文章编号:1007-5321(2017)02-0001-10 **DOI**:10.13190/j.jbupt.2017.02.001

面向 5G 需求的移动边缘计算

田 辉, 范绍帅, 吕昕晨, 赵鹏涛, 贺 硕
(北京邮电大学 网络与交换技术国家重点实验室, 北京 100876)

摘要: 移动边缘计算有助于实现第五代移动通信(5G)新业务超低时延、高能效、超高可靠和超高连接密度的需求,是未来 5G 通信的关键技术. 从细粒度任务卸载算法、高可靠任务卸载与预测算法以及服务器联合资源管理策略 3 个方面,介绍了现有移动边缘计算技术在面向 5G 业务需求的工作进展,分析了未来移动边缘计算面临的挑战,并给出了未来的研究方向和研究热点.

关 键 词: 移动边缘计算; 5G 需求; 任务卸载; 资源管理
中图分类号: TN911.22 **文献标志码**: A

Mobile Edge Computing for 5G Requirements

TIAN Hui, FAN Shao-shuai, LÜ Xin-chen, ZHAO Peng-tao, HE Shuo
(State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Mobile edge computing is a key technique to help push the envelope of performance to provide much lower latency, ultra-high energy efficiency, ultra-high reliability and much higher connectivity density in the fifth generation of mobile technology (5G). Fine-granularity based task offloading algorithms, task offloading and prediction algorithms with high reliability and joint resource management policy were introduced to report the progress of mobile edge computing toward the 5G business requirement. A number of challenges and potential research directions in mobile edge computing were also given.

Key words: mobile edge computing; 5G requirement; task offloading; resource management

0 引言

近年来,随着移动网络和智能化移动设备的不断升级,移动互联网用户数量呈现爆炸式增长. 根据 Cisco 最新发布的预测报告,2020 年全球移动数据流量将是 2015 年的 8 倍,联网设备数量将增加到 263 亿台[1]. 由此可见,移动互联网和物联网必将成为未来移动通信发展的主要驱动力.

目前,第五代移动通信(5G, the fifth generation of mobile technology)面临着爆炸式数据流量增长与海量设备连接并存的新挑战. 与此同时,5G 网络新增的业务场景,如无人驾驶汽车、智能电网、工业通信等,对时延、能效、设备连接数和可靠度等指标也提出了更高的要求. 为了应对移动互联网及物联网的高速发展,5G 需满足超低时延、超低功耗、超高可靠、超高密度连接的新型业务需求.

目前,增强现实(AR, augment reality)、在线游戏、云桌面等新型移动互联网业务飞速发展. 同时,智慧城市、环境监测、智能农业等物联网业务不断涌现. 然而,现有终端设备处理能力很难满足上述低时延、高复杂度、高可靠性的移动应用需求,进而影响用户体验. 移动云计算允许移动设备将本地计算

收稿日期: 2017-04-04
基金项目: 国家自然科学基金项目(61471060); 国家自然科学基金委创新研究群体项目(61421061); 国家留学基金委项目
作者简介: 田 辉(1963—), 女, 教授, 博士生导师, E-mail: tianhui@bupt.edu.cn.

任务部分或完全迁移到云端服务器执行,从而解决了移动设备自身资源紧缺问题,并且节约了任务本地执行能耗. 然而,将任务卸载到位于核心网的云服务器需要消耗回传链路资源,产生额外的时延开销,无法满足5G场景中低时延、高可靠的需求. 移动边缘计算(MEC, mobile edge computing)由欧洲电信标准协会于2014年率先提出,如图1所示. MEC系统允许设备将计算任务卸载到网络边缘节点,如基站、无线接入点等,既满足了终端设备计算能力的扩展需求,同时弥补了云计算时延较长的缺点. MEC迅速成为5G的一项关键技术,有助于达到5G业务超低时延、超高能效、超高可靠性等关键技术指标.

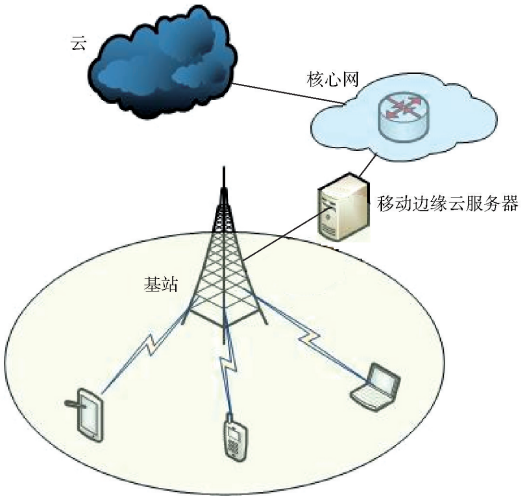

图1 MEC系统基本架构

首先,MEC能缩短任务执行时延. 移动应用任务处理时延包括传输时延、计算时延和通信时延. 传统的移动云计算中,信息需要经过无线接入网、回传链路到达位于核心网的云服务器. MEC将边缘服务器部署在无线接入网侧,缩短了计算服务器与移动设备距离. 由于传输距离的缩短,MEC的任务卸载不需要经过回传链路和核心网,减少了时延开销. 另一方面,边缘服务器的计算处理能力远大于移动设备,从而大幅缩短任务计算时延. 综上,MEC传输距离短、协议扁平化的特点使其能够满足5G网络的超低时延需求.

其次,MEC能大幅提升网络能效. 物联网设备可广泛应用到环境监测、人群感知、智能农业等各种场景. 但部署的物联网设备大多无法通过电网供电,在设备电池能量有限的情况下,MEC缩短了边缘服务器与移动设备的距离,大幅节约了任务卸载、无线传输所耗能量,延长了物联网设备的使用周期. 研究结果表明,对于不同的AR设备,MEC可延长30%~50%的设备电池寿命[2].

最后,MEC能提供更高的服务可靠性. MEC的服务器采用分布式部署,单个服务器服务规模小,不存储过多的有价值信息. 因此,相较于移动云计算的数据中心,不易成为被攻击的目标,可以提供更可靠的服务. 同时,多数移动边缘云服务器属于私有云,信息泄露风险低,也具有更高的安全性.

然而,MEC要满足5G的超低时延、高能效、超高可靠性的业务需求,仍面临以下关键性挑战.

1) 如何利用边缘服务器与移动设备的短距离特性,结合计算任务的属性提高任务卸载效率;

2) 如何在时变的无线环境中保证服务的可靠性与任务卸载效率;

3) 如何解决海量设备连接情况下,无线和计算资源短缺的问题,提升系统可扩展性.

1 MEC技术介绍

针对上述3点挑战,现有MEC研究提出了许多算法以满足上述5G新业务需求. 下面分别介绍提高任务卸载效率的细粒度任务卸载算法、实现服务高可靠性的任务卸载与预测算法以及海量设备连接下的联合资源管理算法.

1.1 高效的细粒度任务卸载算法

MEC由移动云计算演化而来,在靠近移动用户的无线接入网侧提供IT服务环境和云计算能力. 在MEC环境中,用户与边缘服务器的距离更近,数据传输和信令交换更快,任务卸载的通信开销更小,因此成为5G通信网络中的一项关键技术.

在MEC的任务卸载建模当中需要考虑很多因素,如时延和带宽以及计算资源需求量等. 在建模时难以兼顾所有因素,所以,现有文献对任务卸载的建模进行了一定的简化,得出了2种任务卸载模型,即二态的任务卸载和部分任务卸载模型.

在二态的计算卸载模型中,任务高度集成,即任务不能进一步被划分,只能够作为一个整体在本地执行或卸载到边缘服务器执行. 这样的不可划分的任务通常用一个三元素参数$A(L, T, X)$进行表示,其中L为输入数据量,T为任务完成的时延门限,X为完成单位比特任务需要的处理器资源. Miettinen等[3-8]提出了二态的计算卸载模型,上述3种参数可

以通过任务分析器估计得出. 这 3 种参数不仅能描述移动应用的基本需求,如计算和通信需求等,还能辅助估计任务执行的时延和能量消耗. Yuan 等[9]引入软时延门限要求,并且把执行单位比特输入数据的任务所需要的处理器资源建模为一个随机变量. 实际任务执行时延超过设定软时延门限的概率应小于给定阈值.

但是二态的计算卸载模型的卸载粒度大,仅能描述任务在本地执行或被整体卸载. 近几年,随着代码分解技术和并行运算的发展,MEC 中的部分任务卸载模型得到了广泛的关注. 具体地,有很多实际中的复杂移动应用可分解为多个计算任务,这些计算任务可以进一步划分成多个不同的任务模块. 任务模块作为卸载决策的最小单元,可以在本地执行或卸载到边缘服务器. 因此,在部分卸载模型中,移动应用被划分为 2 部分,即本地执行的任务模块和卸载到边缘服务器的任务模块. 由于卸载粒度的细化,部分卸载模型能够充分利用服务器和移动设备并行计算的优势,更加有效地降低移动设备能耗和移动应用的运行时延,提高卸载效率. 因此,学术界针对这种细粒度的部分卸载模型进行了广泛的研究,基于不同的研究目标提出了许多细粒度任务卸载算法. Jia 等[10]将移动应用建模为链状任务调用图,其中相邻的任务模块之间存在固定的先后顺序关系,即执行某一节点任务时必须先完成其先序任务.

考虑到链状任务调用图不能有效利用并行运算的优势,同时比较复杂的移动应用的子任务划分存在分支情况,业界提出了一种基于有向非循环图(DAG, directed acyclic graph)的任务模型. 这种基于 DAG 的任务模型能够将移动应用划分为更细粒度的任务模块,并运行任务模块间的并行处理. Mahmoodi 等[11]利用移动设备和服务器之间的负载均衡概念,设计了一种启发式分解算法达到最小化移动应用完成时间的目的. Kao 等[12]采用整数规划方式联合优化计算卸载、调度和云卸载策略,从而最大化资源节约. Zhang 等[13]研究了在规定的资源限制条件下最小化时延问题,并且提出了能够保证卸载性能的多项式时间近似算法. Khalili 等[14-16]将物理层参数与程序划分机制联合优化. Deng 等[17-18]基于 DAG 模型描述了在时延限制条件下最小化移动设备能耗的细粒度任务卸载算法.

Deng 等[17]将移动应用建模为拥有多个子任务的 DAG(见图 2),并在任务完成时间限制下以最小化能耗为目标,将任务卸载问题构建成非线性 0 ~ 1 规划问题.

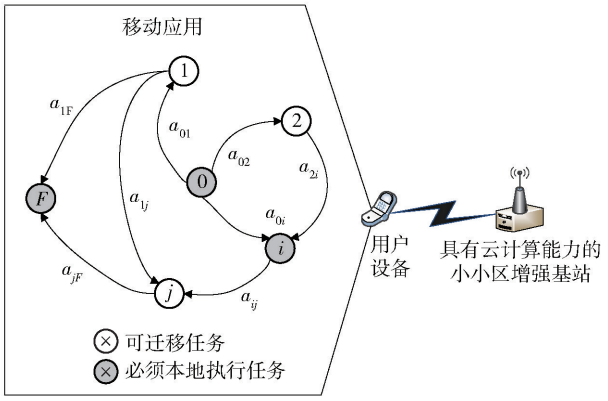

图 2　具有云计算能力的小小区增强基站及移动应用模型

Zhao 等[18]同样研究了 DAG 模型下的带有时延限制条件的移动设备能量消耗最小化问题. 为了更有效地解决该问题,提出了一种基于部分关键路径的贪心算法.

此外,在部分卸载模型当中,存在一种全粒度程序划分方式. 在这种方式当中,任务的输入数据能够以任意的比例划分为本地执行和服务器执行 2 部分. Wang 等[19]联合优化卸载比例、发射功率和处理器资源分配,并分别求解 2 个优化问题:一是在时延限制条件下最小化移动设备能耗;二是在能耗限制条件下最小化时延. 构建的 2 个问题都是非凸问题,Wang 等[19]利用变量替换技术求出第 1 个问题的最优解,利用连续凸逼近技术得到第 2 个问题的次优解.

1.2　高可靠的任务卸载与主动缓存算法

MEC 环境与云计算环境不同,边缘服务器位于无线接入网内,避免了回传链路带来的时延. 因此,MEC 的任务卸载需要考虑无线接入网的随机信道条件和服务器的占用率,如信道、任务到达、资源占用率等. 具体地,由于无线信道的衰落特性不断变化,计算任务随机到达移动设备,边缘服务器系统中的资源占用率也由于多用户竞争而随时间变化. 上述随机环境会影响任务卸载效率. 然而,在传统移动云计算研究中,由于回传链路时延远大于无线传输与计算时延,任务卸载决策没有考虑上述随机特性. 因此,自适应 MEC 随机环境的高可靠任务卸载策略或预测方式成为满足 5G 高可靠性需求的关键

任务.

目前,Huang 等[20-25]采用了随机任务模型,移动用户会随机产生任务,这些卸载的并且未被执行的任务会进入到 MEC 服务器的缓存当中排队. 这种系统模型考虑长期平均的系统性能,相比于确定性任务到达的系统长期平均性能更具优势. Huang 等[20]设计了一种基于李雅普诺夫优化技术的动态卸载算法,来最小化移动设备能耗,同时保证将超出时延限制的任务比例控制在给定门限下. Liu 等[21]假设本地运行和边缘运行同时可用,设计了一种基于马尔可夫决策过程理论的时延最优化任务调度策略,该策略根据信道状态控制本地的处理状态、传输单元以及缓存中的任务队列长度. Chen 等[22]和 Hong 等[23]考虑最小化长期平均运行成本问题,联合优化计算时延和能量消耗. Chen 等[22]根据马尔可夫决策过程理论来优化卸载的数据量. Hong 等[23]根据一个半马尔可夫决策过程框架来联合控制本地 CPU 频率、调制方式以及数据速率. Kwak 等[24]研究了不同类别移动应用场景下 MEC 的能量-时延折中问题,提出了一种基于李雅普诺夫优化的算法来联合决定卸载策略、任务分配、CPU 时钟频率和选择的网络接口. Jiang 等[25]将 Kwak 等[24]的研究工作扩展到了多核移动设备.

为了应对 MEC 的随机环境,提供低时延、高可靠的边缘计算服务,Lü 等[26]将 MEC 中的环境变化建模为随机扰动,并且根据模型预测控制理论设计了一种自适应的在线任务卸载算法. 由于随机扰动存在,用户需要感知环境实时随机扰动情况,决定决策窗口长度. 提出的自适应滚动时域卸载机制通过监测任务执行时偏差来判断环境扰动,并根据扰动频率,调整折扣因子和决策窗口长度. 图 3 所示为窗口长度为 3 的滚动时域卸载机制示意图. 增加决策窗口长度能够优化未来系统状况,减小窗口长度可以在随机扰动频繁时优先考虑即将执行任务的迁移决策. 对于决策窗口中固定的配置信息,Lü 等[26]提出了多目标的动态规划算法,在最小化预估开销的同时,保证用户的时延需求. 仿真结果表明,与静态最优卸载策略相比,自适应的在线任务卸载算法在随机环境当中具有显著的优势.

有预测显示,视频流量将占据未来移动流量数据的绝大部分. 而视频流量的内容集中性和请求重复性也使得缓存视频内容成为一种有效降低网络负担的技术. 但是,移动边缘云服务器的缓存容量有限,无法在边缘服务器部署所有服务和缓存内容. 因此,边缘服务器需要准确预测内容流行度,从而缓存当前流行内容或服务.

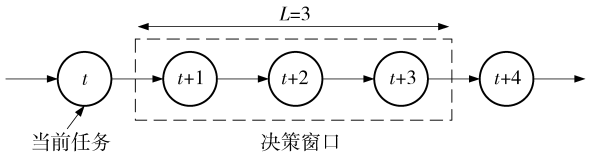

图 3 窗口长度为 3 的滚动时域卸载机制示意图

现有内容流行度预测方式主要基于大数据分析结果. 而在边缘网络中,边缘服务器覆盖用户数目少,用户与大群体分析结果存在较大的差异,无法采用基于大数据的预测方式. 因此,He 等[27]基于社交关系利用传播动力学预测小用户数目条件下的文件的流行度,研究场景如图 4 所示. 采用的传播动力学模型既不需要大数据预测的训练阶段,也不需要内容流行度的先验概率获取,能有效预测小用户数目下的内容流行度. He 等[27]首先基于经典的传染病模型,利用用户间的社交关系建立用户间内容传播网络,提出离散时间马尔可夫链方法来预测从个体角度出发的特定内容被访问的概率. 仿真结果表明,提出的预测模型显著优于其他方案,预测准确性相比对比算法最高提升了 94%.

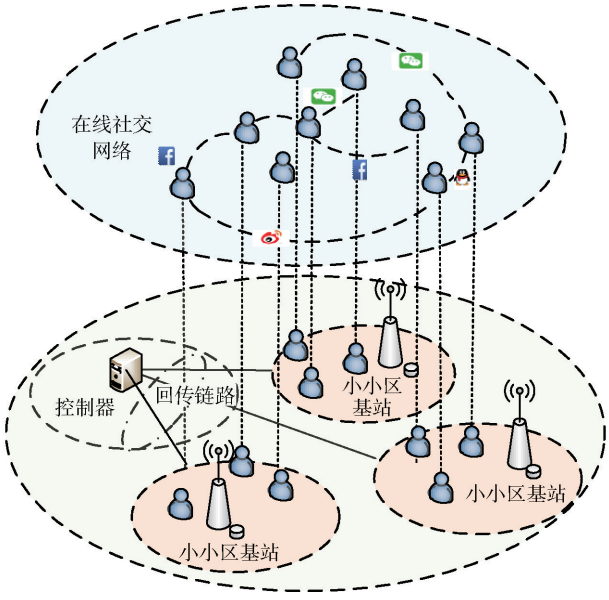

图 4 社交网络与边缘缓存相结合的场景

移动性是移动用户的固有特点. 在 MEC 中,边

缘服务器可以感知用户的移动以及运动轨迹，并获取其个人偏好信息，从而提高任务卸载的可靠性以及资源利用效率，降低反应时间. 另一方面，移动性也给普遍存在的可靠的计算带来了巨大的挑战，因为用户移动性会造成频繁的小区切换、服务中断，导致用户体验恶化. Wang 等[28]定义用户间接触概率来建模分析用户移动性，并通过凸优化设计了一种机会卸载策略来最大化任务成功卸载概率. Zhang 等[29]为了考虑用户移动性，利用随机几何理论，将可用边缘服务器的设备数量建模为同性随机点过程，并通过求解马尔可夫决策过程得到卸载策略. Lee 等[30]和 Rahimi 等[31]分别根据一系列用户可以连接到的网络和一个二维的位置-时间工作流来建模用户的移动性. 此外，Prasad 等[32]将移动性管理与流量控制相结合，通过设计智能小区联合控制机制提高具有时延容忍任务的用户体验. 同时，为了提高任务卸载算法的可靠性，Ko 等[33]设计了移动性感知在线预取算法，Li 等[34]将设备间通信(D2D, device-to-device)应用到移动性感知卸载策略设计当中，Chen 等[35]和 Chen 等[36]设计了移动性感知的错误容忍机制，Chaisiri 等[37]和 Zhang 等[38]设计了移动性感知的服务器调度算法.

1.3　多用户的联合无线与计算资源管理

上面介绍了单用户场景下的任务卸载和计算资源策略. 在 5G 场景中，海量物联网设备位于同一小区范围，并竞争有限的无线信道和计算资源. 此时，联合管理无线和计算资源能减少用户间竞争带来的干扰，提高边缘计算效率. 图 5 则是在多用户场景下，考虑无线资源与计算资源联合管理的示意图[39].

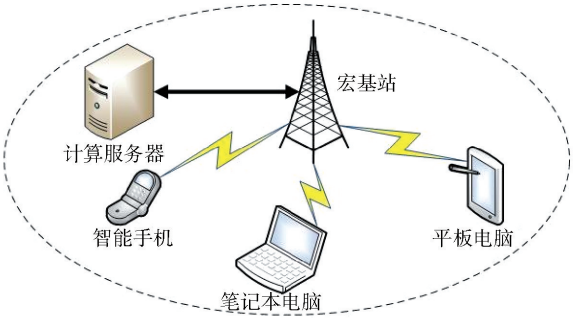

图 5　多用户的联合无线与计算资源管理示意图

You 等[40-47]研究了集中式的联合资源管理策略. 其中，边缘服务器拥有全部的系统信息(如信道条件、任务到达等)，进行集中式的资源分配，并通知用户按分配结果上传任务. You 等[40-42]联合优化无线与计算资源以最小化系统能耗. 在文献[40]中，移动设备的计算任务量和本地处理能力各不相同，并通过时分方式复用边缘云服务器资源. 边缘云服务器联合优化用户卸载数据大小和传输时间，构建的联合优化问题可以通过凸优化求解. 在文献[41]中，边缘云服务器优化设备传输功率和计算资源分配比例，并为每个用户设计了唯一的从传输功率到计算资源分配的单射函数. Lorenzo 等[42]进一步考虑了用户侧的细粒度卸载策略以提升卸载效率，并由边缘云服务器集中式为每个用户进行卸载决策. 不同于文献[40-42]仅考虑系统能耗，Hoang 等[43-47]考虑了用户的服务质量需求. 其中，文献[43]研究了在保障用户服务质量情况下的服务提供商收益最大化问题，假设每个用户均分边缘云服务器计算资源，该问题可以构建为一个半马尔可夫决策问题，并转化成线性规划问题求解. Mao 等[44]研究了随机任务到达情况下的无线与计算资源分配问题，提出的基于李雅普诺夫优化的在线算法表明在多用户 MEC 系统中存在能耗与时延的折中关系，即无法同时实现能耗与时延的最小化. Munoz 等[45-47]研究了云接入网中边缘云服务器的联合资源管理算法.

在未来海量设备连接场景中，文献[40-47]中描述的集中式资源分配策略由于边缘服务器需要获取所有设备的任务到达和信道情况信息，面临着巨大的信令开销. Chen 等[48]研究了不同用户传输互相干扰情况下二态的用户卸载策略，以最大化系统效用. 为了减小集中式处理的信令开销，构建的 NP 难整数规划问题可以利用博弈论转化为多用户的卸载博弈过程，设计的分布式卸载决策可以达到博弈的纳什均衡. 在文献[48]的基础上，Ma 等[49]进一步考虑了多个接入点同时覆盖相同区域，该区域内用户可以选择接入点将任务卸载到边缘服务器，并对用户卸载切换的能耗进行了建模，设计了分布式的卸载策略.

考虑到密集小小区网络中的小区间干扰问题，Deng 等[50]提出了基于顺序博弈的联合迁移决策和资源分配策略. 在密集小小区系统中，同时卸载用户数受小区间干扰和可用计算资源限制. 例如，边缘服务器可分配虚拟机数目有限，随着卸载用户数增加，卸载的任务会在虚拟机侧排队，从而影响用户体验. 具体地，用户监测决策环境，预测计算迁移对

应的能耗和时延,根据自己的效用选择最佳的决策.计算迁移用户选择信道抑制干扰,边缘服务器优化可用计算资源. Deng 等[50]提出的基于顺序博弈的迁移决策能达到纳什均衡,并利用各个移动设备的计算能力,降低系统信令开销.

Chen 等[48-50]提出的基于博弈论的分布式资源分配策略能减小系统信令开销,然而达到纳什均衡需要反复迭代过程,带来额外的能量和时延开销. Lü 等[39]提出了基于去耦理论的分布式卸载决策和无线与计算资源分配算法. 该算法免去了上述卸载博弈达到均衡的迭代过程以及迭代带来的开销. 具体地,构建的联合卸载决策和资源分配的优化问题是一个非线性混合整数规划问题,并且卸载决策和无线与计算资源分配间相互耦合. 基于去耦理论,Lü 等[39]将原问题分解为卸载决策和资源分配2个子问题分别求解. 资源分配子问题可以通过凸优化和准凸优化求得闭式解,从而将原问题退耦,转化为一个整数规划问题,并证明其次模性质. 由于问题去耦,用户可以独立完成传输功率优化,并根据系统状态决定是否上传卸载请求,边缘服务器收到请求后联合优化卸载决策和计算资源分配.

2 未来挑战与展望

下面从 MEC 系统中的缓存应用、移动性管理以及绿色节能3个方面介绍 MEC 研究中存在的机遇与挑战.

2.1 部署缓存的 MEC 系统

Cisco 最新发布的预测报告显示,到2021年,视频流量将占据移动数据流量的82%[1]. 这些视频流量存在以下特点:请求内容具有集中性,并且流行度较高的视频内容会在不同时间被大量重复请求. 目前,许多研究利用缓存技术将流行度高的视频内容缓存到基站侧,从而减少视频下载时延,并减轻网络回传链路的负担. 为进一步满足5G中的超低时延需求,将缓存技术与移动边缘云计算技术相结合,当有多个用户迁移相同的计算任务时,通过在移动边缘云服务器缓存的数据与服务,可有效地降低任务处理时延.

然而,由于移动边缘云服务中的计算资源和存储资源的有限性,用户业务请求与实际资源之间存在偏差,在部署缓存的移动边缘云系统中,为了在满足用户体验的前提下,最大化边缘云服务器中的计算资源和存储资源的利用率,制定有效的缓存策略并提高缓存命中率是需要解决的关键问题.

5G 围绕人和其周围的事物,是一种万物互联的通信. 基于此,首先可以从人出发,利用大数据技术挖掘人与事物之间的依赖关系以及人与人之间的社交关系,从而预测某一服务或者数据在人群中的流行程度,构建准确的人-物流行度预测模型.

其次,从多个维度研究请求内容流行度的演化. 同一时间不同地点的内容请求分布是不同的,如同一时间的博物馆更倾向于请求 AR 服务,而咖啡馆等休闲场所则是超高清视频服务. 同样,同一地点不同时间的内容请求分布也有所差异. 构建准确的时空流行度预测模型是提高缓存命中率,解决移动边缘云计算资源瓶颈的关键方法.

2.2 MEC 系统中的移动性管理

移动性管理是 MEC 系统中需要解决的一个关键问题. 尤其在5G超密集部署场景中,用户移动带来的频繁切换、小区间干扰、用户体验下降等问题愈演愈烈,给 MEC 带来了巨大的挑战. 虽然异构蜂窝网络中已有较成熟的移动性模型,基于这些模型构建的动态移动性管理体系已经可以取得较高的数据速率和低错误率[51-52],但是上述研究主要针对移动环境下的无线资源管理,而 MEC 系统中最主要的计算资源并未考虑,所以 MEC 系统中的移动性管理仍是一个需要逐步完善的领域. 下面将从边缘服务器计算资源预分配以及 D2D 辅助2个潜在研究方向进行简要介绍.

1)基于移动性的边缘服务器计算资源预分配

随着用户的移动,承担任务迁移的边缘服务器频繁切换. 当切换到下一个承担任务的边缘服务器时,切换所需传输的大量数据以及资源的重新获取将导致时延加大,并加重 MEC 网络的负担. 通过预测用户移动轨迹,并准确实现边缘服务器计算资源预分配,可以预先得到所需的任务计算数据,从而有效地减少时延并降低不必要的能耗. 但是该技术面临以下挑战:准确的用户轨迹预测和预取任务数据的选择.

用户数据的不完整性和数据特征的模糊性,给准确预测用户轨迹带来了极大的挑战. 在获取有限数据的基础上,可以从用户行为出发,基于时间和空间2个维度对用户行为进行细分,从而得到多维度的行为特征,并可以选择使用定量和定性方式更准确地描述个体用户的移动行为,通过设计相应算法

自适应地选取倾向性最强的特征作为预测依据，从而提高用户轨迹预测的准确性.

2）移动环境中的 D2D 辅助任务迁移

D2D 通信技术是一种在基站的控制下，允许终端之间通过复用蜂窝网络资源直接进行通信的技术. 它可以有效地卸载基站通信压力，提高网络通信容量. 将 D2D 技术应用到 MEC 中，通过建立 D2D 通信链路，可以将用户的计算任务通过 D2D 链路迁移到邻近且计算资源相对充足的另一设备上. 这种短距离通信可有效降低数据传输能耗. 但是该技术面临以下挑战：

移动用户是否有意愿进行 D2D 通信是使用 D2D 技术的前提. 因此，应设计有效的激励机制，使得用户双方都可以在使用 D2D 通信的情况下获得收益，从而建立 D2D 链路. 该种激励机制可以根据不同情况下用户的需求来建立.

在移动用户愿意进行 D2D 通信的前提下，如何根据用户的移动信息、动态的信道环境和异构用户的计算能力来优化承担任务迁移的用户选择策略是又一关键性问题. D2D 能成功通信的关键是用户之间的距离，因此，可以通过研究与 D2D 通信链路持续时间成正相关的社交相似度，利用随机位点模型结合轨迹相似度预测 D2D 链路持续时间，更加有效地进行 D2D 用户配对.

但是，大量 D2D 链路会带来严重的通信干扰，尤其是在快速变化的无线衰落环境中，这种干扰情况更加复杂，为保证 D2D 在 MEC 系统中的应用，可以结合认知无线电技术设计更加有效的干扰消除方法，在准确预测用户移动轨迹的前提下，结合终端间协作，增加用户任务迁移率，减少服务时延.

2.3　绿色 MEC 系统设计

随着 5G 时代的到来，网络实现超密集部署，移动边缘云服务器的超密集部署将会大大增加系统整体能耗. 为了响应“绿色”发展，MEC 系统应朝着超低能耗的方向发展. 下面简要介绍 2 种潜在的研究方向.

1）建立更加有效的资源管理方案

为更加有效地进行资源管理和分配，可以建立准确的计算工作量预测模型，在最大化资源分配效率的基础上最小化系统能耗，充分利用系统资源，建立合理的资源分配机制.

2）充分利用可再生资源

传统的电网能量大多通过燃煤发电，必然会给环境带来巨大压力. 新能源技术的发展为绿色 MEC 系统带来了新的机遇. 通过能量收集技术，利用风能、太阳能等给移动设备充电，不仅会延长电池的使用寿命，也减少了因为充电而带来的服务受阻或暂停等问题. 同时，新兴的能量收集技术不仅可以利用可再生能源，也可以收集信息传输过程中电磁波携带的能量等，为实现移动边缘云系统的绿色发展提供了可能.

但是利用新能源技术构建新型 MEC 系统同样面临着以下挑战：由于新能源的加入，改变了能耗组成成分，在传统能耗组成的约束条件下的资源分配方案不再适用；由于新能源的免费、实时等特性，如何在满足用户体验、时延等要求下，优化能量资源分配是需要解决的关键问题之一；同时，由于能量收集的不稳定性，会导致任务卸载过程的中断从而导致任务卸载失败，如何提升 MEC 系统任务卸载的鲁棒性也是一项关键挑战.

首先，为解决能量组成成分改变问题，可以设计合适的能量选择算法，根据业务量特点，在不同的时间、空间维度上选用不同的可再生能源，建立稳定高效的时空-能源选择模型.

其次，针对能量收集的不稳定问题，结合 5G 基站密集部署的实际，可以考虑设计更加合理的移动边缘云计算服务器的部署策略，在最小化成本的前提下保证能量收集的连续性和稳定性.

此外，可采用可再生能源与传统能源相结合的方法，并设计相应的资源分配方案，从而保证任务迁移过程的稳定性.

3　结束语

MEC 用分布式部署于接入网的计算节点取代传统位于核心网的数据中心网络，减小移动设备与计算服务器的距离，从而减小任务执行时延与能耗，并提供高可靠的计算服务. 然而，MEC 效率受时变的无线信道条件影响，同时有限的边缘服务器计算能力难以同时服务大量设备. 笔者总结了现有 MEC 成果，分析了 MEC 可以从高效的细粒度迁移算法、高可靠的迁移与预测算法以及多用户联合资源管理 3 个方面满足超低时延、高能效、高可靠性和超高连接密度的 5G 新业务需求. 笔者还预测了未来 MEC 的发展趋势，并分析了部署缓存的 MEC 系统、MEC 系统中的移动性管理以及绿色 MEC 3 个方向的研究点.

参考文献:

[1] Cisco. Cisco visual networking index: global mobile data traffic forecast update, 2016-2021 white paper[EB/OL]. [2017-03-01]. http://10.3.200.202/cache/10/03/cisco.com/89e8529e7886890c828d4a976994f806/mobile-white-paper-c11-520862.pdf.

[2] Shi Bowen, Yang Ji, Huang Zhanpeng, et al. Offloading guidelines for augmented reality applications on wearable devices[C]// ACM International Conference. [S.l.]: ACM, 2015: 1271-1274.

[3] Miettinen A P, Nurminen J K. Energy efficiency of mobile clients in cloud computing[J]. HotCloud, 2010 (10): 4-4.

[4] Melendez S. Computation offloading decisions for reducing completion time[Z]. arXiv, 2016: 1608.05839.

[5] Zhang Weiwen, Wen Yonggang, Guan K, et al. Energy-optimal mobile cloud computing under stochastic wireless channel[J]. IEEE Transactions on Wireless Communications, 2013, 12(9): 4569-4581.

[6] Kumar K, Lu Y H. Cloud computing for mobile users: can offloading computation save energy[J]. Computer, 2010, 43(4): 51-56.

[7] Kumar K, Liu Jibang, Lu Y H, et al. A survey of computation offloading for mobile systems[J]. Mobile Networks and Applications, 2013, 18(1): 129-140.

[8] Barbarossa S, Sardellitti S, Di Lorenzo P. Communicating while computing: distributed mobile cloud computing over 5G heterogeneous networks[J]. IEEE Signal Processing Magazine, 2014, 31(6): 45-55.

[9] Yuan Jibang, Nahrstedt K. Energy-efficient soft real-time CPU scheduling for mobile multimedia systems[C]// ACM SIGOPS Operating Systems Review. [S.l.]: ACM, 2003: 149-163.

[10] Jia M, Cao Jibang, Yang Lei. Heuristic offloading of concurrent tasks for computation-intensive applications in mobile cloud computing[C]//2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). [S.l.]: IEEE, 2014: 352-357.

[11] Mahmoodi S E, Uma R N, Subbalakshmi K P. Optimal joint scheduling and cloud offloading for mobile applications[J]. IEEE Transactions on Cloud Computing, 2016 (99): 1-1.

[12] Kao Y H, Krishnamachari B, Ra M R, et al. Hermes: latency optimal task assignment for resource-constrained mobile computing[C]//2015 IEEE Conference on Computer Communications (INFOCOM). [S.l.]: IEEE, 2015: 1894-1902.

[13] Zhang Weiwen, Wen Yonggang, Wu D O. Collaborative task execution in mobile cloud computing under a stochastic wireless channel[J]. IEEE Transactions on Wireless Communications, 2015, 14(1): 81-93.

[14] Khalili S, Simeone O. Inter - layer per - mobile optimization of cloud mobile computing: a message - passing approach[J]. Transactions on Emerging Telecommunications Technologies, 2016, 27(6): 814-827.

[15] Di Lorenzo P, Barbarossa S, Sardellitti S. Joint optimization of radio resources and code partitioning in mobile edge computing[Z]. arXiv, 2013: 1307.3835.

[16] Mahmoodi S E, Subbalakshmi K P, Sagar V. Cloud offloading for multi-radio enabled mobile devices[C]// 2015 IEEE International Conference on Communications (ICC). [S.l.]: IEEE, 2015: 5473-5478.

[17] Deng Maofei, Tian Hui, Fan Bo. Fine-granularity based application offloading policy in cloud-enhanced small cell networks[C]//2016 IEEE International Conference on Communications Workshops (ICC). [S.l.]: IEEE, 2016: 638-643.

[18] Zhao Pengtao, Tian Hui, Fan Bo. Partial critical path based greedy offloading in small cell cloud[C]// IEEE VTC. [S.l.]: IEEE, 2016: 1-5.

[19] Wang Yanting, Sheng Min, Wang Xijun, et al. Mobile-edge computing: partial offloading using dynamic voltage scaling[J]. IEEE Transactions on Communications, 2016, 64(10): 4268-4282.

[20] Huang Dong, Wang Ping, Niyato D. A dynamic offloading algorithm for mobile computing[J]. IEEE Transactions on Wireless Communications, 2012, 11(6): 1991-1995.

[21] Liu Juan, Mao Yuyi, Zhang Jun, et al. Delay-optimal computation task scheduling for mobile-edge computing systems[C]// 2016 IEEE International Symposium on Information Theory (ISIT). [S.l.]: IEEE, 2016: 1451-1455.

[22] Chen Shuang, Wang Yanzhi, Pedram M. A semi-markovian decision process based control method for offloading tasks from mobile devices to the cloud[C]// Global Communications Conference (GLOBECOM). [S.l.]: IEEE, 2013: 2885-2890.

[23] Hong S T, Kim H. QoE-aware computation offloading scheduling to capture energy-latency tradeoff in mobile

clouds[C]//2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON). [S.l.]: IEEE, 2016: 1-9.

[24] Kwak J, Kim Y, Lee J, et al. DREAM: dynamic resource and task allocation for energy minimization in mobile cloud systems[J]. IEEE Journal on Selected Areas in Communications, 2015, 33(12): 2510-2523.

[25] Jiang Zhefeng, Mao Shiwen. Energy delay tradeoff in cloud offloading for multi-core mobile devices[J]. IEEE Access, 2015(3): 2306-2316.

[26] Lü Xinchen, Tian Hui. Adaptive receding horizon offloading strategy under dynamic environment[J]. IEEE Communications Letters, 2016, 20(5): 878-881.

[27] He Shuo, Tian Hui, Lü Xinchen. Edge popularity prediction based on social-driven propagation dynamics[J]. IEEE Communications Letters, 2017, 21(5): 1-4.

[28] Wang Chuanmeizhi, Li Yong, Jin Depeng. Mobility-assisted opportunistic computation offloading[J]. IEEE Communications Letters, 2014, 18(10): 1779-1782.

[29] Zhang Yang, Niyato D, Wang Ping. Offloading in mobile cloudlet systems with intermittent connectivity[J]. IEEE Transactions on Mobile Computing, 2015, 14(12): 2516-2529.

[30] Lee K, Shin I. User mobility model based computation offloading decision for mobile cloud[J]. JCSE, 2015, 9(3): 155-162.

[31] Rahimi M R, Venkatasubramanian N, Vasilakos A V. MuSIC: mobility-aware optimal service allocation in mobile cloud computing[C]//2013 IEEE Sixth International Conference on Cloud Computing (CLOUD). [S.l.]: IEEE, 2013: 75-82.

[32] Prasad A, Lundén P, Moisio M, et al. Efficient mobility and traffic management for delay tolerant cloud data in 5G networks[C]//2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). [S.l.]: IEEE, 2015: 1740-1745.

[33] Ko S W, Huang Kaibin, Kim S L, et al. Online prefetching for mobile computation offloading[Z]. arXiv, 2016: 1608.04878.

[34] Li Yujin, Sun Lei, Wang Wenye. Exploring device-to-device communication for mobile cloud computing[C]// 2014 IEEE International Conference on Communications (ICC). [S.l.]: IEEE, 2014: 2239-2244.

[35] Chen C A, Won M, Stoleru R, et al. Energy-efficient fault-tolerant data storage and processing in mobile cloud [J]. IEEE Transactions on Cloud Computing, 2015, 3(1): 28-41.

[36] Chen C A, Stoleru R, Xie G G. Energy-efficient and fault-tolerant mobile cloud storage[C]//2016 5th IEEE International Conference on Cloud Networking (Cloudnet). [S.l.]: IEEE, 2016: 51-57.

[37] Chaisiri S, Lee B S, Niyato D. Optimization of resource provisioning cost in cloud computing[J]. IEEE Transactions on Services Computing, 2012, 5(2): 164-177.

[38] Zhang Yuan, Yan Jinyao, Fu Xiaoming. Reservation-based resource scheduling and code partition in mobile cloud computing[C]//2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). [S.l.]: IEEE, 2016: 962-967.

[39] Lü Xinchen, Tian Hui, Zhang Ping, et al. Multi-user joint task offloading and resources optimization in proximate clouds[J]. IEEE Transactions on Vehicular Techology, 2017, 66(4): 3435-3447.

[40] You Changsheng, Huang Kaibin, Chae H, et al. Energy-efficient resource allocation for mobile-edge computation offloading[J]. IEEE Transactions on Wireless Communications, 2017, 16(3): 1397-1411.

[41] Barbarossa S, Sardellitti S, Lorenzo P Di. Joint allocation of computation and communication resources in multiuser mobile cloud computing[C]//IEEE International Workshop Signal Processing Advances Wireless Communications (SPAWC). Darmstadt, Germany: [s.n.], 2013: 26-30.

[42] Lorenzo P D, Barbarossa S, Sardellitti S. Joint optimization of radio resources and code partitioning in mobile edge computing[Z]. arXiv, 2016: 1307.3835v3.

[43] Hoang D T, Niyato D, Wang Ping. Optimal admission control policy for mobile cloud computing hotspot with cloudlet[C]//IEEE Wireless Communications and Networking Conference (WCNC). Paris, France: [s.n.], 2012: 3145-3149.

[44] Mao Yuyi, Zhang Jun, S Song, et al. Power-delay tradeoff in multi-user mobile-edge computing systems [C]//IEEE Global Communications Conference (GLOBECOM). Washington, DC: [s.n.], 2016: 1-6.

[45] Munoz O, P-Iserte A, Vidal J. Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading [J]. IEEE Transactions on Vehicular Technology, 2015, 64(10):

497-508.

[46] Munoz O, P-Iserte A, Vidal J. Joint optimization of radio and computational resources for multicell mobile-edge computing[J]. IEEE Transactions on Signal and Information Processing Over Networks, 2015, 1(2): 89-103.

[47] Wang Kezhi, Yang Kun, Magurawalage C. Joint energy minimization and resource allocation in C-RAN with mobile cloud[J]. IEEE Transactions on Cloud Computing, 2016(99): 1-10.

[48] Chen Xu, Jiao Lei, Li Wenzhong, et al. Efficient multi-user computation offloading for mobile-edge cloud computing[J]. IEEE Transactions on Networking, 2016(24): 2795-2808.

[49] Ma Xiao, Lin Chuang, Xiang Xudong, et al. Game-theoretic analysis of computation offloading for cloudlet-based mobile cloud computing[C]// ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM). Cancun, Mexico: [s. n.], 2015: 271-278.

[50] Deng Maofei, Tian Hui, Lü Xinchen. Adaptive sequential offloading game for multi-cell mobile edge computing [C]// 2016 23rd International Conference on Telecommunications (ICT). Thessaloniki: [s. n.], 2016: 1-5.

[51] Lopez-Perez D, Guvenc I, Chu Xiaoli. Mobility management challenges in 3GPP heterogeneous networks [J]. IEEE Communications Magazine, 2012, 50(12): 70-78.

[52] Kassar M, Kervella B, Pujolle G. An overview of vertical handover decision strategies in heterogeneous wireless networks[J]. Computer Communications, 2008, 31(10): 2607-2620.