：$E_{\max} = mE$，其中m 为队列的长度。因此，电池容量共有$(m+1)$ 个状态：$\{0, 1, 2, \cdots, m\}$，其中状态0表示small cell的电池没电，状态m表示small cell 基站的电池容量达到电池的总容量mE。基于$M/D/1$ 队列理论，电池容量各个状态之间转换可用马尔可夫链来表示。具体来说，可用维度为$(m+1)\times(m+1)$的矩阵表示其转移概率矩阵P_{ij}，如下所示：

$$\mathbf{P} = \begin{bmatrix} p_0 & p_1 & p_2 & \cdots & p_m \\ p_0 & p_1 & p_2 & \cdots & p_m \\ 0 & p_0 & p_1 & \cdots & p_{m-1} \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & p_1 \end{bmatrix}. \tag{4.6}$$

其中

$$p_i = \frac{1}{i}(\frac{\lambda_E}{\mu_E})^{-i} \exp\left(-\frac{\lambda_E}{\mu_E}\right), i = 0, 1, 2, \cdots, m \tag{4.7}$$

根据排队论知识可得：当$\lambda_E \geq \mu_E$ 时，队列的长度是不稳定的，此时可再生的能量很充足；当$\lambda_E \leq \mu_E$，则队列存在着平稳分布。

当队列存在稳定状态时，记$\rho_E = \lambda_E/\mu_E$，根据Pollaczek-Khinchin 公式, 可以得到电池容量的稳定状态π_i为：

$$\begin{aligned}
&\pi_0 = 1 - \rho_E \\
&\pi_1 = (1 - \rho_E)(e^{\rho_E} - 1) \\
&\pi_i = (1 - \rho_E)\left\{ e^{n\rho_E} + \sum_{k=1}^{i-1} e^{k\rho_E}(-1)^{i-k}\left[\frac{(k\rho_E)^{i-k}}{(i-k)!} + \frac{(k\rho_E)^{i-k-1}}{(i-k-1)!}\right]\right\} \quad (i \geq 2)
\end{aligned}$$

进一步可以得到当small cell上没有可再生能量处理计算任务时，即电池的容量状态从$1 \to 0$时概率为：

$$\pi_1 P_{21} = \pi_1 p_0 = (1 - \rho_E)(1 - e^{-\rho_E}) \tag{4.8}$$

基于公式(4.8)，得到当计算任务无法在small cell边缘云上处理，用户需要将计算任务卸载到宏基站边缘云上进行任务的处理的概率。于是，存在这样的一个问题，如何设计计算任务在宏基站和small cell边缘云上的安排策略以及边缘云上计算资源的分配策略，能使得此策略既能最大程度的利用可再生能源，又能使得任务处理延时较低。

4.3.2 模型的建立

对于计算任务的卸载策略，根据上面讨论，本节关注的问题有两个：1）任务安置问题：当small cell与部署在small cell 基站的边缘云由可再生能源供应时，计算任务应该卸载到宏基站边缘云上还是small cell边缘云上。2）计算资源分配问题：宏基站边缘云或small cell 基站边缘云对于每个计算任务应该分配多少计算资源。关注的目标是保证用户计算任务时延的基础上，减少电网的能量消耗，增加可再生能源的使用。可以注意到，计算任务处理的时延和可再生能源的使用存在一个折中：如果所有的用户都将计算任务卸载到small cell边缘云上，则会减少宏基站边缘云上的计算任务处理的和传输任务结果的能量消耗，从而最大程度上使用了可再生能源，减

少了电网能量的供应。但是small cell的计算能力有限并且当small cell 基站上没有可再生能量供电时，就会增加任务处理的时延。

下面根据任务的处理时延（task duration）和能量消耗（energy cost）给出任务卸载问题的定义。根据4.2.2节的讨论, 可以得到处理n 个计算任务的总体时延为: $\sum_{i=1}^{n} t_i = \sum_{i=1}^{n}(\frac{\omega_i}{\kappa_i^{\gamma_i} c_{\gamma_i}} + \frac{s_i}{r_i^u})$。根据4.3.1节的讨论，第$i$ 个任务需要卸载到宏基站边缘云上的概率为$(1-\rho_E)(1-e^{-\rho_E})$，于是可以得到这$n$个任务需要消耗的平均电网能耗为：$\sum_{i=1}^{n} e_i = \sum_{i=1}^{m}(1-\rho_E)(1-e^{-\rho_E})\varepsilon_i$。那么计算任务卸载问题可以描述为如下优化问题：

$$\underset{\kappa,\gamma}{\text{minimize}} \qquad f(\kappa,\gamma) = \sum_{i=1}^{n}(\alpha^t t_i + \alpha^e e_i) \tag{4.9a}$$

$$\text{subject to} \qquad \sum_{i\in o_{\gamma_i}} \kappa_i^{\gamma_i} \le 1, \quad \gamma_i = 0, 1. \tag{4.9b}$$

其中$\alpha^t, \alpha^e \in [0,1]$ 分别表示任务i对于任务时延和能量消耗决策的权重参数。

具体来说：（1）当计算任务是延时敏感型任务时（比如，视频流任务），与网络能量消耗相比，需要更多地考虑任务时延，因此可以设置$\alpha^t > \alpha^e$，在一些特殊的情况可以设置$\alpha^t = 1$，$\alpha^e = 0$。（2）当计算任务不是延时敏感型任务时，可以设置$\alpha^t < \alpha^e$, 从而保证在满足用户延时要求时，最大程度的使用可再生能源，从而减少电网能量的消耗。o_{γ_i} 表示所有安置在γ_i 上的计算任务。优化问题的目标函数(4.9a)是最小化任务的延时和电网的能量消耗，约束条件(4.9b)表示在宏基站和small cell部署的边缘云资源分配不能超过其最大值。

综上所述，此优化问题一方面保证用户任务时延，另一方面，最大程度利用可再生能源，从而减少电网的能量消耗。

4.4 可再生能量供电下任务卸载策略求解

这一节针对优化问题(4.9a)和(4.9b)给出可再生能源供电下的计算任务卸载策略。对于目标函数，将任务的时延和能耗的具体表达式代入(4.9a)，进而得到如下式子：

$$\begin{aligned} f(\kappa,\gamma) &= \sum_{i=1}^{n}(\alpha^t t_i + \alpha^e e_i) \\ &= \sum_{i=1}^{n}\left[\alpha^t(\frac{\omega_i}{\kappa_i^{\gamma_i} c_{\gamma_i}} + \frac{s_i}{r_i^u}) + \alpha^e(1-\rho_E)(1-e^{-\rho_E})\left(\frac{\omega_i}{\kappa_i^{\gamma_i} c_{\gamma_i}} P_E^{\gamma_i} + \frac{\beta s_i}{r_i^d}\left(\frac{1}{\eta} P_T^{\gamma_i} + P_C^{\gamma_i}\right)\right)\right]. \end{aligned}$$

对于优化问题的求解，定义$\mathbb{K}$和$\mathbb{H}$ 是κ和γ的可行集。由于目标函数$f(\kappa,\gamma)$是非凸的，所以此优化问题是关于κ和γ 的混合整数非凸优化问题，但是可以通过交替优化技术转化为下面两个子问题求得原问题的解。1）**计算资源分配子问题：**给定$\gamma=\gamma^0\in\mathbb{H}$，当$\gamma_i$ 固定时，优化问题(4.9a) 和(4.9b)关于κ_i是凸优化问题。基于此，可以利用KTT求得关于κ 的最优解并且记为$f(\kappa^*,\gamma_0)$。2）**任务安置子问题：**基于上面的最优解κ^*，子问题$f(\kappa^*,\gamma)$ 是关于γ的$0-1$ 整数规划问题。在本节中利用任务安置算法给出此子问题的解。下面给出这两个子问题的具体求解过程并给出这两个子问题的解收敛于原问题的解。

4.4.1 计算资源的分配问题

引理 4.1 给定$\gamma=\gamma^0\in\mathbb{H}$，优化问题(4.9a)和(4.9b)是关于$\kappa_i$ 的凸优化问题。

证明. 给定$\gamma=\gamma^0=(\gamma_1^0,\gamma_2^0,\cdots,\gamma_n^0)\in\mathbb{H}$，也就是，任务安置在宏基站边缘云还是small cell边缘云已经给定。定义$j=\gamma_i^0$，$A=(1-\rho_E)(1-e^{-\rho_E})$，于是优化问题的目标函数(4.9a) 可以表示为下式：

$$f(\kappa,\gamma^0)=\sum_{i=1}^{n}\alpha^t(\frac{\omega_i}{\kappa_i c_j}+\frac{s_i}{r_i^u})+\sum_{i=1}^{n}\alpha^e A\left[\frac{\omega_i}{\kappa_i c_j}P_E^j+\frac{\beta s_i}{r_i^d}\left(\frac{1}{\eta}P_T^j+P_C^j\right)\right] \tag{4.10}$$

其中$f(\kappa,\gamma^0)$是关于$(\kappa_1,\kappa_2,\cdots,\kappa_n)$ 的函数。于是可以得到函数$f(\kappa,\gamma^0)$的海森矩阵(Hessian matrix)$\mathbf{H}$ 如下：

$$\mathbf{H}=\begin{bmatrix}\frac{\partial^2 f}{\partial^2\kappa_1} & \frac{\partial^2 f}{\partial\kappa_1\partial\kappa_2} & \cdots & \frac{\partial^2 f}{\partial\kappa_1\partial\kappa_n}\\ \frac{\partial^2 f}{\partial\kappa_2\partial\kappa_1} & \frac{\partial^2 f}{\partial^2\kappa_2} & \cdots & \frac{\partial^2 f}{\partial\kappa_2\partial\kappa_n}\\ \vdots & \vdots & & \vdots\\ \frac{\partial^2 f}{\partial\kappa_n\partial\kappa_1} & \frac{\partial^2 f}{\partial\kappa_n\partial\kappa_2} & \cdots & \frac{\partial^2 f}{\partial^2\kappa_n}\end{bmatrix}.$$

海森矩阵$\mathbf{H}$中每一个元素值为：

$$\frac{\partial^2 f}{\partial\kappa_i\partial\kappa_j}=\begin{cases}\frac{2\omega_i(\alpha^t+\alpha^e P_E^j)}{\kappa_i^3 c_j} & \text{如果}\quad i=j\\ 0 & \text{其他}\end{cases} \tag{4.11}$$

由于$\frac{2\omega_i(\alpha^t+\alpha^e P_E^j)}{\kappa_i^3 c_j}$中的分子和分母都大于0，于是可得$\frac{\partial^2 f}{\partial\kappa_i\partial\kappa_j}\geq 0$。继而可以得出海森矩阵$\mathbf{H}$所有的特征值为$\frac{2\omega_i(\alpha^t+\alpha^e P_E^j)}{\kappa_i^3 c_j}$且均为正值，从而可以得到海森矩阵$\mathbf{H}$ 是对称正定矩阵。根据参考文献[101]中凸函数的二阶条件判定定理，可以得到优化问题

的目标函数(4.9a) 关于κ_i 是凸函数。由于优化问题的限制条件(4.9b)关于κ_i是线性的，因此该优化问题是关于κ_i 的凸优化问题。 □

根据引理 4.1，利用KKT条件，可以得到定理 4.1。

定理 4.1 给定$\gamma = \gamma_0 \in \mathbb{H}$，优化问题(4.9a)和(4.9b)关于$\kappa$ 的最优解如公式(4.14)所示。

证明. 根据引理 4.1，可以得到优化问题(4.9a)和(4.9b)的拉格朗日（Lagrange）函数为：

$$\begin{aligned} L(\kappa,\nu) = &\sum_{i=1}^{n} \alpha^e A \left[\frac{\omega_i}{\kappa_i^j c_j} P_E^j + \frac{\beta s_i}{r_i^d} \left(\frac{1}{\eta} P_T^j + P_C^j \right) \right] \\ &+ \sum_{i=1}^{n} \alpha^t \left(\frac{\omega_i}{\kappa_i^j c_j} + \frac{s_i}{r_i^u} \right) + \sum_{j=0}^{1} \nu_j \left(\sum_{i \in o_j} \kappa_i^j - 1 \right) \end{aligned} \tag{4.12}$$

假设$\tilde{\kappa}, \tilde{\nu}$是任意满足KTT条件的点，基于KTT条件可以得到下面两个等式成立：

$$\begin{aligned} &\nabla f(\tilde{\kappa}_1^j, \tilde{\kappa}_2^j, \cdots, \tilde{\kappa}_n^j) + \sum_{j=0}^{1} \tilde{\nu}_j \nabla \left(\sum_{i \in o_j} \tilde{\kappa}_i^j - 1 \right) = 0 \\ &\sum_{i \in o_j} \tilde{\kappa}_i^j - 1 = 0 \end{aligned}$$

通过求解上面的式子，可以得到变量κ_i的最优解为：$\kappa_i^* = \tilde{\kappa}_i^j$，具体如下：

$$\kappa_i^* = \frac{\sqrt{\alpha^t \omega_i + \alpha^e A \omega_i P_E^j}}{\sum_{i \in o_j} \sqrt{\alpha^t \omega_i + \alpha^e A \omega_i P_E^j}} \tag{4.13}$$

由于$\sum_{i=1}^{m} f(\kappa) = \sum_{j=1}^{n} \sum_{i \in o_j} f(\kappa)$，将方程(4.13) 带入到到方程(4.10) 中，可以得到最优的$f(\kappa^*, \gamma^0)$值如下：

$$\begin{aligned} f &= \sum_{j=0}^{1} \sum_{i \in o_j} \left[\frac{\alpha^e A P_E^j \sqrt{\omega_i} \sum_{i \in o_j} (\sqrt{\omega_i} b_j)}{b_j c_j} + \frac{s_i}{r_j} \left(\frac{1}{\eta} P_T^j + P_C^j \right) + \frac{\alpha^t \sqrt{\omega_i} \sum_{i \in o_j} (\sqrt{\omega_i} b_j)}{b_j c_j} + \frac{s_i}{r_j} \right] \\ &= \sum_{j=0}^{1} \left[\frac{\alpha^e A P_E^j \left(\sum_{i \in o_j} \sqrt{\omega_i} \right)^2}{c_j} + \sum_{i \in o_j} \frac{\beta s_i}{r_j^d} \left(\frac{1}{\eta} P_T^j + P_C^j \right) + \frac{\alpha^t \left(\sum_{i \in o_j} \sqrt{\omega_i} \right)^2}{c_j} + \sum_{i \in o_j} \frac{s_i}{r_j^u} \right] \end{aligned} \tag{4.14}$$

其中$b_j = \sqrt{\alpha^t + \alpha^e A P_E^j}$ □

4.4.2 任务安置问题

通过上面的讨论，当给定$\gamma = \gamma^0 \in \mathbb{H}$，可以得到优化问题(4.9a)和(4.9b) 关于变量κ_i的最优解如方程(4.14)所示。在此基础上，优化问题(4.9a)和(4.9b) 关于γ 变成了$0-1$ 规划问题，如下所示：

$$\underset{\gamma}{\text{minimize}} \quad f(\kappa^*, \gamma) \tag{4.15a}$$

$$\text{subject to} \quad \sum_{i \in o_{\gamma_i}} \kappa_i^{*\gamma_i} \leq 1, \quad \gamma_i = 0, 1. \tag{4.15b}$$

在这一节中提出一个任务安置算法去解决此问题。

首先将优化问题(4.15a)和(4.15b)进行改写。定义$\mathbf{X} = (x_{ij})_{n\times 2}$，$x_{ij} \in \{0, 1\}$ 为需要求解的矩阵，其中x_{ij} 表示任务i是否被安置在边缘云j上：如果$x_{ij} = 0$，则任务i不会被安置在宏基站边缘云或small cell边缘云；如果$x_{ij} = 1$，则任务i会被安置在宏基站边缘云上或small cell边缘云上。于是可以看出，变量$\mathbf{X}$ 和变量γ的含义是相同的，都是任务的安置问题。此外，定义一些其他变量如下：

$$W = (\sqrt{\omega_1}, \sqrt{\omega_2}, \cdots, \sqrt{\omega_n}), \quad S = (s_1, s_2, \cdots, s_n)$$

$$\mathbf{C} = \begin{bmatrix} \sqrt{\alpha^e A P_E^0 c_0^{-1} + \alpha^t c_0^{-1}} & 0 \\ 0 & \sqrt{\alpha^e A P_E^1 c_1^{-1} + \alpha^t c_1^{-1}} \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} (\eta^{-1} P_T^0 + P_C^1) r_0^{-1} + r_0^{-1} \\ (\eta^{-1} P_T^1 + P_C^1) r_1^{-1} + r_1^{-1} \end{bmatrix}$$

因此，可以将关于变量γ优化问题(4.15a)和(4.15b)改写为关于变量X 的优化问题，如下所示：

$$\underset{\mathbf{X}}{\text{minimize}} \quad \|(W\mathbf{X}\mathbf{C})^T\|_2^2 + S\mathbf{X}\mathbf{R} \tag{4.16a}$$

$$\text{subject to} \quad x_{ij} \in \{0, 1\} \tag{4.16b}$$

$$\sum_{j=1}^{2} x_{ij} = 1, i = 1, 2, \cdots, n. \tag{4.16c}$$

其中优化的目标函数(4.16a)是减少任务延时和电网能量的消耗，约束条件(4.16b)和(4.16c) 是保证每个任务要么放在宏基站边缘云上，要么放在small cell

基站边缘云上。对于这个问题求解来说，假设需要处理的任务有n 个，则X_{ij} 有2^n个选择，根据文献[67]中的证明，此问题时一个NP难问题。

为了解决这个问题，首先定义连续的正变量y_{ij}, 其中y_{ij} 满足以下条件：

$$y_{ij} \geq 0, \quad \sum_{j=1}^{2} y_{ij} = 1, i = 1, 2, \cdots, n.$$

基于变量y_{ij}，定义线性函数$\phi(y_{ij}) = \frac{y_{ij}}{y_{ij}^{t-1}+\epsilon}$，其中$\epsilon$是一个非常小整数，$t$ 为迭代次数。参照Fazel等人[102]和Liu等人[103]的工作，利用线性函数$\phi(y_{ij})$去代替目标函数(4.16a)中的变量X，利用变量y_{ij}代替约束条件(4.16b)和(4.16c)中的变量X。类似于定理 4.1的证明，可以得到修改后的优化问题目标函数是凸函数，约束条件为线性函数，因此，修改后的问题是凸优化问题并且可以求解。具体的求解算法如算法 4.1所示。

算法 4.1: 任务安置算法

输入: 给定初始的任务安置, y_{ij}^0;

给定一个很小的误差限, δ;

输出: 在宏基站边缘云和small cell边缘云上的任务安置

1: $t := 0$。$y_{ij}^0 := 1 - \epsilon$，其中ϵ 是一个很小的正数;

2: $t := t + 1$;

3: 假设已经给定前面迭代的结果$\{y_{ij}^{t-1}\}$，定义$\phi_{ij}^t(y_{ij}) := \frac{y_{ij}}{y_{ij}^{t-1}+\epsilon}$，$i = 1, \cdots, n$，$j = 1, 2$;

4: 将上面得到的变量$\phi_{ij}^t(y_{ij})$代替优化问题(4.16a)，(4.16b)和(4.16c)中的变量X，求解优化问题可以得到$\{y_{ij}^t\}$;

5: 如果$|y_{ij}^t - y_{ij}^{t-1}| < \delta$，可以得到$y_{ij}^* = y_{ij}^t$；其他的话, 执行步骤2;

4.4.3 收敛性分析

在这一小节中，首先给出算法 4.1的收敛性分析，基于算法 4.1的步骤4，可以得到下式：

$$y_{ij}^t = \underset{y_{ij}}{\operatorname{argmin}} f(\frac{y_{ij}}{y_{ij}^{t-1} + \delta})$$

根据参考文献[104]中的全局收敛定理，得到算法 4.1是收敛的。然而，这里存在一个问题，即对于算法 4.1中修改后的优化问题的解为什么近似等于优化问题(4.16a)，

(4.16b)和(4.16c)的解。现在给出其说明，当两个优化问题近似相等时，也就是说，当$|y_{ij}^t - y_{ij}^{t-1}| < \delta$，变量$\phi_{ij}^t(y_{ij})$ 和变量X 是相等的。由于δ 是一个很小的正数，于是可以得到$y_{ij}^{t-1} \approx y_{ij}^t = y_{ij}^*$。因此，对于$\phi_{ij}^t(y_{ij}^*)$有如下等式成立。

$$\phi_{ij}^t(y_{ij}^*) = \frac{y_{ij}^*}{y_{ij}^{t-1} + \epsilon} \approx \begin{cases} 1 & \text{如果} y_{i,j} > 0 \\ 0 & \text{如果} y_{ij} = 0 \end{cases}$$

因此，可以得到变量$\phi_{ij}^t(y_{ij}^*)$ 近似等于变量X，即，修改后关于变量y_{ij} 优化问题的解近似等于原$0-1$ 整数规划问题的解。

其次给出由两个子问题求得的最优解也是原优化问题(4.9a)和(4.9b)的最优解。根据上面的讨论，原优化问题目标函数(4.9a) 可以表示为函数$f(\kappa, \gamma)$，其中κ和γ 是变量。当给定$\gamma = \gamma^0 \in \mathbb{H}$，根据4.4.1节的讨论，$f(\kappa, \gamma_0)$ 关于变量κ是凸的。因此, 存在$\kappa^* \in \mathbb{K}$，使得下面等式成立：

$$f(\kappa^*, \gamma^0) \leq f(\kappa, \gamma^0)$$

当给定$\kappa = \kappa^* \in \mathbb{K}$, 根据4.4.2节的讨论，$f(\kappa^*, \gamma)$ 关于变量γ 是一个$0-1$ 整数规划问题。根据算法 4.1，可以得到，存在$\gamma^* \in \mathbb{H}$，使得下面等式成立：

$$f(\kappa^*, \gamma^*) \leq f(\kappa^*, \gamma)$$

根据上面两个不等式，可以得到，对于任意的$\kappa \in \mathbb{K}$ 和$\gamma \in \mathbb{H}$，有下面等式成立：

$$f(\kappa^*, \gamma^*) \leq f(\kappa, \gamma^*) \leq f(\kappa, \gamma)$$

因此，可以通过求解两个子问题得到的$f(\kappa^*, \gamma^*)$，其中$f(\kappa^*, \gamma^*)$是原优化问题的最优解。

4.5 实验结果与分析

在这一节中将给出由可再生能源供电下任务卸载策略的仿真实验。实验结果分为以下三个部分：1）从计算延时和电网能量消耗两个角度与其他几种策略进行了对比；2）研究了计算量对任务卸载策略的影响；3）研究了数据量对任务卸载策略的影响。

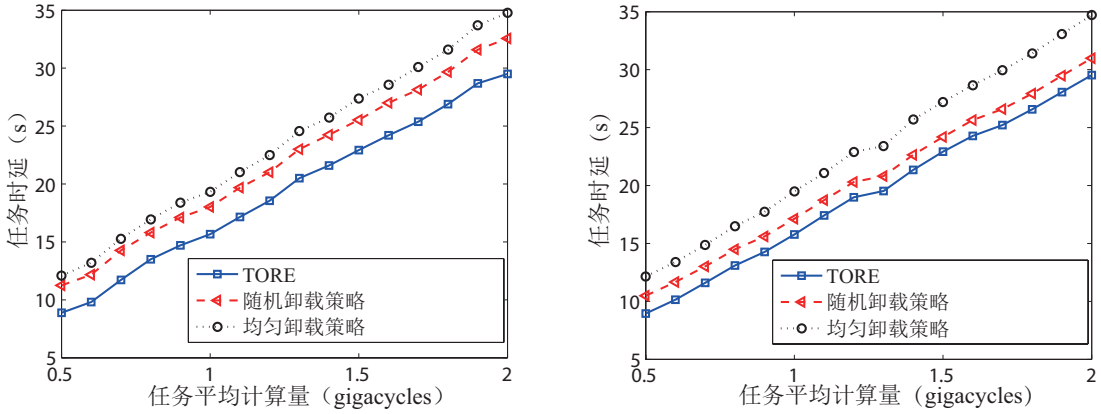
表 4.3 宏基站和small cell的仿真参数

参数	取值
宏基站的发射功率，P_T^0	20 W
宏基站的固定功率，P_C^0	130 W
small cell的发射功率，P_T^1	6.3 W
small cell的固定功率，P_C^1	56 W
宏基站的传输带宽，B_0	10 MHz
small cell的传输带宽，B_1	5 MHz
基站的高斯噪声，σ^2	10^{-7} W
信道功率增益，h	10^{-5}
宏基站的功率放大器效率因子，η_0	0.21
small cell的功率放大器效率因子，η_1	0.38
宏基站边缘云的计算资源消耗，P_E^0	90 W/Gigacycles
small cell边缘云的计算资源消耗，P_E^1	40 W/Gigacycles

4.5.1 仿真设置

本小节给出5G网络的设置、计算任务和可再生能量的设置。对于5G网络的参数设置，根据Zhang等人[96] 的工作，给出了基站的功率参数的设置，对于边缘云资源的处理任务时能量消耗的参数，根据Chen等人[82] 的工作，具体如表 4.3 所示。对于用户端的参数设置，根据Chen 等人[93] 的工作，给出了用户端的仿真参数，设置用户终端的传输带宽$B_m = 1$ MHz，用户端的发射功率$P_T^m = 0.2$ W，高斯信道噪声$\sigma^2 = 10^{-9}$ W，信道增益$h = 10^{-5}$。对于计算任务i，采用Tong等人[99]对计算任务计算量和数据量的设置，假设计算任务的计算量ω_i 和数据量s_i 均由概率分布产生。宏基站和small cell边缘云的计算资源分别为20吉赫（GHz）和10吉赫（GHz）。对于可再生能源的到达速率，根据Zhang 等人[96] 的研究工作，设置$\lambda_E = 40$J/s。

在本实验中，设置任务数据量处理前后的比例$\beta = 0.1$，并且讨论以下两种情形：1）$\alpha_t = 1(\alpha_e = 0)$，此种情形仅考虑任务时延；2）$\alpha_t = 1(\alpha_e = 0)$，此种情形仅考虑电网的能量消耗。



(a) 计算量：正态分布　　(b) 计算量：均匀分布

图 4.2 任务计算量为正态分布和均匀分布对任务时延的影响

4.5.2 对比卸载策略

本章提出的基于可再生能量的任务卸载策略(TORE)，与随机卸载策略(Random offloading scheme)和均匀卸载策略(Uniform offloading scheme) 进行比较。从任务时延和能量的消耗对这三种卸载策略进行评价。

随机卸载策略的设置包括两部分，（1）计算任务的安置问题：将计算任务将随机卸载到宏基站边缘云上或small cell 基站边缘云上。设置一个能够以相同概率产生0 和1 的随机生成器，按照随机生成器生成的数分配计算任务到宏基站边缘云上或small cell 基站边缘云上。（2）计算资源的分配问题：对于安置在宏基站边缘云或small cell 边缘云上计算任务，以最优化的分配策略分配计算资源。

均匀卸载策略的设置也包括两部分，（1）计算任务的安置问题：将计算任务均匀卸载到宏基站边缘云和small cell边缘云上。（2）计算资源的分配问题：对于安置在基站边缘云或small cell边缘云上的计算任务，仍以最优的分配策略分配计算任务。

4.5.3 计算量对任务卸载的影响

首先考虑计算量对计算任务卸载的影响。在实验中，和Tong等人[99]的设置一样，设置任务的数据量服从均值为5 兆字节（MB）的正态分布。设置任务的计算量服从以下三种分布：均匀分布（Uniform distribution），正态分布（Normal distribution）和帕累托分布（Pareto distribution）。

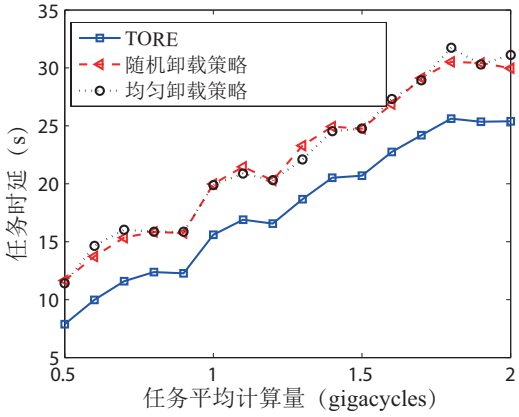
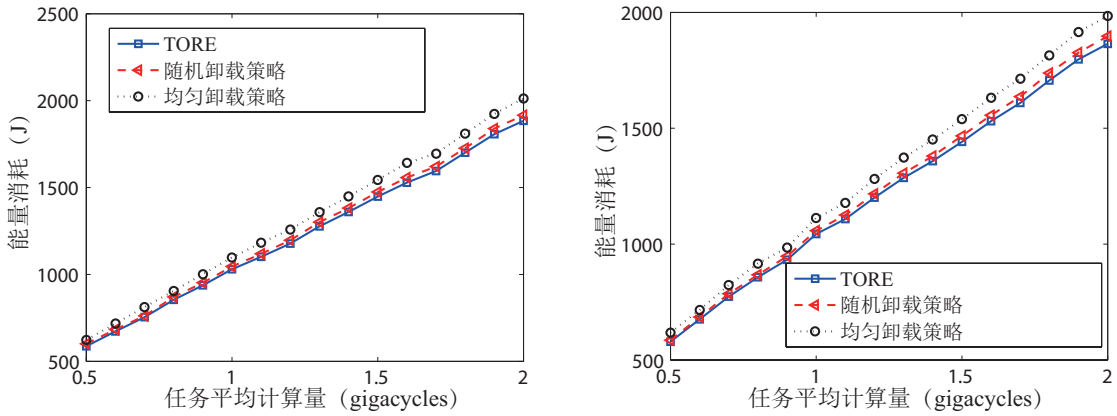


图 4.3 任务计算量为帕累托分布对时延的影响



(a) 计算量：正态分布　　(b) 计算量：均匀分布

图 4.4 任务计算量为正态分布和均匀分布对能量消耗的影响。

根据图4.2，图4.3，图4.4和图4.5可以得到，计算任务的计算量越大会导致越长的时间延时和消耗更多的电网能耗。这是因为任务的计算量越大，需要处理的时间越长，进而消耗的能量也越多。对比随机卸载策略和均匀卸载策略，可以看出，本章提出的任务卸载策略（TORE）表现出更短的时间延迟和更少的电网能量消耗。这是因为本章提出的计算卸载策略最大程度上考虑了可再生能源，并且对任务进行了合理的安置，对边缘云资源进行了优化的分配，从而最大程度上减少了计算延时，降低了电网供电的能量消耗。

根据图 4.2a和图 4.4a可以得到，当任务的计算量服从正态分布（Normal distribution）时，与随机卸载策略相比较，本章提出的任务卸载策略（TORE）能够减

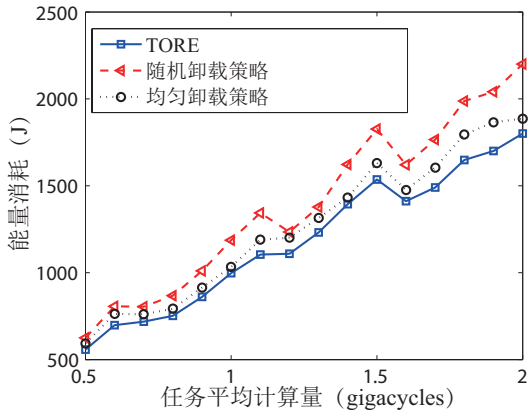
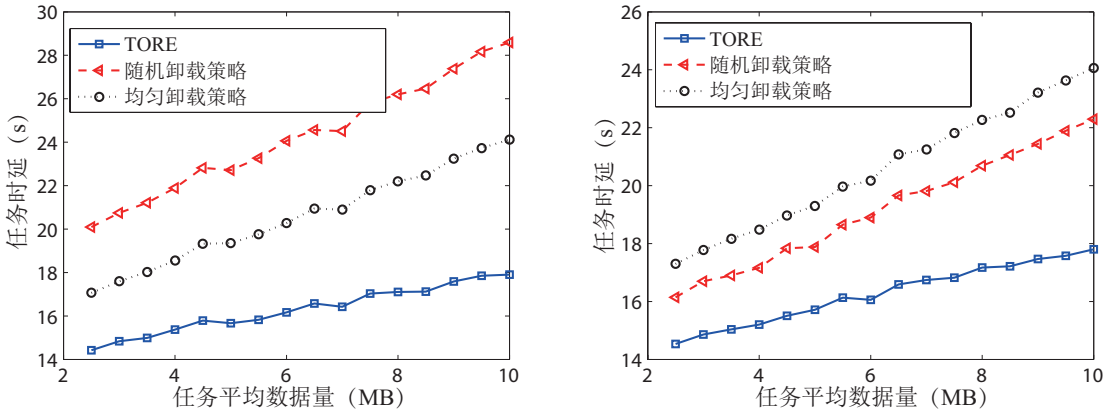


图 4.5 任务计算量为帕累托分布对能量消耗的影响



(a) 数据量：正态分布　　(b) 数据量：均匀分布

图 4.6 任务数据量为均匀分布和正态分布对时延的影响。

少30% 的任务延时和节省10% 的能量消耗。与均匀卸载策略相比较，本章提出的任务卸载策略（TORE）能够减少41%的任务延时和节省11%的能量消耗。

根据图 4.2b 和图 4.4b可以得到，当任务的数据量服从均匀分布（Uniform distribution）时，其结果和正态分布类似，本章提出的任务卸载策略能够在最大程度上减少任务的时延和节省能量的消耗。但是当计算量服从均匀分布时，虽然时延相差不大，但消耗了更少的电网能耗。

由图 4.3 和图 4.5 可知，当任务的计算量服从帕累托分布（Pareto distribution）时，其曲线走势没有正态分布和均匀分布平稳，这是因为帕累托分布具有长尾现象，并且在仿真帕累托分布时，设置了相当小的帕累托分布的间隔为0.1。对比于正态分

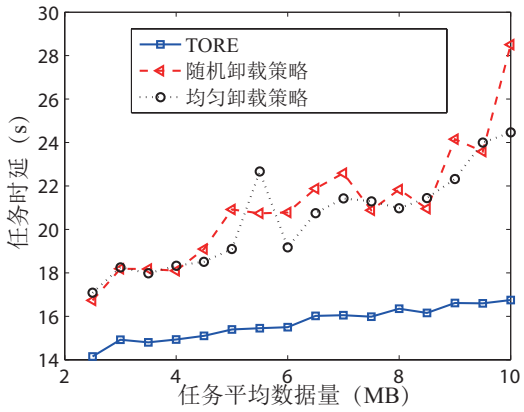


图 4.7 任务数据量为帕累托分布时对时延的影响。

布，当计算量服从帕累托分布时，和上面的结论类似，虽然时延相差不大，但消耗了更多的电网能耗。从上面的分析可以得出，任务间的计算量相差越大，会导致消耗更多的能耗，这是因为对于需要大计算量的任务，一般需要卸载到宏基站边缘云完成处理，从而导致较多的电网能量消耗。

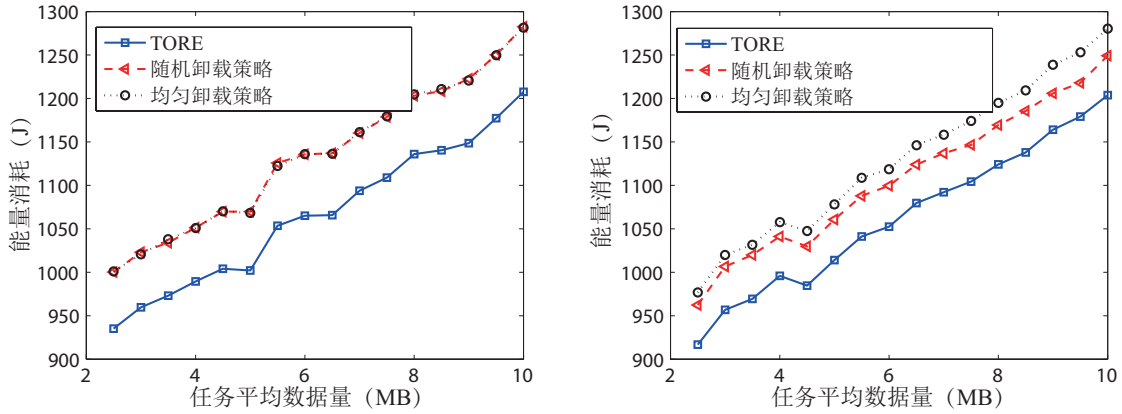
4.5.4 数据量对计算卸载的影响

这一小节考虑数据量对计算任务卸载的影响。和Tong等人[99]的设置一样，实验中设置任务的计算量服从均值为1千兆周（Gigacycle）的正态分布。设置任务的数据量服从以下三种分布：均匀分布（Uniform distribution），正态分布（Normal distribution）和帕累托分布（Pareto distribution）。

根据图4.6，图4.7，图4.8和图4.9可以得到，计算任务的数据量越大会导致时间延迟越长和电网的能量消耗越高。此外，还可以得到，对比随机卸载策略和均匀卸载策略，本章提出的任务卸载策略（TORE）表现出更短的时间延时和更少的电网能量消耗。

根据图 4.6a和图 4.8a可以得到，当任务的数据量服从正态分布（Normal distribution）时，和随机卸载方案相比较，本章提出的任务卸载策略（TORE）能够减少30% 的任务延时和节省5% 的能量消耗。和均匀卸载策略相比较，本章提出的任务卸载策略（TORE）能够减少28% 的任务延时和节省5%的能量消耗。

进一步，根据图 4.6b 和图 4.8b可以得到，当任务的数据量服从均匀分布（Uniform distribution）时，其结果和正态分布类似，本章提出的可再生能源卸载策略能够在最大程度上减少任务的时延和节省能量的消耗。但当计算量服从均匀分布



(a) 数据量：正态分布　　(b) 数据量：均匀分布

图 4.8 任务数据量对能量消耗的影响

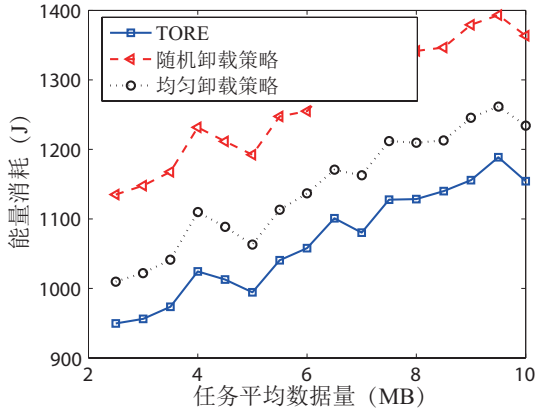


图 4.9 任务数据量为帕累托分布时对能量消耗的影响

时，虽然在能耗方面与正态分布相差不大，但其任务时延更少。

最后，根据图 4.7 和图 4.9 所示，当任务的数据量是服从帕累托分布（Pareto distribution）时，可以看出曲线走势没有正态分布和均匀分布平稳，这是由于帕累托分布具有长尾现象，并且在仿真帕累托分布时，设置了帕累托分布的间隔为0.5。与数据量服从正态分布相比，其任务时延更长。从上面的分析可以得出，任务间的数据量相差越大，会导致越长的计算时延。此外，从图中还可以可以得到：1）任务的计算量服从分布不同，产生的结果不同。2）任务数据量对时延和能量消耗的影响要低于计算量对其的影响。

4.6 本章小节

本章首先提出基于可再生能源供电下移动边缘云计算框架，其次基于可再生能量的分析，给出了用户任务延时和电网供电能耗的最小化问题，并将此问题转化为任务的安置和计算资源的分配两个子问题进行求解，从而提出了基于可再生能源供电的计算任务卸载策略，此策略不仅能够最大程度上利用可再生能源，而且能够减少任务的时延。最后给出了仿真实验，通过对比实验证明本章提出的策略能够至少减少20% 的任务延时，降低30%的能量消耗。

5 5G网络边缘计算卸载策略研究

考虑到基于D2D的边缘计算(如移动微云)模式的不可靠性，本章提出了一种新的计算任务卸载模式，即基于机会主义的移动自组微云计算卸载模式，同时分析了这种模式下的任务时延和能量消耗，并给出了计算任务如何在远端云，移动微云和基于机会主义的移动自组微云的卸载策略选择算法。最后实验验证了此模式要优于远端云和移动微云模式。

5.1 引言

随着移动设备和数据流量的爆炸性增长，以及移动设备的越来越智能化，大量的移动设备应用需要强大的计算能力来支撑。然而由于移动设备的计算能力、内存、存储、通信和电池容量有限，使得计算密集型任务很难在移动设备上进行处理。为了解决这个问题，研究者提出了移动云计算的概念[31,105]，即将计算量繁重的应用程序转移到云端进行处理，使得移动终端获得额外的计算与存储资源，并且降低了移动设备的能量消耗。对于计算量繁重的任务，用户可以通过蜂窝网（cellular network）或者WiFi将其卸载到远端云（remote cloud）处理。对于这两种卸载方式，WiFi 虽能够达到较高的数据传输率，但是在用户移动环境下，WiFi不能提供持续性的连接；而蜂窝网络能够支持任意时间和任何地点的无线接入，但是大量设备的连

表 5.1 几种计算任务卸载模式的比较

Structure	**通信方式**	**能耗**	**移动支持性**	**服务节点自由度**	**计算时延**
远端云	蜂窝网络	高	高	无	中
	WiFi	低	低	无	中
移动微云	D2D	低	低	低	高
移动	D2D	低	中	中	低
自组	D2D 和蜂窝网络	中	高	高	低
微云	D2D 和WiFi	低	高	高	低

接会导致蜂窝网络的负载过重，从而造成较低的数据传输速度。

近年来，设备到设备（D2D）通信作为同地区内不同设备间的一种直接的短程通信模式，已经在技术、应用和商业模式等方面得到了深入研究[106][107][108]。而且随着具有高存储和计算能力的移动设备的激增，Li等人[32]提出了一种新型的基于D2D的边缘计算卸载模式，称之为移动微云（mobile cloudlets）。在移动微云中，移动设备既可以作为计算服务的提供者（即服务节点），也可以作为计算服务请求者（即任务节点）。当移动微云的D2D连接可用时，任务节点可以将计算任务卸载到微云上进行处理。基于移动微云的计算可以使通信能耗变少、传输延迟变短，然而由于用户移动性使得D2D网络具有动态的特性，导致了移动微云计算的不可靠性。

综上所述，在用户移动环境下，远端云和移动微云模式在任务卸载方面各有利弊。如表5.1所示，远端云模式具有高能耗的缺点。移动微云模式对用户的移动性的支持度较低。为了解决这个问题，本章利用云端云和移动微云，提出了一种新的基于D2D的边缘计算任务卸载模型，即基于机会主义的移动自组微云模式（OCS）。具体来说，本章的主要贡献如下：

- 本章提出了一种新的基于D2D的计算任务卸载模式：基于机会主义的移动自组微云模式（OCS），此模式不仅能够支持用户的高移动性和保障任务时延的需求，而且能够尽可能多地节省系统能量的消耗。
- 本章基于移动微云和边缘云模式，结合移动性模型给出了基于机会主义的移动自组微云模式的时延和能耗分析，提出了时延和能耗最小的任务分配优化问题，继而给出了计算任务如何在远端云，移动微云和基于机会主义的移动自组微云的选择算法。
- 通过实验得出，本章提出的卸载模式在一定情形下要优于其他两种模式，并得到如下的卸载策略，给具有较高移动性和较大计算能力的服务节点分配的工作量越多，越能够减少计算和通信的能量消耗，进而提高移动自组微云模式的性能。

本章节组织如下。第5.2节描述了移动自组微云模式。第5.3节对此模式进行分析。第5.4节给出了卸载策略模型的建立与求解。第5.5节进行了实验，并分析结果。第5.6节对本章进行了小节。

5.2 移动自组微云模式描述

在本节中，首先给出了本章的研究动机，其次给出了移动自组微云模式的系统框架。

5.2.1 研究动机

群智感知作为一种新型的移动计算方式[78][109]，指的是针对劳动密集型或耗时性任务，考虑到周围存在大量的普通用户，可以将任务随机分配给普通用户，让他们协助你完成任务。群智感知已被应用于多个领域：包括定位，导航，城市交通的感知，市场预测，舆论挖掘等等。但是鲜有工作将群智感知应用于计算任务的卸载。基于此，本章就是利用群智感知的思想来完成用户本身设备没有办法完成的任务。

举例来说，假设用户需要处理一系列图片，图片本身的尺寸非常大，但是用户仅仅对某些特定领域感兴趣，比如对于整幅图像来说，用户仅仅对头像感兴趣。与整幅图片相比较，用户感兴趣区域的尺寸是非常小的。当用户在本地利用移动设备处理时，受电池容量、处理能力的限制，自己很难完成此任务。当通过蜂窝网络将图片传递到远端云时，传输这些图片可能导致较高的时延和能量消耗。当利用移动微云处理时，由于用户的移动性，导致了接触时间的有限性，从而降低了任务的完成的概率。也就是说，要保证任务的完成度，需要限制了用户的移动性。所以，设计一种既能支持用户的高移动性，又能保证时延和降低能耗的计算卸载模式是一个挑战性的问题。

5.2.2 移动自组微云模式的提出

考虑到用户之间有限的接触时间，那么假设两个节点碰面一次就能完成整个任务[109]或传递整个任务[87] 是与实际不符的。为了克服这个问题，本章假设进行D2D通信的任务节点和服务节点在接触时间内能够完成子任务的传递。当任务节点和服务节点在碰面完成离开后，服务节点会一直处理子任务，直到它完成子任务的处理。当子任务处理完成后，根据服务节点处理完子任务的位置，可以分为以下三种情形将子任务结果反馈给任务节点：（1）又一次进入了任务节点的D2D的通信范围内；（2）不在任务节点的D2D通信范围内，但是WiFi 是可行的；（3）既不在任务节点的D2D 通信范围内，WiFi 也是不可行的，但是在蜂窝网的覆盖之下，称这种服务模式为基于机会主义移动自组微云模式（OCS）。移动自组微云模式的一个重要特点就是服务节点和任务节点的接触时间可以长，也可以短。

具体来说，基于机会主义移动自组微云计算模式的任务反馈可以分为以下三种方式：

- D2D方式：在这种方式下，当服务节点在完成子任务的计算后，一旦服务节点和任务节点再次碰面，服务节点便将子任务的处理结果通过D2D的方式传递给计算节点。我们称这种计算任务反馈的方式为D2D方式。然而这种方式需要服务节点和任务节点至少能够碰面两次。
- WiFi方式：考虑到服务节点可以移动到了其他区域，而且此时WiFi是可用的。举例来说，在处理完子任务时，服务节点回到了家里。此时，子任务的结果可以通过WiFi的方式上传到云端，进而将结果传递给任务节点。
- 蜂窝网方式：在这种方式下，当服务节点移动到了一些没有WiFi的区域时，他只能通过蜂窝网络将计算结果上传到云端，进而传递给任务节点。

图 5.1给出了移动自组微云模式下三种方式的一个典型例子。David有一个本地手机无法处理的计算密集型的任务，在他的D2D 通信范围内，David有三个朋友Smith，Alex 和Bob均处于空闲状态。于是David将计算任务分为3个子任务，并且将子任务通过D2D 通信的方式传递给这三个用户。其中Smith是与David的一起玩耍的好朋友，他一直在David的D2D通信范围内，当他完成子任务处理后，他将子任务的处理结果直接通过D2D的方式传递给David。Alex因为有事回到家里，于是Alex 将结果通过WiFi的方式传递给David。而Bob 的移动性较高，在完成子任务处理前，他已经移动到了其他蜂窝，于是他通过蜂窝网的方式将结果传递给David。

根据上面的讨论，基于机会主义移动自组微云模式（OCS）在一些情形中是非常有效的。仍考虑图像分割的例子，与远端云模式相比较，移动自组微云模型可以通过D2D通信将整幅图片传递给服务节点，因此带来了较高的带宽和较少的能量。和移动微云相比较，移动自组微云模式有更高的扩展能力，这是因为他不要求服务节点和计算节点在一定时间内或一定区域内一直在接触，从而给了任务节点和服务节点更高的自由度。因此，移动自组微云可以看成远端云和移动微云的折中，能够实现更高的灵活性和能效的优化。据我们所知，本章是第一次提出机会主义移动自组微云模式。接下来，将分析此模式下的时延和能耗模型。

5.3 移动自组微云模式分析

本节从用户的移动性模型，计算任务的类型，计算任务的分配方案三个方面给出机会主义移动自组微云模型建立。表5.2 给出了本章所用的符号和默认值。

表 5.2 模型的变量和符号

变量名	默认值	解释
M	500	小区内的节点数
cn_i	N/A	有计算任务需要处理的节点（任务节点）i
sn_k	N/A	帮助任务节点处理任务的节点（服务节点）k,
N	45	小区内任务节点的总数目
n	10	每个任务节点包括的子任务数目
K	450	小区内所有子任务的数据
$X(t)$	N/A	t时刻服务节点的数目
$S_i(t)$	N/A	任务节点cn_i在t时刻分发出的子任务的数目
λ	0.0001	小区内任意两个节点平均碰面次数
$r_{t,t+\Delta t}(i)$	1/0	在Δt时间内任务节点cn_i是否成功分配子任务
$\theta^i_{t,t+\Delta t}(k)$	1/0	在Δt时间内任务节点cn_i是否成功将子任务分配给服务节点sn_k
t^*	N/A	完成非克隆任务的平均时间
t^*_s	N/A	完成可克隆任务的平均时间
Q	200	计算任务的总体大小
x_i	N/A	分配给服务节点sn_i任务量的大小
r	0.5	子任务处理后S^{result}_{sub-tk} 和处理前S^{recv}_{sub-tk}比例
$E^{cell}_{n\to c}$	2	通过蜂窝网方式从节点到云端的单位能量消耗
$E^{cell}_{c\to n}$	2	通过蜂窝网方式从云端到节点的单位能量消耗
E^{cloud}_{proc}	0.1	云端处理任务的单位能量消耗
E_{D2D}	1	D2D通信的单位能量消耗
$E^{node}_{proc}(k)$	0.2	在服务节点处理任务的单位能量消耗
ρ	0.001	探索服务节点的能量消耗
t_d	4000	计算任务完成的时间限制
ν_i	N/A	服务节点sn_i的平均处理速度
C_{Cloud}	N/A	在远端云的总体能量消耗
$C_{cloudlet}$	N/A	在移动微云的总体能量消耗
C_{OCS}	N/A	在移动自组微云的总体能量消耗
ω	0.5	计算时延和能量消耗的权重值

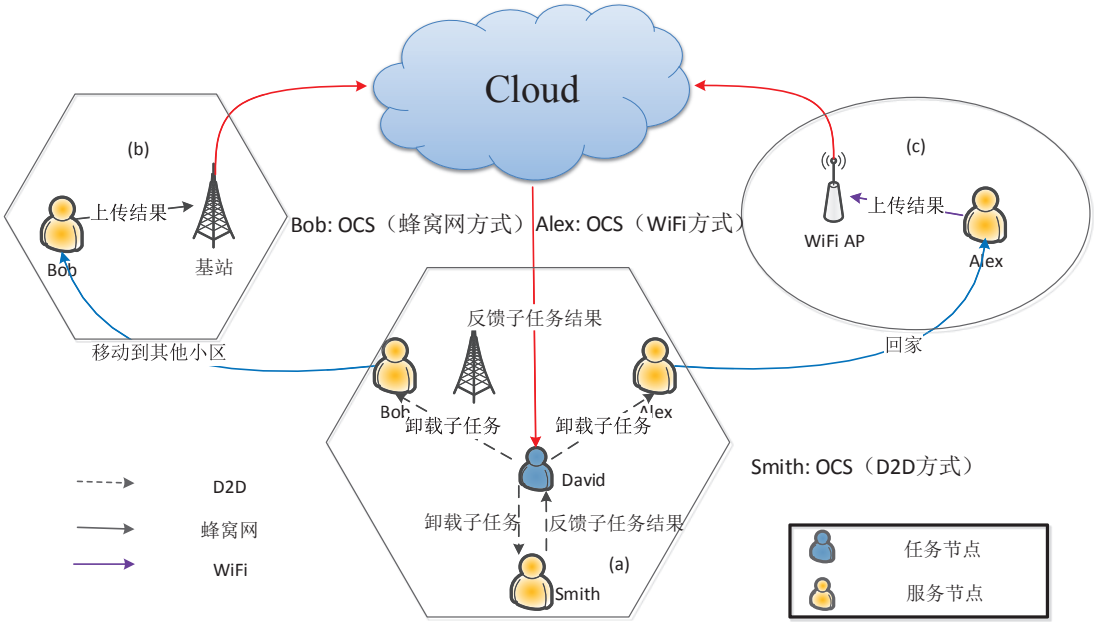


图 5.1 移动自组微云计算模式：(a) Smith 将子任务计算结果通过D2D的通信传递给David; (b) Bob通过蜂窝网将子计算结果上传到云端。(c) Alex通过WiFi将子任务结果上传到云端

5.3.1 用户移动模型

假设在移动云计算网络中有M个移动节点，其中有N个任务节点，每个任务节点包括的子任务数目为n。当且仅当服务节点在任务节点的通信半径R内时，任务节点和服务节点能够进行通信。也就是说，定义L_i和L_j分别表示任务节点cn_i和服务节点sn_j在时刻t的位置，当$||L_i(t) - L_j(t)|| < R$，两个节点可以进行通信。在本章中，假设节点的移动性是相互独立的。对于节点移动性的描述，参照Li等人[87]和Gao等人[62]的工作，假设任意两个节点的接触时间间隔服从参数为λ的指数分布，参数λ可以理解为两个节点的平均碰面次数。根据指数分布性质，于是得知任意两个节点在Δt时间内没有碰面的概率是：$P\{t > \Delta t\} = e^{-\lambda \Delta t}$。对于任务节点的任务描述如下：假设计算任务能够分为n个子任务，即每个任务节点的总计算量为Q，其能够分为n个子任务，每个子任务的任务量记为x_i，可以得出

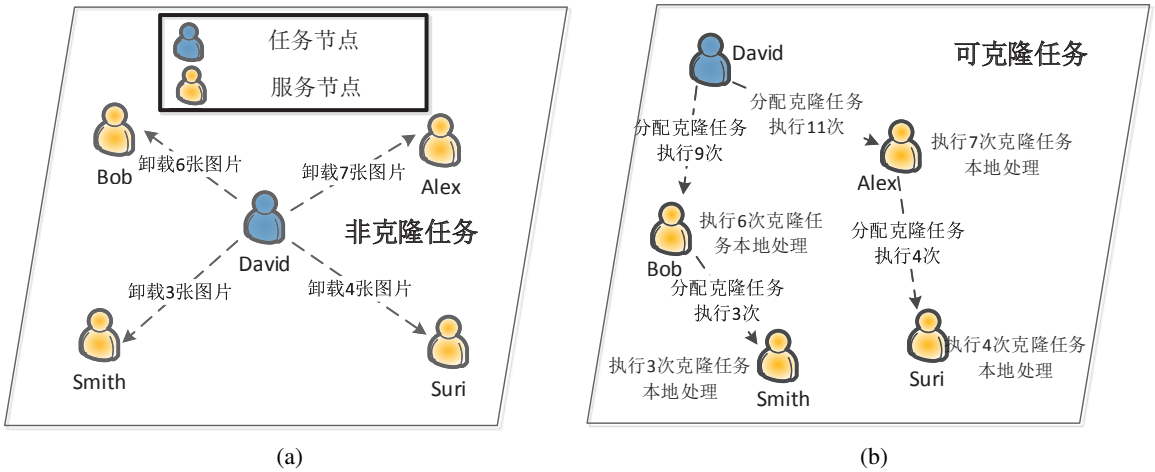
$$Q = \sum_{i=1}^{n} x_i \tag{5.1}$$

服务节点sn_i可以处理每个子任务，记其单位时间的处理速度为ν_i。

5.3.2 计算任务类型

对于计算任务来说，计算任务（简称为任务）一般由处理代码，数据和参数三部分构成。基于任务类型（比如应用、属性等）的不同，可将任务分为非克隆的任务（cloned task）和可克隆任务（non-cloned task）。对于非克隆任务，一个计算任务可以分为一定数量的子任务，每一个子任务是对处理代码、数据和参数的特定组合。举例来说，非克隆任务的两个子任务的数据和参数是不同的，但是其处理代码是相同的。所以当任务节点将非克隆子任务分配给服务节点时，服务节点只能再将其处理，由于任务的数据和参数的不同，而没有办法再将其分配。对于可克隆的任务，一般是随机参数的计算任务。当计算节点将任务分发给服务节点后，服务节点可以将其复制，再分发给其他服务节点。

为了进一步给出非克隆任务和可克隆任务的区别，下面给出了可克隆任务的具体特征，如下所示：(1) 当服务节点收到一个可克隆任务时，它可以复制可克隆任务，并将其分发给其他服务节点。本章假设服务节点只能接受一次可克隆任务，而且在传输过程中不考虑数据包的丢失。(2) 一个可克隆的任务一般仅包括处理的代码而不包括提前分配的参数和数据。当服务结点收到一个可克隆任务时，可以根据服务节点随机产生的参数去执行处理的代码。(3) 任务节点由于一般受限于其处理能力和电池容量，所以需要将执行多次的基于随机参数的处理代码分配给服务节点。(4) 尽管克隆任务的处理代码存在着高度冗余，但在每个服务节点的处理是参数不同，所以处理的结果也是不同的。



(a) (b)

图 5.2 计算任务的类型: (a) 非克隆任务；(b) 可克隆任务

接下来给出非克隆任务和可克隆任务的具体例子。图5.2a给出了一个非克隆任务的例子。David (任务节点) 有20 张图片需要处理，其中每幅图片的内容不同且需要提取特定的感兴趣的区域。在David的D2D通信范围内，有Bob，Alex，Smith和Suri 四个服务节点。由于这四个服务节点拥有不同设备，其计算任务处理能力也是不同的，所以David 将计算任务分为不相等的四份子任务，然后将四份子任务分别分发给四个服务节点。举例来说，Bob被分配了6幅图片，Alex 被分配了7幅图片，Smith 被分配了3幅图片，Suri 被分配了4幅图片。当服务节点将图片特定的感兴趣的区域分割完成后，服务需要将特定的感兴趣区域（任务结果）反馈给任务节点（David）。在这个例子中，Bob，Alex，Smith和Suri均是免费贡献自己的计算能力，即David免费使用这些服务节点的计算能力。事实上，服务节点本身的自私性是机会自组微云计算卸载的主要障碍，换句话说，大部分用户都想把自己的任务分发给其他用户，而不愿意接受别人分发的任务（分享自己闲置的计算资源）。这个事实可能导致移动自组微云任务卸载的失败。为了解决这个问题，需要对服务节点设置激励机制，对于这个例子来说，根据Bob，Alex，Smith和Suri处理的子任务量不同，分配不同的奖励。这些奖励可用于自己有任务需要处理时，给其他服务节点作为奖励，从而加快自己任务的处理速度。文献[26]给出了D2D计算任务卸载激励机制的讨论。然而，由于激励机制的设置不是本章关注的重点，因此不作详细讨论。

图5.2b展示了可克隆任务的场景。举例来说，David有一个可克隆的任务需要处理20次。在David的通信范围内，有Bob 和Alex 两个服务节点。考虑到服务节点处理能力的不同，David分别分配给Bob和Alex处理9次和11次。当Bob收到任务的分配后，他自己处理6 次，由于任务是可克隆的，他将剩余的3次处理分配给在他D2D通信范围的Smith进行处理。同理，Alex自己处理7 次，将剩余的4 次处理分配给了Suri。总体来说，需要处理20次的可克隆任务被分配给Bob处理了6次，分配给Alex处理了7次，分配给Smith处理了3 次，分配给Suri 处理了4 次。从这个例子可以看出，当服务节点收到可克隆任务后，可克隆任务可以被复制并且被分发到其他服务节点，这个类似于在线社交网络中的传染病模型。

5.3.3 任务分配方案

考虑到服务节点计算能力的异构处理能力，本章主要讨论两种子任务的分配方案：(1) 静态分配（static allocation）：通常情况下，任务节点并不知道服务节点的处理能力，因此可以假设所有服务节点的处理能力并没有差别，也就是说，服务节点

具有相同的处理速度（记为v_i），并且每个服务节点可被任务节点分配相同的子任务量$x_i = Q/n$。然而，这种假设的缺点是服务节点处理能力弱的，可能会导致任务总体时延变大。

(2)动态分配（dynamic allocation）：为了缩短计算任务时延，任务节点不应该忽略服务节点的异构性。动态分配方案考虑到了服务节点处理能力的异构性，任务节点能够以更加智能的方式分配子任务。比如当服务节点拥有较强的处理能力时，它将会被分配到更多的子任务量。

5.4 卸载策略模型的建立与求解

本节首先对移动自组微云计算模式的任务的时延（task duration）和能耗（energy cost）分析，其次给出了时延和能耗的优化问题，最后给出了计算任务如何在远端云，移动微云和基于机会主义的移动自组微云模式下的选择算法。

5.4.1 任务时延分析

本小节主要分析系统的任务时延。考虑到任务节点的任务是计算密集型任务，那么任务由本地处理的计算时延为Q/v，它将远远大于在移动自组网络处理的时延，其中v为任务节点的处理速度。具体来说，移动自组微云的计算任务时延包括三个方面：（1）子任务分发时延；（2）子任务的处理时延；（3）子任务结果的反馈时延。其中子任务的分发表示任务节点将子任务分配给服务节点，即将子任务通过D2D 的通信方式传递给服务节点。一般情况下，子任务处理和结果反馈的时延要远远小于子任务分发的时延，因此为了简单起见，本节仅考虑任务的分发时延。

非克隆任务的时延分析：首先给出非克隆任务的时延分析。定义Δt为一个能够满足两个节点一次接触的很小的时间间隔，如表5.2所示，如果$r_{t,t+\Delta t}(i)$等于1，这意味着任务节点m_i在Δt 时间内能够成功的将子任务分发给服务节点，反之亦然，因此，可以将$r_{t,t+\Delta t}(i)$表示为：

$$r_{t,t+\Delta t}(i) = \begin{cases} 1 & \Delta t\text{时间内任务节点}m_i\text{成功分配子任务}, \\ 0 & \text{其他情况}. \end{cases} \tag{5.2}$$

根据5.3.1小节的讨论，计算节点和服务节点的接触时间服从参数为λ的指数分布，那

么任务节点m_i将子任务成功分配的概率是：

$$P\{r_{t,t+\Delta t}(i)=1\}=1-(e^{-\lambda\Delta t})^{X(t)}. \tag{5.3}$$

其中$X(t)$为t时刻服务节点的数目。根据(5.3)，可以得到变量$r_{t,t+\Delta t}(i)$的数学期望为：$E(r_{t,t+\Delta t}(i))=1-(e^{-\lambda\Delta t})^{X(t)}$。进一步可以得到没有被分配子任务的服务节点的数目为：

$$X(t+\Delta t)=X(t)-\sum_{i=1}^{N}r_{t,t+\Delta t}(i). \tag{5.4}$$

对(5.4)两边同时求期望，于是可以得到：

$$E(X(t+\Delta t)=E(X(t))-NE(r_{t,t+\Delta t}(i)). \tag{5.5}$$

当Δt 趋于0时，基于极限理论，可以得到$E(X(t))$的导数如下所示：

$$E'(X(t))=\lim_{\Delta t\to 0}\frac{E(X(t+\Delta t)-E(X(t))}{\Delta t}=-N\lambda E(X(t)). \tag{5.6}$$

通过求解上面常微分方程（ODE）(5.6)，得出$E(X(t))$表达式如下：

$$E(X(t))=E(X(0))e^{-N\lambda t}. \tag{5.7}$$

通过对方程(5.7)的反函数进行求解，可以得到非克隆任务的平均时延（记为t^*）为：

$$t^*=\frac{\ln\frac{M-N}{E(X(t^*))}}{N\lambda}. \tag{5.8}$$

相应的，可以得到$E(X(t^*))=M-Nn$。

可克隆任务的时延分析：下面给出可克隆任务的时延分析。首先考虑系统中仅有一个任务节点，定义$S(t)$为时刻t拥有子任务的服务节点的数目，$\delta_{t,t+\Delta t}(k)$为在Δt时间内任务节点m_k是否完成了子任务的分配，和非克隆任务时延分析类似，可以得到

$$E(S(t))=\frac{S(0)Me^{M\lambda t}}{M-S(0)-S(0)e^{M\lambda t}}. \tag{5.9}$$

其中$S(0)=1$。于是可以得到可克隆任务的时延t为：

$$t=\frac{\ln(\frac{S(t)(M-S(0))}{S(0)(M-S(t))})}{M\lambda}. \tag{5.10}$$

其次，考虑有多个任务节点。假设系统中有N个任务节点，并且每个任务节点的任务都是可克隆任务。定义$S_i(t)$为t 时刻已经由任务节点m_i分配子任务的服务节点数目，于是可以得到$S_i(t+\Delta t)$，如下所示：

$$S_i(t+\Delta t) = S_i(t) + \sum_{k=1}^{M-NS_i(t)} \theta^i_{t,t+\Delta t}(k). \tag{5.11}$$

其中$\theta^i_{t,t+\Delta t}(k)$表示在$\Delta t$ 时间内任务节点是否将子任务分配给了服务节点，对于方程(5.11)，利用类似于方程(5.5)和方程(5.6)的方法，可以得到：

$$E'(S_i(t)) = (M - NE(S_i(t)))\lambda E(S_i(t)). \tag{5.12}$$

通过求解常微分方程（ODE）(5.12)，可以得到$E(S_i(t))$为：

$$E(S_i(t)) = \frac{e^{\lambda M t}M}{M - N + e^{\lambda M t}N}. \tag{5.13}$$

最后，通过对方程(5.13)的反函数求解，可以得到可克隆任务的平均处理时延(记为t^*_s)如下：

$$t^*_s = \frac{\ln\left(\frac{E(S_i(t^*_s))(M-N)}{M-NE(S_i(t^*_s))}\right)}{M\lambda}. \tag{5.14}$$

5.4.2 能耗最小的卸载策略

本小节将给出远端云，移动微云和移动自组微云三种模式下计算卸载的系统能量消耗分析。其中系统的能耗可分为三部分：（1）子任务分发消耗的能量；（2）子任务执行消耗的能量；（3）计算任务结果反馈消耗的能量。具体来说，考虑一个任务节点，其计算总量为Q，并且可分为n个子任务。对于节点的计算任务分配来说，由于静态分配可以看成动态分配的特殊情况，所以本节主要关注计算任务是动态分配的。由于能量的消耗和传输量有关，假设云端和服务节点接受到的子任务大小为S^{recv}_{sub-tk}，子任务处理后结果的大小为S^{result}_{sub-tk}，记r为S^{result}_{sub-tk} 和S^{recv}_{sub-tk}的比值。

远端云模式能耗分析：首先给出远端云模式下计算任务卸载的系统能量消耗，这里考虑的场景均是WiFi不可用的场景。根据上面的讨论，远端云的系统能耗主要包括任务卸载的能耗，任务处理的能耗和任务反馈的能耗，即：

$$\begin{aligned} C_{cloud} &= \sum_{i=1}^{n}(E^{cell}_{n\to c}x_i + E^{cloud}_{proc}x_i + rE^{cell}_{c\to n}x_i), \\ &= Q(E^{cell}_{n\to c} + E^{cloud}_{proc} + rE^{cell}_{c\to n}). \end{aligned} \tag{5.15}$$

其中$E_{n\to c}^{cell}$表示服务节点将任务处理完成后，将结果上传到远端云的能量消耗，$E_{c\to n}^{cell}$表示计算任务结果从远端云反馈给任务节点的能量消耗，E_{proc}^{cloud}表示云端处理的能量消耗。

移动微云模式能耗分析：其次给出移动微云模式下计算任务卸载的能量消耗。移动微云计算任务卸载消耗的能量主要包括D2D 的通信能量消耗，服务节点处理任务的能耗和周期性的检查周围服务节点的能量消耗。于是可以得到移动微云模式下系统的能量消耗为：

$$\begin{aligned} \mathrm{C}_{\text{cloudlet}} &= \sum_{i=1}^{n}(E_{D2D}x_i + E_{proc}^{node}(i)x_i + rE_{D2D}x_i) + M\rho t^*, \\ &= Q(1+r)E_{D2D} + \sum_{i=1}^{n} E_{proc}^{node}(i)x_i + M\rho t^*. \end{aligned} \tag{5.16}$$

其中E_{D2D}为任务节点和服务节点之间D2D通信的能量消耗，E_{proc}^{node}表示服务节点处理任务的能量消耗，ρ为探索服务节点的能量消耗，t^*表示任务完成的时间。

定义$\vec{X} = \{x_1, x_2, \ldots, x_n\}$为需要求解的移动微云计算模式下的任务分配方案，最小化其能量消耗的优化问题可以表示为：

$$\begin{aligned} \underset{\vec{X}}{\text{minimize}} \quad & \mathrm{C}_{\text{cloudlets}} \\ \text{subject to} \quad & \sum_{i=1}^{n} x_i = Q \\ & x_i \geq 0 \qquad i = 1, 2, \ldots, n. \end{aligned} \tag{5.17}$$

其中优化变量为任务节点的分配方案，优化目标为最小化系统能量消耗。约束条件为需要将任务处理完成。

移动自组微云模式能耗分析：最后给出移动自组微云下任务卸载能量消耗分析。考虑到服务节点的移动性，它可能移动到没有WiFi 的区域。在WiFi不可用的条件下，基于服务节点将计算任务结果反馈方式不同分为以下两种情形：（1）服务节点移动到任务节点的附近，并且可以进行D2D 通信，在这种情况下，D2D 通信可以用于任务结果的传递，对应上面提到的D2D 方式。（2）通过蜂窝网络将计算结果反馈是唯一的选择，这种方式称为蜂窝网方式。

首先给出服务节点sn_i和任务节点在T_d时间内碰面两次的概率值P_i的计算，其中T_d表示计算任务的最大时延。定义$t_{i,1}$ 为初始时刻到服务节点sn_i 第一次碰到任务节点的时间间隔，定义$t_{i,2}$ 为从第一次碰面后到第二次碰面的时间间隔。由于碰面的

时间间隔服从独立同分布的指数分布，于是可以得到在T_d时间内服务节点和任务节点碰面两次的概率为：

$$P_i = P(t_{i,1} + t_{i,2} \leq t_d) = \int_0^{t_i} P(t_{i,1} + t_{i,2} \leq t_d | t_{i,1} = x)\lambda e^{-\lambda x} dx. \tag{5.18}$$

由于$P(t_{i,1}+t_{i,2} \leq t_d|t_{i,1} = x) = P(t_{i,2} \leq t_d - x) = 1 - e^{-\lambda(t_d - x)}$，那么$P_i$可以表示为：

$$\begin{aligned} P_i = P(t_{i,1} + t_{i,2} \leq t_d) &= \int_0^{t_i} (1 - e^{-\lambda(t_d - x)})\lambda e^{-\lambda x} dx, \\ &= 1 - e^{-\lambda t_i} - \lambda t_i e^{-\lambda t_i}. \end{aligned} \tag{5.19}$$

其中$t_i = t_d - x_i/\nu_i$。于是，在机会主义的移动自组微云计算模式下任务卸载的能量消耗可以表示为：

$$C_{OCS} = \sum_{i=1}^{n} (x_i E_{D2D} + E_{proc}^{node}(i)x_i + r x_i P_i E_{D2D} + r x_i (1 - P_i)(E_{n\to c}^{cell} + E_{c\to n}^{cell})) + M\rho t^*. \tag{5.20}$$

和上面类似，定义$\vec{X} = \{x_1, x_2, \ldots, x_n\}$为需要求解的机会主义移动自组微云计算模式下的任务分配方案，因此关于能耗的优化问题可以表示为：

$$\begin{aligned} \underset{\vec{X}}{\text{minimize}} \quad & \mathrm{C_{OCS}} \\ \text{subject to} \quad & \sum_{i=1}^{n} x_i = Q \\ & x_i \geq 0 \qquad i = 1, 2, \ldots, n. \end{aligned} \tag{5.21}$$

其中优化的变量是具体的分配方案，优化的目标是最小化能量的消耗，约束的条件是需要将总的任务分配出去。由于此问题的目标函数是非线性函数，约束条件为线性约束，所以是一个非凸优化问题。对这个优化问题的求解分为两部分，首先优化变量P_i，其次优化移动自组微云的能量消耗，从而给出问题的解。

一般来说，通过蜂窝网传递任务处理结果消耗的能量要大于D2D的能量消耗。因此，考虑到不同情形下任务处理时延和能量消耗，针对特定的应用，给出如何在远端云、移动微云和移动自组微云模式下的选择算法 5.1，具体来说，(1) 远端云：如果计算任务所处的环境有稳定的WiFi，将其卸载到云端是一个不错的选择。(2) 移动微云：如果任务节点是能耗敏感型任务而且小区内的节点移动性比较弱，移动微云是一个比较好的选择。(3) 机会主义移动自组微云模式：如果r比较小，而且服务节点的自由度比较大，机会主义移动自组微云是一个好的选择。

算法 5.1: 模式选择算法

输入: S_{sub-tk}^{result} 和S_{sub-tk}^{recv}的比例: r;
用户设备之间平均碰面率：λ

输出: 远端云，移动微云和移动自组微云

if *用户所处环境有稳定的WiFi* **then**
  用户将任务通过WiFi上传到远端云处理;
end
if *λ较小并且任务是能耗敏感性* **then**
  用户通过移动微云处理计算任务;
end
if $r < 1$, *λ 较大并且用户具有较大的自由度* **then**
  用户通过OCS卸载计算任务;
end

5.4.3 时延和能耗最优的任务卸载策略

在本小节给出时间延迟和能量消耗的联合优化。考虑到不同的任务对时间时延和能量消耗的关注点不同，我们引入权重因子ω，表示对能量消耗和时间时延不同的关注度。最小化时延和能量的问题可以建立为下面的问题：

$$\begin{aligned} \underset{\vec{X}}{\text{minimize}} \quad & \mathrm{t}^* + \omega \cdot \mathrm{C_{OCS}} \\ \text{subject to} \quad & \sum_{i=1}^{n} x_i = Q \\ & x_i \geq 0 \qquad\qquad i = 1, 2, \ldots, n. \end{aligned} \tag{5.22}$$

其中目标函数是综合考虑任务处理的时延和能耗，约束条件是需要完成任务的处理。由于此问题是非凸的优化问题，本章利用启发式的遗传算法去解决上面的优化问题，通过启发式的搜索最优的x_i 可以求解此问题的有效解。

5.5 实验结果与分析

这一节中将给出机会主义移动自组微云模式下计算任务卸载的评估。首先给出实验参数的设置：基于Chen等人的工作[110]，设置两个节点单位时间内碰面的次数λ为0.00004到0.00032，设置小区总结点数M为300到3000，设置ρ 为0.001。其次给出

实验的分析。对于实验的分析，主要关注两个方面：可克隆任务和非克隆任务的时延分析；静态分配方案和动态分配方案的能耗分析。最后，给出本文提出模型的对比方案：比较了Chun等人[73]提出的远端云模式和Li 等人[32]提出移动微云模式。

5.5.1 任务时延分析

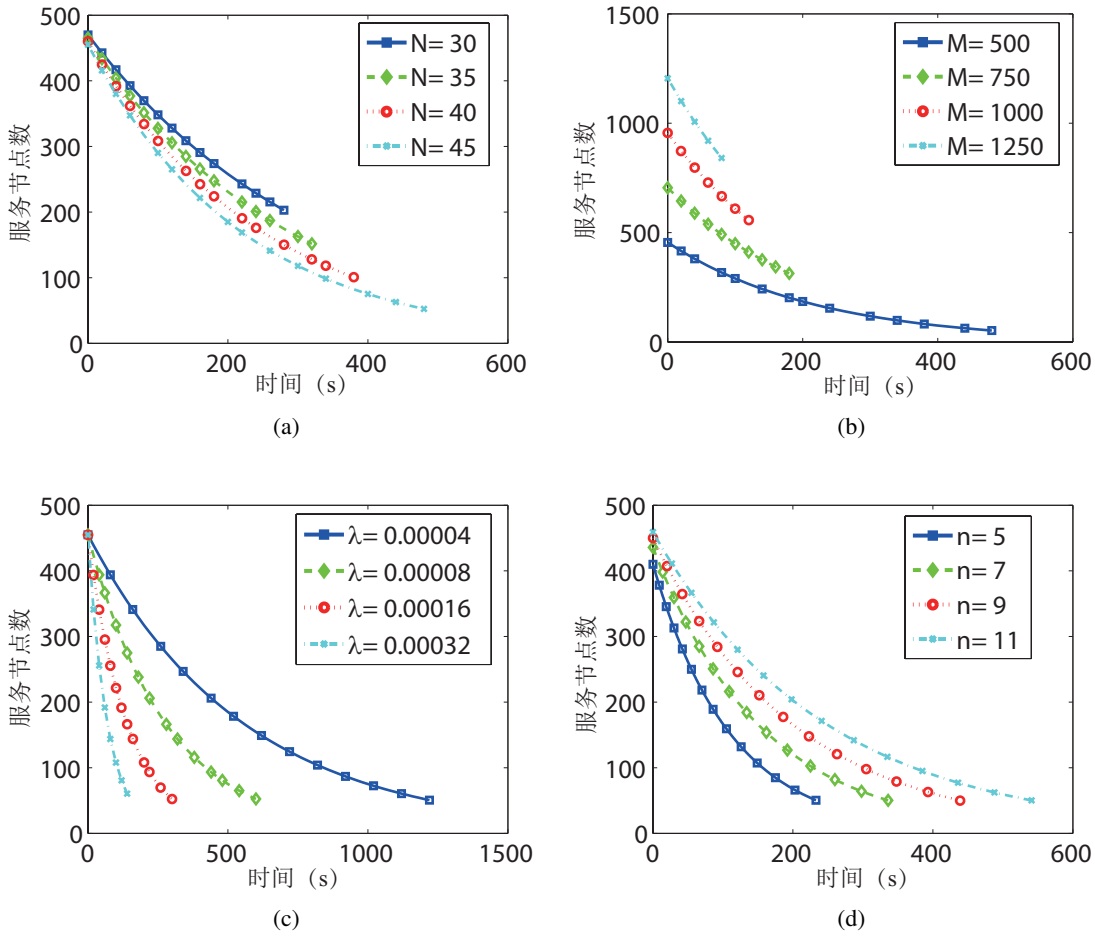


图 5.3 $X(t)$的评估：(a) N对$X(t)$的影响；(b) M对$X(t)$的影响；(c) λ对$X(t)$的影响；(d) n对$X(t)$ 的影响

非克隆任务时延分析： 由于t时刻服务节点的数目$X(t)$和小区内任务节点的总数N，小区内的总节点的数目M，小区内任意两个节点碰见碰面次数λ和子任务的个数n均有关，所以在实验中，我们将评估N，M，λ，n对$X(t)$的影响。

在图5.3a, 将小区总结点数M设置为500,任务包含的子任务数n设置为10，两个节

点单位时间碰面的次数λ 设置为0.0001，小区内任务节点的数目N取不同的值，分别为30，35，40 和45。这里注意到，任务的时延是任务节点将他们所有的子任务分配给那些没有任务的移动用户。从图 5.3a中可以看出，初始时刻服务节点的数目$X(0)$要比M 小，这是因为N个用户已经有计算任务，因此$X(0)$取值为$M-N$。

图5.3b中，将N设置为45，n设置为10，λ设置为0.00001，而变化M的取值，分别为500，750，1000，和1250。从图 5.3b中可以看出，在不同的M场景中，当$M=1250$时，任务完成的时间最短。这是因为小区内的节点数越多，在相同的时间内，对于有任务节点来说，他们有更多的机会碰到服务节点来卸载任务。相比之下，当$M=500$ 时，$M-N$ 个服务节点对任务卸载过程来说是不够的，从而导致了更长的任务时延。

在图5.3c中，将M设置为500，N设置为45，n设置为10，λ分别取为0.00004，0.00008，0.00016，和0.00032，以此来得到λ对t和$X(t)$的影响。如图 5.3c中所示，当λ取值越大时，表示用户的移动性越大，任务节点碰到计算节点的概率越大，从而使得子任务的分配越快，进而降低了用户的时延。因此，当λ减小时，任务时延会随之增加。

图5.3d中，设置M为500，λ设置为0.0001，小区内子任务总量K设置为450，n取不同的值，分别为5，7，9，11，因此N 为K/n。根据图 5.3d，可以得到：随着n的变大，任务时延就会增加。这是因为在k一定的前提下，n的增加预示着N 的减少，即任务节点的总数变少，从而使得时延增加。这表明，在固定M和λ取值的情况下，n越小，N越大能够越快的完成任务。此外，从 5.3d中还可以看出，初始时刻服务节点的数目$X(0)$不相等。这是因为当n 变化时，N也会变化。与图 5.3a类似，$X(0)$等于$M-N$。

图5.4a中，将M设置为500，K设置为450，变化λ 的取值，分别为0.00004，0.00008，0.00016和0.00032。如图5.4a所示，由于子任务的总量是固定的，随着N的增大和λ取值的增加，任务完成时间会不断减少。然而当N 达到40 以后，效果就不那么明显了。这就说明在小区内节点总数和子任务量一定的前提下，拥有任务的节点存在最优值，过少任务节点会导致每个任务节点的任务量过多，而过多的任务节点会导致服务节点数过少。从图5.4a 中还可以看出，当λ取值变大时，随着N不断增大而产生的时延减少的效果就越不明显。这是因为随着λ取值变大，λ时间时延的影响比重越来越大。

图5.4b中将λ设置为0.0001，将K设置为450，变化M的取值，分别取为500，

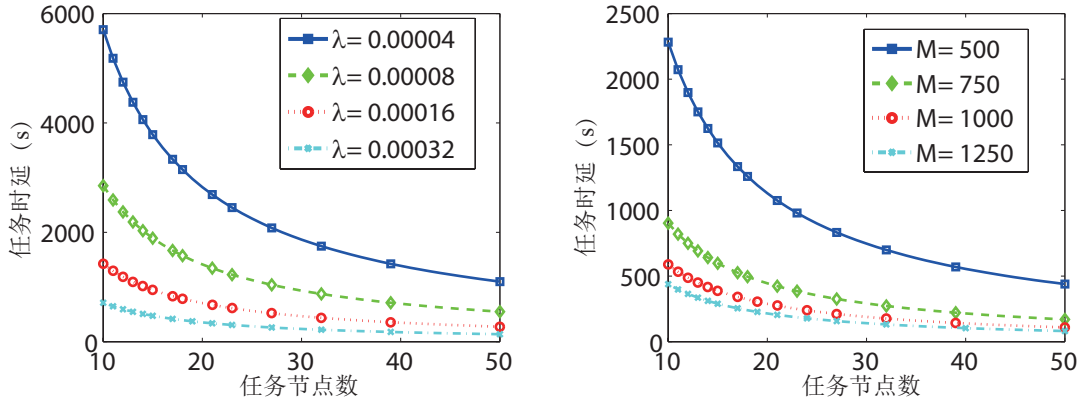


图 5.4 非克隆任务时延t^*：(a) 不同λ和N对t^*的影响；(b) 不同M和N对t^*的影响

750，1000和1250。从图5.4b 中可以看出，随着N的增大，任务完成时间会不断减少。从图5.4b中还可以看出，当M 达到1000的时候，增加N 所带来的时延减少是不明显的，这是因为当总任务量固定时，只需要一部分服务节点就可以完成任务的处理。

可克隆任务时延分析： 图5.5a中设置M为500，设置n为10，设置λ为0.0001，变化N的取值，分别取为30，35，40和45。从图5.5a 中可以得出，N对$S_i(t)$的影响是不明显的。这是因为在可克隆任务的前提下，每个服务节点都有可能转化为任务节点，所以N对其影响不明显。图5.5b中设置N为45，设置n为10，设置λ为0.0001，变化M 的取值，分别取为500，750，1000，1250。如图 5.5b中所示，M越大代表任务节点更有可能遇到任务节点。因此，当M 取为1250 时，$S_i(t)$ 增长最快，达到最大值10。图5.5c中，M的值设置为500，N的值设置为45，n的值设置为10，变化λ 的取值，分别取为0.00004，0.00008，0,00016，和0.00032。如图5.5c中所示，λ越大表示任务节点碰到服务节点的概率越大。因此，当λ取0.00032时，$S_i(t)$ 增长得最快，达到其最大值10。

在图5.6a中，将M设置为500，取K为450，变化λ的取值，分别取为0.00004，0.00008，0.00016和0.00032。在图5.6b中，将λ设定为0.0001，将K设定为450，变化M的取值，分别取为500，750，1000，1250。将图5.6与图 5.4进行对比，可以得到在相同的N值，λ值下，或相同的N 值，M值下，图5.6的任务时延t_s^* 比图5.4的任务时延t^*要小的多。这是因为与非克隆任务相比，可克隆任务允许任务的复制，也就是说当任务节点遇到服务节点时，这个服务节点就可能变成了任务节点。换句

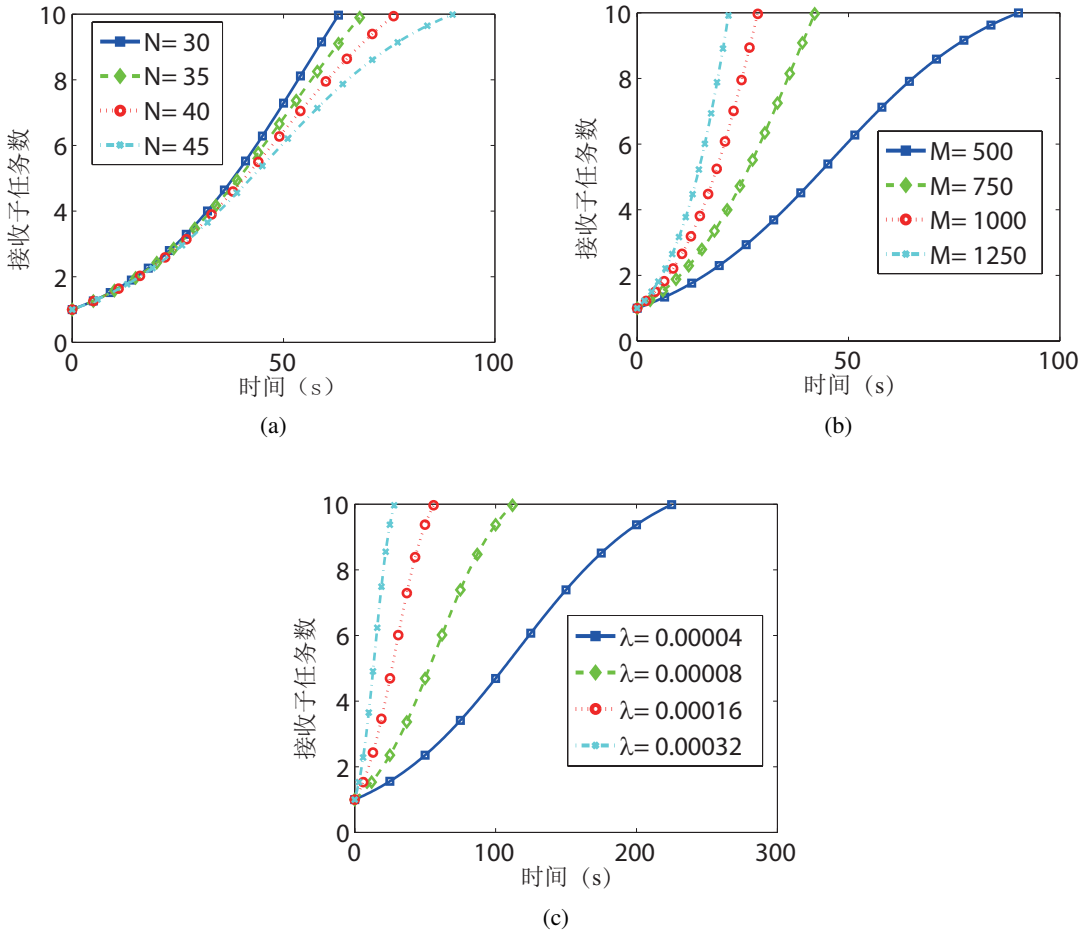
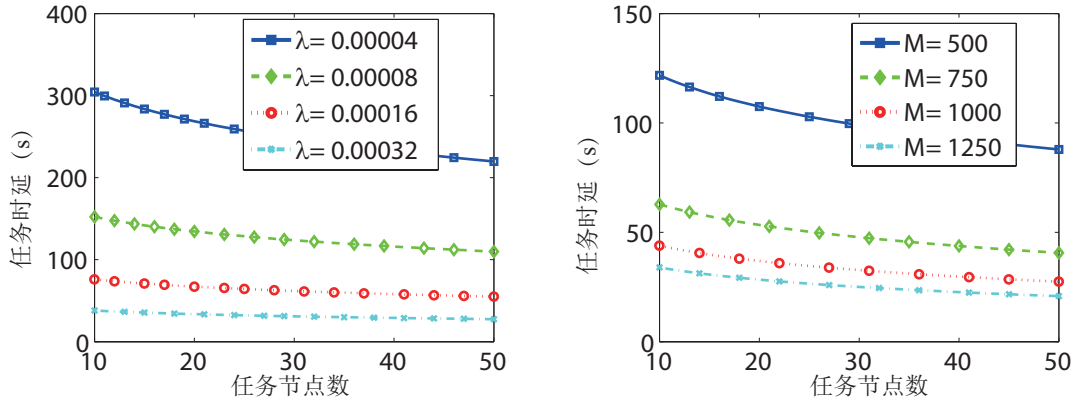


图 5.5 $S_i(t)$的评估：(a)N对$S_i(t)$的影响；(b)M对$S_i(t)$的影响；(c)λ对$S_i(t)$的影响

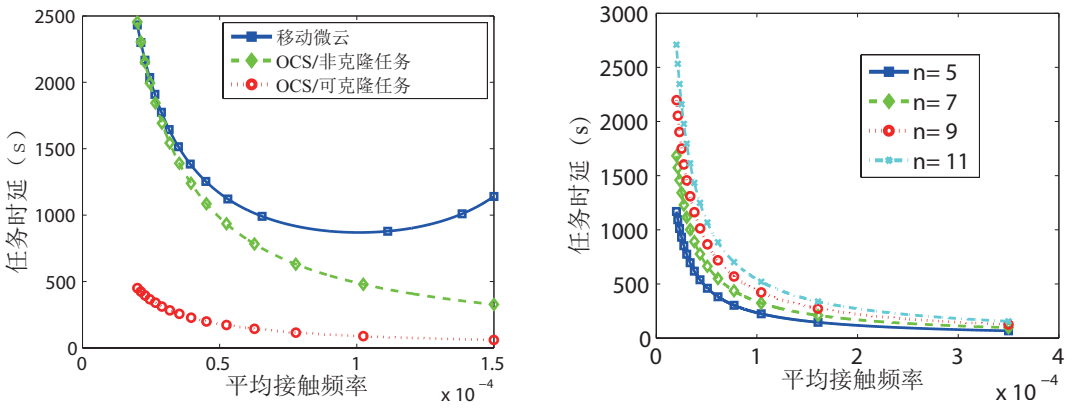
话说，如果任务是可克隆的，会导致任务节点的数目变大，从而减少了任务的时延。然而，非克隆任务的任务节点的数目是保持不变的。

计算任务的时延分析：图5.7a比较了移动自组微云计算模式和移动微云计算模式的任务时延分析。从图中可以看出，相同的λ 值下，移动自组微云计算模式下当任务可克隆时，任务时延是最短的，并且移动自组微云计算模式的任务时延比移动微云计算的任务时延要短一些。而且随着λ 的增大，移动自组微云计算模式呈现了较好的时延性能，这是因为随着λ 的增大，任务节点会更加频繁地遇到服务节点。对于移动微云计算模式，当λ 较小时（比如λ 从0.00002 到0.0001），其计算任务时延会逐渐减少，但是当λ 继续增大，移动微云的任务时延就会开始增加。这是因为随着任务节点和服务节点接触频率的增加，导致了接触时间的变短，从而导致子任务



(a) (b)

图 5.6 可克隆任务时延t_s^*：(a) 不同λ和N对t_s^*的影响；(b) 不同M和N对t_s^* 的影响



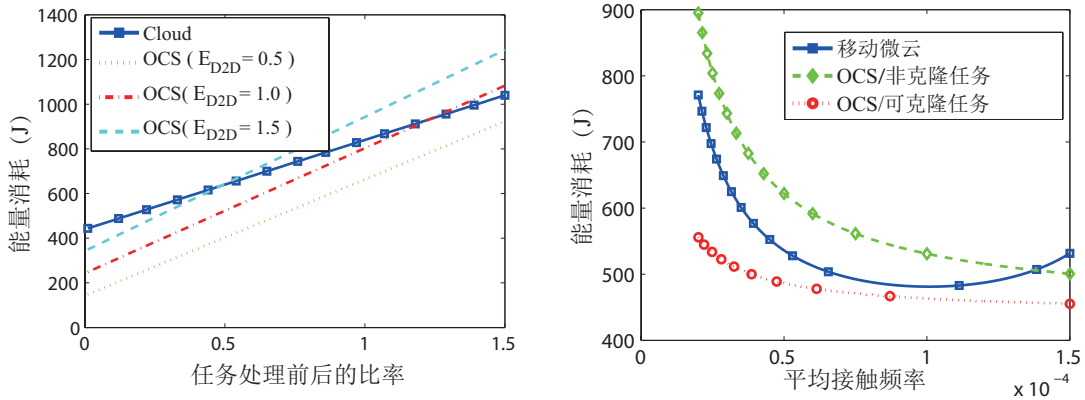
(a) (b)

图 5.7 计算任务时延分析：(a) 移动微云和移动自组微云模式下的任务时延分析；(b) 不同的n和不同的λ 对任务时延的影响

没有足够时间来实现任务结果的反馈。从图5.7b中可以得出，当λ 大于0.0003 时，移动自组微云计算模式的性能开始变得不明显。这是因为节点的接触时间太短，以致于无法保证子任务成功地卸载。

5.5.2 计算任务的能耗分析

图5.8a比较了远端云模式和移动自组微云模式下的能量消耗，其中四条直线分别代表了远端云模式下的能量消耗，以及移动自组微云模式下不同的E_{D2D}值下能量



(a) (b)

图 5.8 计算任务的能耗分析：(a) 远端云模式和移动自组微云模式的能耗分析；(b) 移动微云和移动自组微云模式的能耗分析

消耗。因为$E^{cell}_{n\to c}, E^{cell}_{c\to n} > E_{D2D}$，所以当$r < 1$时，移动自组微云模式的能量消耗要比远端云的能量消耗小一些。当$r > 1$时，随着r的增大，移动自组微云模式下的能量消耗也会增加，并且其增长速度比远端云的增长速度要快。此外，从图中我们还可以看出，当E_{D2D}增加时，移动自组微云模式的能耗也会随之变大。

图5.8b比较了移动微云模式和移动自组微云模式下的能量消耗。从图中可以得出，当λ值固定并且当计算任务可克隆（即可克隆任务）的情况下，移动自组微云的能耗比其他模式下的能耗要小。这是因为当任务可克隆时，移动自组微云能够更快的完成子任务计算处理。当$0.00002 \leq \lambda \leq 0.00014$时，并且任务不可克隆时（即非克隆任务），移动微云的能耗比移动自组微云的能耗小。这是因为当计算任务不能克隆时，移动自组微云可能需要将子任务结果上传到云端，而移动微云仅仅通过D2D通信，从而能够更加节省能耗。当λ继续增大时，导致了任务节点和服务节点的接触时间变短，从而导致在接触时间内无法完成子任务结果的反馈。因此，当$\lambda \geq 0.00014$时，并且任务是不可克隆时，移动自组微云模式要比移动微云模式好。

5.5.3 优化问题分析

这一小节考虑了静态和动态分配方案对实验结果的影响，并且对非克隆任务和可克隆任务都会进行评估。在能耗和任务时延方面，利用遗传算法来解决优化问题。在具体的实验中，任务时延和能耗的权重因子ω设置为0.5。

图5.9a比较了移动微云模式下静态分配和动态分配方案的能量消耗。从图中可

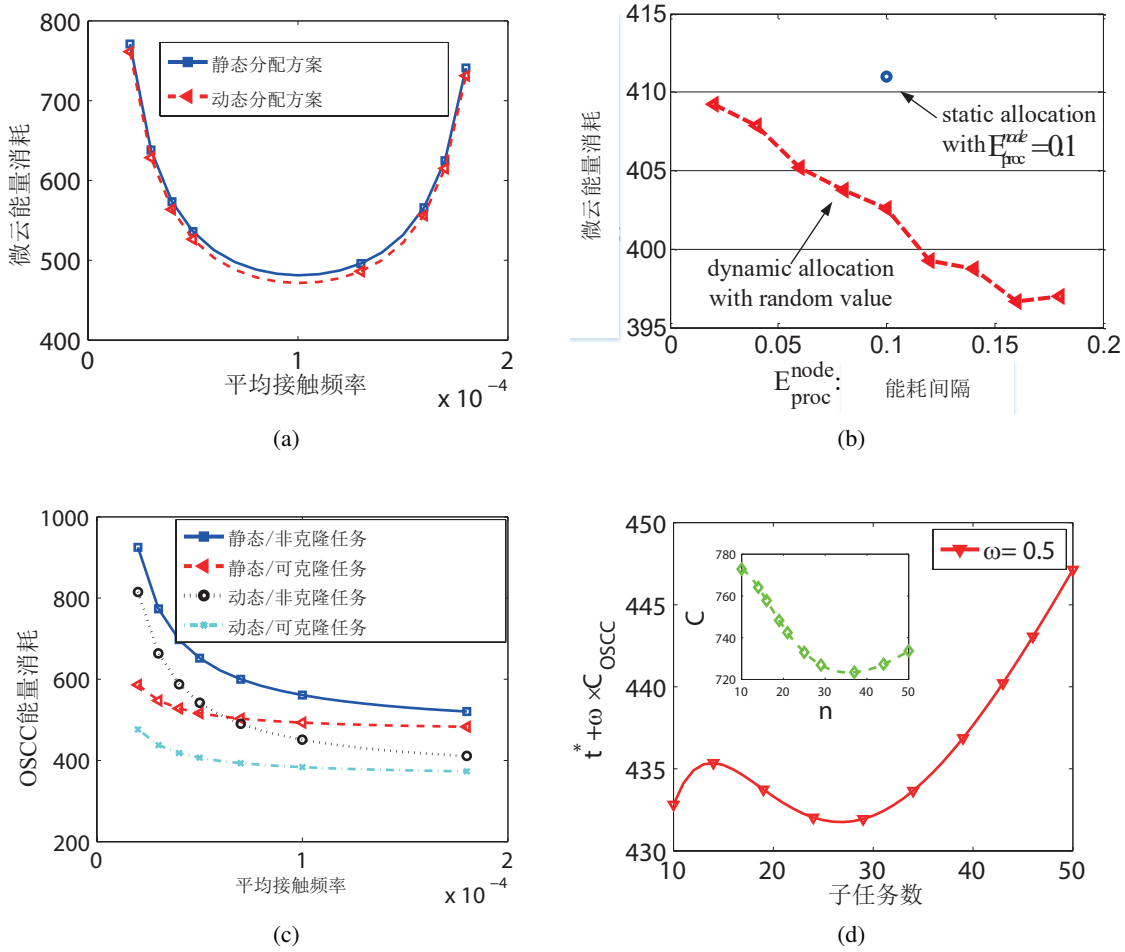


图 5.9 对于优化问题的评估：(a) 比较了移动微云模式下静态和动态分配方案的能量消耗；(b) 比较了移动微云模式下E_{proc}^{node} 对于其能量消耗的影响；(c) 比较了移动自组微云模式下静态和动态分配方案的能量消耗；(d) 联合最小化能量和时延的分析

以看出，动态分配方案的能耗比静态分配方案的能耗都要小一些，这是因为在动态分配的方案下，任务节点知道每个服务节点的处理能力，这样任务节点可以将计算量大的任务分发给处理能耗低的服务节点，从而降低能量的消耗。从图中还可以得出，当$\lambda < 0.00005$ 和$\lambda > 0.00017$ 时，动态分配的收益并不大。这是因为当$\lambda < 0.00005$时，用户的移动性比较低，当$\lambda > 0.00017$时，用户的移动性又比较高。这两种情况下用户的移动性均对能量的消耗影响比较大。

图5.9b给出了在静态分配和动态分配方案下，E_{proc}^{node}对能耗的影响。对于静态分配来说，其中蓝色的圆圈表示当E_{proc}^{node}固定为0.1时，即服务节点具有相同的处理能力时，静态分配的能量消耗情况。对于动态分配来说，为了描述动态分配，考虑

使用随机值来给出节点的处理能力。即X轴上的数据点的值表示平均值取0.1，方差为X轴上的点。举例来说，X 轴上的0.2 意味着E_{proc}^{node}的取值在0.01和0.19之间随机获得，X 轴上的0.01 意味着取值在0.09 到0.11之间变化。从图 5.9b 可以看出，计算节点处理能力的间隔越大，动态分配的性能越好。

图5.9c给出了移动自组微云模式下任务在可克隆和非克隆条件下，静态分配和动态分配方案下的能量消耗。从图5.9c 中可以看出，可克隆任务在动态分配方案下的能量消耗是最小的。随着λ的增大，四种情况下的能耗都是减少的。其中非克隆任务在动态分配方案下，其能耗减少地最快。这是因为λ的值对非克隆任务的影响比可克隆任务的影响大。当λ的取值达到0.00018 时，任务是否克隆对能耗的影响变得比较小。这是因为单位时间内的任务节点和服务节点的碰面时间增加，因此加速了子任务的分配，而任务是否克隆对其影响相应变得较小。

图5.9d展示了联合最小化任务时延和能量的消耗。在本实验中设置$\omega = 0.5$。通过图5.9d中嵌入的小图中可以看出，当$n < 35$ 时，随着子任务数目n的增加，能耗逐渐减少。这是因为当总任务Q固定时，随着子任务的增加，越来越多的子任务可以D2D通信时，分配给每个服务节点的子任务量就会变小，从而降低了能量的消耗。然而当子任务数目再增加时，需要更长时间来传递任务内容和定期地探索服务节点，因此任务时延会增加。当$n > 35$ 时，由于任务时延导致的周期性探索消耗的能量过多，从而增加了系统的能耗。因此这里存在着权衡，我们通过解决优化函数来减少时间和能耗。如图 5.9d 所示，利用遗传算法，当n等于26 时，达到了最优的性能。

5.6 本章小节

本章首先提出了基于机会主义的移动自组微云计算卸载模式。其次分析了此模式下计算任务的时延和能耗，给出了计算任务如何在远端云，移动微云和此模式下进行选择卸载，实验得出，本章提出的卸载模式在一定情形下要优于其他两种模式，并且能够在时延和能量方面具有较好的灵活性和更高的性能。

6 总结与展望

本章对5G移动边缘缓存与计算进行了总结，并对未来在缓存与计算方面的研究进行了展望。

6.1 研究工作总结

在5G超密蜂窝网络中，移动边缘缓存与计算能够克服回传链路容量有限的瓶颈，同时能够减少核心网的能量消耗。本文针对用户移动性、可再生能源到达的随机性以及移动边缘计算的不可靠性三大问题，提出了几种新颖的移动边缘缓存与计算卸载策略。本文的主要内容和研究成果如下：

（1）提出5G网络移动性缓存策略。基于用户移动性分析，本文将在small cell和用户设备上的缓存部署方案建模为0-1非线性规划问题，并证明该问题是一个NP难问题。进一步，利用子模态优化给出问题的近似解。通过实验验证了此策略比其他现有的缓存策略更有效，并得出当用户的移动性较低时，small cell 和移动设备应缓存流行的文件；当其移动性较高时，small cell和移动设备应注意缓存一些流行度较低的文件。

（2）提出5G网络绿色移动编码缓存策略。本文从编码缓存的角度出发，研究了在接触时间动态变化的条件下，small cell 和用户设备缓存内容的安置和传输问题。针对此问题，本章通过子模态优化给出了缓存文件的安置，进一步给出了small cell和移动设备的最优发射功率，从而得出网络的最小能量消耗。实验表明，与其他缓存策略进行比较，绿色移动编码缓存策略具有最高缓存命中率和最低传输能耗。同时得出当用户移动性较高时，缓存文件的small cell和用户设备消耗的传输能量较少；当用户移动性较低时，传输能量消耗比较多。

（3）提出可再生能量供电下的5G移动边缘云计算框架和卸载策略。通过对可再生能量进行分析，本文建立了用户时延和电网能耗最小化的优化问题，并将此问题分解为任务安置和计算资源分配两个子问题。通过对子问题的求解得到了此框架下最优的计算任务卸载策略（TORE），此策略既能保障用户任务的时延，又能最大程度的使用可再生能量，减少电网供电的能耗。实验结果表明，与随机计算卸载和均匀计算卸载策略相比，该策略至少能缩短20% 任务延时，降低30% 能量消耗。

（4）提出一种新颖的计算任务卸载模式并给出其卸载策略分析。本文提出基于机会主义的移动自组微云计算模式，基于用户移动性模型给出了此模式的时延和能耗分析，提出了任务时延和系统能耗最小化的任务卸载策略，继而给出了计算任务在远端云、移动微云和移动自组微云的选择算法。实验结果表明，本文提出的卸载模式在任务前后处理比例小于1、用户接触频率大于0.0014时，优于其他两种模式。同时得出：当给具有较高移动性和较强计算能力服务节点分配的任务越多，越能够降低任务计算和传输能量消耗，进而提高系统能效。

6.2 研究工作展望

结合本文对5G移动边缘缓存与计算的研究，未来主要以下几个方面继续展开研究工作：

（1）面向用户行为分析的缓存策略部署。本论文主要讨论了用户移动性对缓存策略的影响及带来的缓存增益，但是用户的移动性仅仅是用户习惯和行为的一个特征，对于用户的其他行为特征，比如用户对内容的喜好程度，现有的缓存策略都是基于文件流行性来描述的，但是文件流行性描述的是文件库中每个文件被所有用户请求的概率，并不能真实的反映用户的兴趣，如何根据用户习惯和行为对内容进行缓存仍是挑战性研究问题。比如可以通过对用户历史数据进行学习分析，得到基于用户兴趣度的缓存策略的部署。

（2）可再生能源量供电下的主动计算。本论文主要讨论了基于可再生能量供电下的移动边缘云计算，研究了可再生能量到达的随机性给计算卸载带来的挑战，但是，在实际应用中，存在可再生能源供给比较充足而计算任务相对较少的情况，如凌晨过后的基于风能供电的small cell 基站，造成了可再生能源的浪费。因此，如何通过用户卸载行为的分析，提前对用户的任务进行计算（比如在增强现实游戏中，根据用户的习惯，可以提前卸载场景的相关参数）是值得探讨的问题。比如，可以根据用户在边缘云已经卸载的任务来估计用户在未来可能需要的计算资源等，提前进行计算。

（3）基于边缘缓存与计算融合的研究。本文讨论了基于移动设备的边缘缓存和计算，但是没有考虑到缓存与计算相融合。有研究者对缓存与计算融合进行了初步研究，比如研究了绿色无线网络的通信与计算融合[111]。但是鲜有研究考虑移动边缘缓存与计算融合，因此也值得深入探讨。

致 谢

值此博士论文完成之际，回首博士求学历程，自己付出了很多，但同时也收获了丰硕的成果。

首先，感谢我的恩师陈敏教授。感谢您为我提供如此好的科研平台，让我有机会接触到科研的最前沿。感谢您在我科研遇到困难时，耐心地指导，您敏锐的洞察力总能把握问题的关键。感谢您经常带领大家一块出去玩，将生活和科研的热情结合在一起，您就像个大家长对待实验室的每个人。感谢您以身作则为我们树立了榜样，让我充满了学习的动力和激情。

其次，要感谢实验室的每一位成员。实验室就像一个大家庭，而我们就像是兄弟姐妹。感谢博士期间有你们陪伴，你们不仅在科研上给我支持，而且在生活上给予我很多照顾。感谢和我一起共同学习的博士生，他们是马玉军、王军锋、胡龙、史霄波、钱永峰、阳俊、李伟、周萍、缪一铭等，感谢你们在科研上对我的支持，和你们交流时思路更加清晰。感谢王露、刘梦宸、韩超、魏则如、徐意、卢佳毅等硕士生在课题研究、实验仿真中对我提供的帮助。感谢我选择加入嵌入与普适计算实验室，实验室的每个人都对科研充满激情，与你们交流总是有收获，每每都让我受益匪浅。

再次，感谢黄承明教授。感谢您不仅教会我读书，还教会我做人要踏实肯干。您的一言一行都是我学习的典范，感谢您给我们树立了榜样。值此毕业之际，向您表达我最衷心的感谢。

同时我要感谢我的父母，您们的理解和支持是我前进的动力！对自己多年一直求学在外很少回家表示愧疚！特感谢我的夫人，在我读博期间，几乎承担了所有的家务，还经常给我准备一顿大餐，让我感受到生活是如此的美好。

参考文献

[1] Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021. Technical report, 2017

[2] Andrews J G, Buzzi S, Choi W, et al. What will 5G be? IEEE Journal on Selected Areas in Communications, 2014, 32(6):1065–1082

[3] Ge X, Tu S, Mao G, et al. 5G Ultra-dense Cellular Networks. IEEE Wireless Communications, 2016, 23(1):72–79

[4] Ding M, Wang P, López-Pérez D, et al. Performance Impact of LoS and NLoS transmissions in Dense Cellular Networks. IEEE Transactions on Wireless Communications, 2016, 15(3):2365–2380

[5] Vondra M, Becvar Z. Distance-based Neighborhood Scanning for Handover Purposes in Network with Small Cells. IEEE Transactions on Vehicular Technology, 2016, 65(2):883–895

[6] Ge X, Yang B, Ye J, et al. Spatial Spectrum and Energy Efficiency of Random Cellular Networks. IEEE Transactions on Communications, 2015, 63(3):1019–1030

[7] Xia P, Jo H S, Andrews J G. Fundamentals of Inter-cell overhead Signaling in Heterogeneous Cellular Networks. IEEE Journal of Selected Topics in Signal Processing, 2012, 6(3):257–269

[8] Zhang N, Cheng N, Gamage A T, et al. Cloud assisted HetNets toward 5G Wireless Networks. IEEE Communications Magazine, 2015, 53(6):59–65

[9] Ge X, Cheng H, Guizani M, et al. 5G Wireless Backhaul Networks: Challenges and Research Advances. IEEE Network, 2014, 28(6):6–11

[10] Ismail M, Zhuang W, Serpedin E, et al. A Survey on Green Mobile Networking: From the Perspectives of Network Operators and Mobile Users. IEEE Communications Surveys & Tutorials, 2015, 17(3):1535–1556

[11] Bastug E, Bennis M, Debbah M. Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks. IEEE Communications Magazine, 2014, 52(8):82–89

[12] Liu A, Lau V K. Cache-enabled Opportunistic Cooperative MIMO for Video Streaming in Wireless Systems. IEEE Transactions on Signal Processing, 2014, 62(2):390–402

[13] Gregori M, Gómez-Vilardebó J, Matamoros J, et al. Wireless Content Caching for Small Cell and D2D Networks. IEEE Journal on Selected Areas in Communications, 2016, 34(5):1222–1234

[14] Liu A, Lau V K. Asymptotic Scaling Laws of Wireless ad hoc Network with Physical Layer Caching. IEEE Transactions on Wireless Communications, 2016, 15(3):1657–1664

[15] Mao Y, Luo Y, Zhang J, et al. Energy Harvesting Small Cell Networks: Feasibility, Deployment, and Operation. IEEE Communications Magazine, 2015, 53(6):94–101

[16] Touzri T, Ghorbel M B, Hamdaoui B, et al. Efficient Usage of Renewable Energy in Communication Systems Using Dynamic Spectrum Allocation and Collaborative Hybrid Powering. IEEE Transactions on Wireless Communications, 2016, 15(5):3327–3338

[17] Han T, Ansari N. Powering Mobile Networks with Green Energy. IEEE Wireless Communications, 2014, 21(1):90–96

[18] Marsan M A, Bucalo G, Caro A D, et al. Towards Zero Grid Electricity Networking: Powering BSs with Renewable Energy Sources. in: Proceedings of IEEE International Conference on Communications (ICC), 2013, 596-601

[19] Dhillon H S, Li Y, Nuggehalli P, et al. Fundamentals of Heterogeneous Cellular Networks with Energy Harvesting. IEEE Transactions on Wireless Communications, 2014, 13(5):2782–2797

[20] Huawei. Mobile Networks go Green. http://www1.huawei.com/enapp/198/hw-082734.htm

[21] Kim J M, Kim Y G, Chung S W. Stabilizing CPU Frequency and Voltage for Temperature-aware DVFS in mobile devices. IEEE Transactions on Computers, 2015, 64(1):286–292

[22] Liu S, Striegel A D. Exploring the Potential in Practice for Opportunistic Networks Amongst Smart Mobile Devices. in: Proceedings of the 19th Annual International Conference on Mobile Computing & Networking, 2013, 315–326

[23] Yu C H, Doppler K, Ribeiro C B, et al. Resource Sharing Optimization for Device-to-Device Communication Underlaying Cellular Networks. IEEE Transactions on Wireless communications, 2011, 10(8):2752–2763

[24] Asadi A, Wang Q, Mancuso V. A Survey on Device-to-Device Communication in Cellular Networks. IEEE Communications Surveys & Tutorials, 2014, 16(4):1801–1819

[25] Wu X, Tavildar S, Shakkottai S, et al. FlashLinQ: A synchronous Distributed Scheduler for Peer-to-Peer ad hoc Networks. IEEE/ACM Transactions on Networking, 2013, 21(4):1215–1228

[26] Pu L, Chen X, Xu J, et al. D2D Fogging: An Energy-efficient and Incentive-aware Task Offloading Framework via Network-assisted D2D Collaboration. IEEE Journal on Selected Areas in Communications, 2016, 34(12):3887–3901

[27] Poularakis K, Iosifidis G, Sourlas V, et al. Exploiting Caching and Multicast for 5G Wireless Networks. IEEE Transactions on Wireless Communications, 2016, 15(4):2995–3007

[28] Wang R, Peng X, Zhang J, et al. Mobility-aware Caching for Content-centric Wireless Networks: Modeling and Methodology. IEEE Communications Magazine, 2016, 54(8):77–83

[29] Karamshuk D, Boldrini C, Conti M, et al. Human Mobility Models for Opportunistic Networks. IEEE Communications Magazine, 2011, 49(12):157–165

[30] Liu Y, Lee M, Zheng Y. Adaptive Multi-resource Allocation for Loudlet-based Mobile Cloud Computing System. IEEE Transactions on Mobile Computing, 2016, 15(8):2398–2410

[31] Fernando N, Loke S W, Rahayu W. Mobile Cloud Computing: A survey. Future Generation Computer Systems, 2013, 29(1):84–106

[32] Li Y, Wang W. Can mobile cloudlets support mobile applications? in: Proceedings of the 33th IEEE International Conference on Computer Communications (INFOCOM), Toronto, Canada, 2014, 1060–1068

[33] Wang X, Chen M, Taleb T, et al. Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems. IEEE Communications Magazine, 2014, 52(2):131–139

[34] Golrezaei N, Dimakis A G, Molisch A F. Scaling Behavior for Device-to-Device Communications With Distributed Caching. IEEE Transactions on Information Theory, 2012, 60(7):4286–4298

[35] Sodagar I. The MPEG-DASH Standard for Multimedia Streaming over the Internet. IEEE MultiMedia, 2011, 18(4):62–67

[36] Maddah-Ali M A, Niesen U. Fundamental Limits of Caching. IEEE Transactions on Information Theory, 2014, 60(5):2856–2867

[37] Li J, Chen Y, Lin Z, et al. Distributed Caching for Data Dissemination in the Downlink of Heterogeneous Networks. IEEE Transactions on Communications, 2015, 63(10):3553–3568

[38] Liu A, Lau V. Exploiting Base Station Caching in MIMO Cellular Networks: Opportunistic Cooperation for Video Streaming. IEEE Transactions on Signal Processing, 2015, 63(1):57–69

[39] Maddah-Ali M A, Niesen U. Decentralized Coded Caching attains Order-optimal Memory-rate Tradeoff. IEEE/ACM Transactions on Networking, 2013, 23(4):1029–1040

[40] Shanmugam K, Golrezaei N, Dimakis A G, et al. Femtocaching: Wireless Content Delivery through Distributed Caching Helpers. IEEE Transactions on Information Theory, 2013, 59(12):8402–8413

[41] Breslau L, Cao P, Fan L, et al. Web caching and Zipf-like Distributions: Evidence and Implications. in: Proceedings of IEEE INFOCOM, 1999, 126–134

[42] Golrezaei N, Molisch A F, Dimakis A G, et al. Femtocaching and Device-to-Device Collaboration: A new Architecture for Wireless Video Distribution. IEEE Communications Magazine, 2013, 51(4):142–149

[43] Ahlehagh H, Dey S. Video-aware Scheduling and Caching in the Radio Access Network. IEEE/ACM Transactions on Networking, 2014, 22(5):1444–1462

[44] Szabo G, Huberman B A. Predicting the Popularity of Online Content. Communications of the ACM, 2010, 53(8):80–88

[45] Shi Y, Larson M, Hanjalic A. Collaborative Filtering beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges. ACM Computing Surveys, 2014, 47(1):1–45

[46] Song J, Song H, Choi W. Optimal Caching Placement of Caching System with Helpers. in: Proceedings of IEEE International Conference on Communications (ICC), London,UK, 2015, IEEE, 1825–1830

[47] Golrezaei N, Mansourifard P, Molisch A F, et al. Base-station assisted Device-to-Device Communications for High-throughput Wireless Video Networks. IEEE Transactions on Wireless Communications, 2014, 13(7):3665–3676

[48] Ji M, Caire G, Molisch A F. Wireless Device-to-Device Caching Networks: Basic Principles and System Performance. IEEE Journal on Selected Areas in Communications, 2016, 34(1):176–189

[49] Ji M, Caire G, Molisch A F. The Throughput-outage Tradeoff of Wireless One-hop Caching Networks. IEEE Transactions on Information Theory, 2015, 61(12):6833–6859

[50] Malak D, Al-Shalash M. Optimal Caching for Device-to-Device Content Distribution in 5G networks. in: Proceedings of IEEE Global Communications Conference (GLOBECOM). IEEE, 2014, 863–868

[51] Song C, Qu Z, Blumm N, et al. Limits of Predictability in Human Mobility. Science, 2010, 327(5968):1018–1021

[52] Rhee I, Shin M, Hong S, et al. On the Levy-walk Nature of Human Mobility. IEEE/ACM transactions on networking, 2011, 19(3):630–643

[53] Bettstetter C, Hartenstein H, Rez-Costa X. Stochastic Properties of the Random Waypoint Mobility Model: Modeling and Analysis of Wireless Networks (Guest Editors: Michela Meo and Teresa A. Dahlberg). Wireless Networks, 2004, 10(5):555–567

[54] Zhao M, Li Y, Wang W. Modeling and Analytical Study of Link Properties in Multihop Wireless Networks. IEEE Transactions on Communications, 2012, 60(2):445–455

[55] Poularakis K, Tassiulas L. Code, Cache and Deliver on the Move: A Novel Caching Paradigm in Hyper-dense Small-cell Networks. IEEE Transactions on Mobile Computing, 2017, 16(3):675–687

[56] Taghizadeh M, Micinski K, Biswas S, et al. Distributed Cooperative Caching in Social Wireless Networks. IEEE Transactions on Mobile Computing, 2013, 12(6):1037–1053

[57] Xiao M, Wu J, Huang L, et al. Multi-task Assignment for Crowdsensing in Mobile Social Networks. in: Proceedings of the 33th IEEE International Conference on Computer Communications (INFOCOM), Toronto, Canada, 2015, 2227–2235

[58] Wang D, Pedreschi D, Song C, et al. Human Mobility, Social Ties, and link prediction. in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Miningt, San Diego, USA, 2011, 1100-1108

[59] Wang T, Song L, Han Z. Dynamic Femtocaching for Mobile Users. in: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), 2015, 861–865

[60] Poularakis K, Tassiulas L. Exploiting User Mobility for Wireless Content Delivery. in: Proceedings of IEEE International Symposium on Information Theory Proceedings (ISIT), 2013, 1017–1021

[61] Li Y, Jiang Y, Jin D, et al. Energy-efficient Optimal Opportunistic Forwarding for Delay-tolerant Networks. IEEE Transactions on Vehicular Technology, 2010, 59(9):4500–4512

[62] Gao W, Cao G. User-centric Data Dissemination in Disruption Tolerant Networks. in: Proceedings of the 32th IEEE International Conference on Computer Communications (INFOCOM), Shanghai, China, 2011, 3119–3127

[63] Wang Z, Shah-Mansouri H, Wong V. How to Download More Data from Neighbors? A Metric for D2D Data Offloading Opportunity. IEEE Transactions on Mobile Computing, 2017, 16(6):1658–1675

[64] Wang R, Zhang J, Song S, et al. Mobility-aware Caching in D2D Networks. arXiv preprint arXiv:1606.05282, 2016

[65] Lu Z, Sun X, La Porta T. Cooperative Data Offloading in Opportunistic Mobile Networks. in: Proceedings of the 35th IEEE International Conference on Computer Communications (INFOCOM), San Francisco, USA, 2016, IEEE, 1–9

[66] González M C, Hidalgo C A. Understanding Individual Human mobility Patterns. Nature, 2008, 453(7196):779

[67] Lan R, Wang W, Huang A, et al. Device-to-device Offloading with Proactive Caching in Mobile Cellular Networks. in: Proceedings of IEEE Global Communications Conference, San Diego, USA, 2015, 1–6

[68] Zhou L. Specific-Versus Diverse-Computing in Media Cloud. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(12):1888–1899

[69] Kim W J, Kang D K, Kim S H, et al. Cost Adaptive VM Management for Scientific Workflow Application in Mobile Cloud. Mobile Networks and Applications, 2015, 20(3):328–336

[70] Vasile M A, Pop F, Tutueanu R I, et al. Resource-aware Hybrid Scheduling Algorithm in Heterogeneous Distributed Computing. Future Generation Computer Systems, 2015, 51:61–77

[71] Chandrasekhar V, Andrews J G, Gatherer A. Femtocell Networks: a Survey. IEEE Commun Magzine, 2008, 46(9):59–67

[72] Lei L, Zhong Z, Lin C, et al. Operator Controlled Device-to-Device Communications in LTE-advanced Networks. IEEE Wireless Communications, 2012, 19(3):96–104

[73] Chun B G, Ihm S, Maniatis P, et al. Clonecloud: Elastic Execution between Mobile Device and Cloud. in: Proceedings of the 6th Conference on Computer Systems, Salzburg, Austria, 2011, 301–314

[74] Kosta S, Aucinas A, Hui P, et al. Thinkair: Dynamic Resource Allocation and Parallel Execution in the Cloud for Mobile Code Offloading. in: Proceedings of the 32th IEEE International Conference on Computer Communications (INFOCOM), Orlando, USA, 2012, 945–953

[75] Taleb T, Ksentini A. Follow me Cloud: Interworking Federated Clouds and Distributed Mobile Networks. IEEE Network, 2013, 27(5):12–19

[76] Flores H, Srirama S. Mobile Code offloading: Should it be a Local Decision or Global Inference? in: Proceedings of the 11th annual international conference on Mobile systems, applications, and services, Taipei,Taiwan, 2013, 539–540

[77] Barbera M, Kosta S, Mei A, et al. To offload or not to offload? The bandwidth and energy costs of mobile cloud computing. in: Proceedings of the 32th IEEE International Conference on Computer Communications (INFOCOM), Turin,Italy, 2013, 1285–1293

[78] Flores H, Hui P, Tarkoma S, et al. Mobile Code Offloading: From Concept to Practice and Beyond. IEEE Communications Magazine, 2015, 53(3):80–88

[79] Jia M, Cao J, Yang L. Heuristic Offloading of Concurrent Tasks for Computation-intensive Applications in Mobile Cloud Computing. in: Proceedings of the 33th IEEE International Conference on Computer Communications (INFOCOM), Toronto, Canada, 2014, 352–357

[80] Lei L, Zhong Z, Zheng K, et al. Challenges on Wireless Heterogeneous Networks for Mobile Cloud Computing. IEEE Wireless Communications, 2013, 20(3):34–44

[81] Patel M, Naughton B, Chan C, et al. Mobile-edge computing introductory technical white paper. White Paper, Mobile-edge Computing (MEC) industry initiative, 2014

[82] Chen X, Jiao L, Li W, et al. Efficient Multi-user Computation Offloading for Mobile-edge Cloud Computing. IEEE/ACM Transactions on Networking, 2016, 24(5):2795–2808

[83] Chen M, Hao Y, Qiu M, et al. Mobility-Aware Caching and Computation Offloading in 5G Ultra-Dense Cellular Networks. Sensors, 2016, 16(7):974

[84] Satyanarayanan M, Bahl P, Caceres R, et al. The Case for Vm-based Cloudlets in Mobile Computing. IEEE Pervasive Computing, 2009, 8(4):14–23

[85] Miettinen A P, Nurminen J K. Energy Efficiency of Mobile Clients in Cloud Computing. in: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, Boston, MA, 2010

[86] Sardellitti S, Scutari G, Barbarossa S. Joint Optimization of Radio and Computational Resources for Multicell Mobile-edge Computing. IEEE Transactions on Signal and Information Processing over Networks, 2015, 1(2):89–103

[87] Wang C, Li Y, Jin D. Mobility-assisted Opportunistic Computation Offloading. IEEE Communications Letters, 2014, 18(10):1779–1782

[88] Calinescu G, Chekuri C, Pál M, et al. Maximizing a Monotone Submodular Function subject to a Matroid Constraint. SIAM Journal on Computing, 2011, 40(6):1740–1766

[89] Passarella A, Conti M. Analysis of Individual Pair and Aggregate Intercontact Times in Heterogeneous Opportunistic Networks. IEEE Transactions on Mobile Computing, 2013, 12(12):2483–2495

[90] Liu D, Yang C. Energy Efficiency of Downlink Networks with Caching at Base Stations. IEEE Journal on Selected Areas in Communications, 2016, 34(4):907–922

[91] MacKay D J. Fountain Codes. IEE Proceedings-Communications, 2005, 152(6):1062–1068

[92] Zhao J, Zhuo X, Li Q, et al. Contact Duration Aware Data Replication in DTNs with Licensed and Unlicensed Spectrum. IEEE Transactions on Mobile Computing, 2016, 15(4):803–816

[93] Chen B, Yang C. Energy Costs for Traffic Offloading by Cache-enabled D2D Communications. in: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), Doha, Qatar, 2016, 1–6

[94] Ross S M. Introduction to Probability Models, 10 ed. USA: Academic press, 2014

[95] Krause A, Golovin D. Submodular Function Maximization. Tractability: Practical Approaches to Hard Problems, 2012, 3(19):8–18

[96] Zhang S, Zhang N, Zhou S, et al. Energy-aware Traffic Offloading for Green Heterogeneous Networks. IEEE Journal on Selected Areas in Communications, 2016, 34(5):1116–1129

[97] Guan Y, Xiao Y, Feng H, et al. MobiCacher: Mobility-aware Content Caching in Small-cell Networks. in: Proceedings of IEEE Global Communications Conference (GLOBECOM), Austin,USA, 2014, 4537–4542

[98] Ahmed A, Ahmed E. A survey on mobile edge computing. in: Proceedings of the 10 th IEEE International Conference on Intelligent Systems and Control, Coimbatore, India, 2016, 1–8

[99] Tong L, Li Y, Gao W. A Hierarchical Edge Cloud Architecture for Mobile Computing. in: Proceedings of the 35th IEEE International Conference on Computer Communications (INFOCOM), San Francisco, USA, 2016, 399–400

[100] Chen M, Hao Y, Li Y, et al. On the Computation Offloading at ad hoc Cloudlet: Architecture and Service modes. IEEE Communications Magazine, 2015, 53(6):18–24

[101] Boyd S, Vandenberghe L. Convex Optimization, 1 ed. USA: Cambridge university press, 2004, pages 539–610

[102] Fazel M, Hindi H, Boyd S P. Log-det Heuristic for Matrix Rank Minimization with Applications to Hankel and Euclidean Distance Matrices. in: Proceedings of American Control Conference, Colorado,USA, 2003, 2156–2162

[103] Liu Y, Niu D, Li B. Delay-Optimized Video Traffic Routing in Software-Defined Interdatacenter Networks. IEEE Transactions on Multimedia, 2016, 18(5):865–878

[104] Luenberger D G, Ye Y, et al. Linear and nonlinear programming, 4 ed. Germany: Springer, 2016, pages 179–230

[105] Leung V, Taleb T, Chen M, et al. Unveiling 5G Wireless Networks: Emerging Research Advances, Prospects, and Challenges [Guest Editorial]. IEEE Network, 2014, 28(6):3–5

[106] Lei L, Zhang Y, Shen X, et al. Performance Analysis of Device-to-Device Communications with Dynamic Interference using Stochastic Petri Nets. IEEE Transactions on Wireless Communications, 2013, 12(12):6121–6141

[107] Han B, Hui P, Kumar V A, et al. Mobile Data Offloading through Opportunistic Communications and Social Participation. IEEE Transactions on Mobile Computing, 2012, 11(5):821–834

[108] Wang X, Chen M, Han Z, et al. TOSS: Traffic Offloading by Social Network Service-based Opportunistic Sharing in Mobile Social Networks. in: Proceedings of the 33th IEEE International Conference on Computer Communications (INFOCOM), Toronto, Canada, 2014, 2346–2354

[109] Li Q, Yang P, Yan Y, et al. Your Friends are More Powerful than You: Efficient Task Offloading through Social Contacts. in: Proceedings of IEEE International Conference on Communications (ICC), Sydney, Australia, 2014, 88–93

[110] Wang X, Chen M, Han Z, et al. Content Dissemination by Pushing and Sharing in Mobile Cellular Networks: An Analytical Study. in: Proceedings of the 9th International Conference on Mobile Ad hoc and Sensor Systems (MASS), Shanghai, China, 2012, 353–361

[111] Huo R, Yu F R, Huang T, et al. Software Defined Networking, Caching, and Computing for Green Wireless Networks. IEEE Communications Magazine, 2016, 54(11):185–193

附录 1　攻读博士学位期间发表的学术论文目录

[1] Chen M, **Hao Y**, Lai C, et al. On The Computation Offloading at Ad Hoc Cloudlet: Architecture and Service Models, IEEE Communications, 2015, 53(6):18–24 (第二作者, 导师一作)

[2] Chen M, **Hao Y**, Li Y, et al. Opportunistic Workflow Scheduling over Co-Located Clouds in Mobile Environment, IEEE Transactions on Services Computing, DOI: 10.1109/ TSC.2016.2589247, 2016 (第二作者, 导师一作)

[3] Chen M, **Hao Y**, Qiu M, et al. Mobility-aware Caching and Computation Offloading in 5G Ultradense Cellular Networks, Sensors, 2016, 16(7): 974–987 (第二作者, 导师一作)

[4] **Hao Y**, Chen M, Hu L, et al. Wireless Fractal Ultra-dense Cellular Networks, Sensors, 2017, 17(4):841–847 (第一作者)

[5] Chen M, **Hao Y**, Hwang K, et al. Green and Mobility-aware Caching in 5G Networks, IEEE Transactions on Wireless Communications, in revision, 2017 (第二作者，导师一作)

[6] Chen M, **Hao Y**, Mao S, et al. User Intent-oriented Video QoE with Emotion Detection Networking, in: Proceedings of the 59th IEEE Global Communications Conference (Globecom), Washington, USA, 2016, 1552–1559 (第二作者，导师一作)

[7] Chen M, **Hao Y**, Li Y, et al. Demo: LIVES: Learning through Interactive Video and Emotion-aware System, in: Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), Hangzhou, China, 2015, 399–400 (第二作者，导师一作)

[8] Chen M, **Hao Y**, Hwang K, et al. Disease Prediction by Machine Learning over Big Healthcare Communities, IEEE Access, DOI: 10.1109/ACCESS.2017.2694446, 2017 (第二作者)

[9] Shi X, **Hao Y**, Zeng D, et al. Cloud-Assisted Mood Fatigue Detection System, ACM/Springer Mobile Networks and Applications, 2016, 21(5): 744-752 (第二作者)

[10] Zhou P, **Hao Y**, Yang J, et al. Cloud-assisted Hugtive Robot for Affective Interaction, Multimedia Tools and Applications, 2017, 76(8): 10839–10854 (第二作者)

[11] Ma Y, **Hao Y**, Qian Y, et al. Cloud-assisted Humanoid Robotics for Affective Interaction, in: Proceedings of the 2nd IEEE International Conference on Control, Automation and Robotics (ICCAR), Hong Kong, China, 2016, 15–19 (第二作者)

[12] Chen M, Ma Y, **Hao Y**, et al. CP-Robot: Cloud-assisted Pillow Robot for Emotion Sensing and Interaction, in: Proceedings of the EAI International Conference on Industrial IoT Technologies and Applications (IndustrialIoT), Guangzhou, China, 2016, 81–93 (第三作者)

[13] Li B, Peng L, **Hao Y**, et al. Energy-Efficient Multiperiod Planning of Optical Core Network to Support 5G Networks, Transactions on Emerging Telecommunications Technologies, DOI: 10.1002/ett.3147, 2017 (第三作者)

附录 2　博士期间主持或参与的课题研究情况

1. 华中科技大学创新研究院技术创新基金：基于移动云计算的微云卸载机制研究（HUST: CX-15-055）
2. 科技部国际科技合作计划项目：基于机器人和云计算的新一代智能健康物联网合作研发（No. 2014DFT10070）
3. 国家自然科学基金面上项目：基于人体局域网跨网协作的情感交互理论与方法研究（No. 61572220）