软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(8):2491−2516 [doi: 10.13328/j.cnki.jos.005861]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

移动边缘网络中计算迁移与内容缓存研究综述*

张开元[1,2], 桂小林[1,2], 任德旺[1,2], 李 敬[1,2], 吴 杰[1,2], 任东胜[1,2]

[1](西安交通大学 电子与信息工程学院,陕西 西安 710049)

[2](陕西省计算机网络重点实验室(西安交通大学),陕西 西安 710049)

通讯作者: 桂小林, E-mail: xlgui@mail.xjtu.edu.cn

摘 要: 随着移动设备数量的爆炸性增长以及许多新兴应用的出现,移动网络的流量呈指数级增长.传统的集中式网络架构由于回程链路负载过重、时延较长,无法满足移动用户的需求.因此,提出了将网络能力从核心网开放至边缘网的新体系结构,即移动边缘计算(MEC).移动边缘计算能够在移动蜂窝网络的边缘提供轻量级的云计算和存储能力.对移动边缘计算相关的最新研究成果进行了详尽的回顾:首先,概述了移动边缘计算的发展历程、关键问题和支撑技术;然后,针对 MEC 架构、计算迁移、边缘缓存和服务编排这 4 个关键研究问题进行了全面的综述,并讨论了增强现实、虚拟现实、动态内容交付、车联网和物联网等移动边缘计算中的典型应用案例;最后,从移动边缘计算功能增强、服务质量保障和安全可用性这 3 个方面展望了移动边缘计算的开放式研究挑战和未来的发展趋势.

关键词: 移动边缘计算;MEC 架构;计算迁移;边缘缓存;服务编排

中图法分类号: TP393

中文引用格式: 张开元,桂小林,任德旺,李敬,吴杰,任东胜.移动边缘网络中计算迁移与内容缓存研究综述.软件学报,2019,30(8):2491−2516. http://www.jos.org.cn/1000-9825/5861.htm

英文引用格式: Zhang KY, Gui XL, Ren DW, Li J, Wu J, Ren DS. Survey on computation offloading and content caching in mobile edge networks. Ruan Jian Xue Bao/Journal of Software, 2019,30(8):2491−2516 (in Chinese). http://www.jos.org.cn/1000-9825/5861.htm

Survey on Computation Offloading and Content Caching in Mobile Edge Networks

ZHANG Kai-Yuan[1,2], GUI Xiao-Lin[1,2], REN De-Wang[1,2], LI Jing[1,2], WU Jie[1,2], REN Dong-Sheng[1,2]

[1](School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

[2](Shaanxi Province Key Laboratory of Computer Network (Xi'an Jiaotong University), Xi'an 710049, China)

Abstract: With the explosive growth of mobile devices and the advent of many new applications, mobile network traffic volume has been growing exponentially. The traditional centralized network architecture cannot accommodate such user demands due to heavy burden on the backhaul links and long latency. Therefore, new architecture, which brings network capability to the edge of network, is proposed, i.e., mobile edge computing (MEC). MEC provides lightweight cloud computing and caching capabilities at the edge of cellular networks. In this survey, an exhaustive review on the state-of-the-art research efforts on mobile edge computing is carried out. First, an overview of mobile edge computing, including development process, research hotpots, and key enablers, is given. Then, a comprehensive survey of issues on mobile edge computing architecture, computation offloading, edge caching and service orchestration at the mobile edge computing is presented. The applications and use cases of mobile edge computing, such as augmented reality, virtual reality, dynamic content delivery, Internet of vehicles, and Internet of things, are discussed. Finally, from the prspectives of function enhancement, quality

* 基金项目: 国家自然科学基金(61472316); 陕西省重大基础研究项目(2016ZDJC-05); 陕西省重点研发项目(2017ZDXM-GY-011, 2019GY-005)

Foundation item: National Natural Science Foundation of China (61472316); Major Basic Research Plan of Shaanxi Province (2016ZDJC-05); Key Development Program of Shaanxi Province (2017ZDXM-GY-011, 2019GY-005)

收稿时间: 2018-04-18; 修改时间: 2018-09-23, 2019-01-29; 采用时间: 2019-04-27; jos 在线出版时间: 2019-05-22

CNKI 网络优先出版: 2019-05-22 15:26:24, http://kns.cnki.net/kcms/detail/11.2560.TP.20190522.1525.014.html

of service assurance, security, and availability, the open research challenges and future direction of mobile edge computing are presented as well.

Key words: mobile edge computing; MEC architecture; computation offloading; edge caching; service orchestration

1 引 言

近年来,随着移动设备(智能手机、笔记本电脑和平板电脑等)在人们的学习、娱乐、社交等日常生活中扮演着越来越重要的角色[1],移动用户对于数据传输速率和服务质量的需求也在日益增长.尽管新的移动设备具备的计算能力越来越强大,但即使是这些新设备也可能无法在短时间内处理计算密集型的应用程序(如虚拟现实、增强现实、人脸识别等).此外,运行高计算能力需求的应用所带来的电量消耗仍然是限制移动用户充分享受该类应用的重大障碍.这激发了移动云计算(mobile cloud computing,简称 MCC)概念的发展,将云计算能力集成到移动网络,使移动用户能够访问并使用云计算服务[2].在 MCC 中,移动设备(mobile device,简称 MD)可通过移动运营商的核心网(core network,简称 CN)访问并使用功能强大的远程云中心(cloud center,简称 CC)上的计算资源和存储资源.MCC 具有以下 3 个优点[3]:(1) 通过将高能量消耗的计算任务/应用程序迁移到云中心去执行,以达成延长移动设备电池寿命的目的;(2) 支持在移动设备上运行计算密集型的应用程序;(3) 为移动设备提供更强的数据存储能力.然而从网络拓扑的角度来看,MCC 与移动用户之间的距离过于遥远,这给移动网络的核心网增加了很大的负载,同时引发了高网络时延.

为了解决 MCC 环境下时间延迟过高的问题,新出现的边缘计算范例考虑将云服务部署在 MD 的附近,即移动网络的边缘.边缘计算可以被理解为是 MCC 的一种特殊情况.在传统的 MCC 中,MD 通过移动运营商的核心网来访问并使用云服务[4];而在边缘计算的环境下,计算和存储资源被假设存在于靠近 MD 的位置.因此,与 MCC 相比,MEC 可以提供更低的服务时延.此外,MCC 是一种完全集中式部署的服务,而边缘计算则是以分布式和集中式结合的方式部署.另一方面,边缘计算相对于 MCC 仅提供有限的计算和存储资源.表 1 概述了 MCC 和边缘计算在关键技术方面的比较.

Table 1 Comparison of MCC and edge computing concepts

表 1 MCC 和边缘计算概念的比较

技术方面	MCC	边缘计算
部署方式	集中式	分布式和集中式
到 MD 的距离	远	近
时延	高	低
计算能力	强大	有限
存储能力	强大	有限

1.1 移动边缘计算的发展历程

第一个使计算和存储资源靠近 MD 的概念是 2009 年提出的 Cloudlet[5],它是一个由移动设备、边缘云平台(即 Cloudlet)和集中式数据中心组成的 3 层架构.Cloudlet 的想法是:将具有强大计算能力的服务器放置在无线网络的关键位置,以便为周边的 MD 提供计算资源和存储资源.在 Cloudlet 中提出了一种与 WiFi 热点类似的“计算热点”的概念,由 Cloudlet 代替互联网连接为移动用户提供轻量级的云计算服务.Cloudlet 可以被部署在人口密集的社区场所附近,如购物中心、火车站、展览会等.但是,由于 Cloudlet 并不是移动网络的固有部分,因此在 Cloudlet 的场景下,MD 的服务质量(QoS)很难像 MCC 那样得到保障.

与 Cloudlet 相比,边缘计算中另一个更广为人知的概念被称为雾计算.思科在 2012 年提出了雾计算,以便在网络边缘处理数 10 亿智慧互联设备上的应用[6].因此,雾计算被认为是物联网(IoT)和大数据应用的关键推动因素之一,因为它具备以下 4 个特点[7]:(1) 低时间延迟和位置感知;(2) 地理位置分布广泛;(3) 大量网络节点(例如无线传感器)的互连;(4) 支持流数据传输和实时应用.此外,雾计算的特点可以在许多应用场景中被利用,例如智能电网、智能交通、无线传感器网络等[8−10].雾计算的典型架构通常包含 3 层:云层、雾层和设备层,其中,雾层可以根据实际需求拓展为多个层.雾节点可以是小型基站、WiFi 接入点,甚至可以是用户终端.用户设备选

择最合适的雾节点使用雾计算服务.

从移动用户的角度来看,因 Cloudlet 和雾计算的计算能力没有天然地集成到移动网络架构中,因此当移动用户频繁地在多个小区之间移动时,QoS 和 QoE 难以得到保障.欧洲电信标准协会(European Telecommunications Standards Institute,简称 ETSI)新成立的 2014 行业规范组(Industry Specification Group,简称 ISG)提出了将边缘计算范例集成到移动网络架构中的新概念——移动边缘计算(mobile edge computing,简称 MEC)[11].与 MEC 有关的标准化工作是由著名的移动运营商(如 DOCOMO,Vodafone,TELECOM Italia)和制造商(如 IBM,Nokia,Huawei,Intel)推动的.MEC 被认为是未来 5G 移动网络架构的重要组成部分,其主要目的是实现云计算功能与移动网络的高效无缝集成,并为所有的利益相关者(移动运营商、服务提供商和移动用户)提供便利.根据 ETSI 首次发布的白皮书[12],移动边缘计算被定义为:“移动边缘计算能够在移动设备附近的无线接入网络(ratio access network,简称 RAN)范围内提供 IT 服务环境和云计算能力.”后来,为了适应多种多样的接入技术,MEC 的定义略有扩大[13]:“边缘计算是指一系列广泛的技术,旨在将计算和存储移出远程云(公有云或私有云)并更接近数据源.”图 1 展示了 MEC 的典型架构. MEC 服务器位于基站附近,他们可以处理用户请求,并直接做出响应或将请求转发到 Internet 中的数据中心.

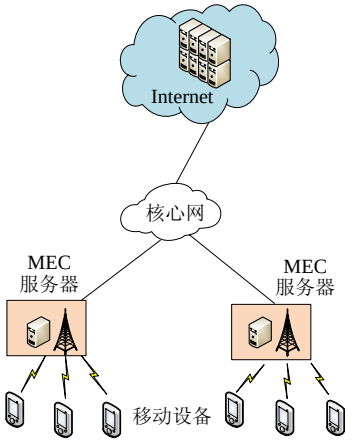

Fig.1 Typical mobile edge computing architecture

图 1 MEC 的典型架构

1.2 移动边缘计算的主要研究范畴

本文对当前移动边缘计算的主要研究工作进行归类和梳理,形成了如图 2 所示的移动边缘计算研究体系.

Fig.2 Mobile edge computing research system

图 2 移动边缘计算研究体系

该研究体系主要包括 4 层:最顶层介绍了移动边缘计算的应用前景,如增强现实、虚拟现实、物联网等;第 2 层则是移动边缘计算中的关键研究问题,主要包括 MEC 架构、计算迁移、边缘缓存和服务编排;第 3 层为关键研究问题对应的技术要点、MEC 架构设计中的系统及服务器管理、计算迁移研究中的迁移决策及资源分配

机制、边缘缓存研究中的数据存储及内容交付技术、服务编排研究中的异构资源协同调度技术;最底层则是支撑移动边缘计算研究的基础技术,如云计算和容器技术、软件定义网络、网络功能虚拟化等.

1.2.1 关键问题

(1) MEC 架构

移动边缘计算的引入,与传统移动网络向 5G 网络的自然演进是同时进行的.MEC 能够在网络边缘实现基于软件的 MEC 应用程序和云计算服务.移动边缘计算并不打算为单个特定的应用程序提供独立的解决方案,与云计算相似,边缘计算的目的也是提供基础设施,以便在用户和设备附近提供计算和存储能力,满足广泛的应用需求.特别是,MEC 实现了模块化、开放式的解决方案,提供了可编程的生态系统,改善了用户体验,同时允许应用服务提供商获取与用户相关的更多信息.因此,在所有的关键研究问题中,MEC 架构设计是最基础的.

(2) 计算迁移

在资源受限的移动设备上运行计算密集型的应用程序会消耗大量的资源和能量,为了解决该问题,计算迁移的概念应运而生[14].通过将移动设备的计算任务/应用程序迁移到网络中的服务器去执行,可以增强移动设备的计算能力,减少移动设备运行应用程序时的能量消耗.为了达到缩短服务时延、节省移动设备能耗的目的,学术界提出了一系列的计算迁移算法.然而,为了设计高效可靠的计算迁移方案,还需要综合考虑移动设备负载、任务属性、网络状态等动态变化的因素.

(3) 边缘缓存

尽管现有的内容交付技术可以优化内容传输服务,提高内容服务器的可用性,减少网络延迟,但是传统的内容交付服务无法跟随用户状态的改变而迅速做出相应的调整.利用 MEC 服务器作为边缘缓存节点,可以基于网络状态、无线信道负载动态地优化内容交付服务.而从移动用户的角度来看,由于 MEC 服务器位于网络的边缘,非常接近移动设备的位置,因此可以结合用户移动性和内容访问日志来优化使用体验.

(4) 服务编排

在移动网络环境中集成 MEC 平台,带来了与服务编排相关的诸多挑战.因为网络中的服务器节点增多,导致系统需要对各个 MEC 服务器的资源(计算资源、存储资源)进行有效的管理.同时,由于用户的移动性,引发了无线电网络资源的动态变化.同时,MEC 应支持应用程序的生命周期管理,即,按需授权第三方应用的实例化或终止应用程序的服务请求.以用户的体验质量和服务可靠性为目标,结合资源管理和服务部署方案来编排 MEC 平台的服务是至关重要的.

1.2.2 支撑技术

(1) 云计算和容器技术

简单地来说,移动边缘计算的概念是将云计算功能延伸到移动网络的边缘.云计算技术的进步,使得在诸如基站和网关等大量通用服务器上部署虚拟机变得更加容易.云计算能够提供强大的处理能力和海量的存储资源.云计算和物联网的整合已被证明有利于提供新的服务[15],因此,将云计算功能集成到移动网络,能够为新兴移动服务的供应和管理提供高效的解决方案.

与传统云计算中的虚拟机(virtual machine,简称 VM)技术不同,新兴的容器技术是一种内核轻量级的操作系统层虚拟化技术.它能够划分物理机的资源,创建多个与 VM 相比尺寸小得多的隔离用户空间实例.由于容器的轻量级特性,其能够在执行应用程序或服务时,提供简单的实例化.借助于容器技术的使用,能够实现 MEC 服务的便携式运行,为移动用户带来很大的便利.此外,由于容器技术提供了快速打包的机制,服务端也能非常方便地将服务部署到大规模互联的 MEC 平台.

(2) 软件定义网络

软件定义网络(software defined network,简称 SDN)技术使边缘网络具有智能化、可编程和易于管理的特点[16].SDN 的主要思想是分离网络的控制面和数据面,它的优势主要包括在通用硬件上创建网络控制面、通过 API 开放网络功能、远程控制网络设备以及将网络智能从逻辑上解耦和为不同的基于软件的控制器.借助 SDN 技术可以实现移动边缘计算平台所需的分层管理[17].

(3) 网络功能虚拟化

网络功能虚拟化(network function virtualization,简称 NFV)技术是未来 5G 网络的重要组成部分,它和 SDN 技术是相辅相成的.NFV 的目的是借助软件编程技术将若干网络功能模块虚拟化,并将其从专用的硬件平台转移到通用计算平台.被虚拟化的网络功能模块可以提供与传统移动网络相同的服务.因此,移动网络的可扩展性和灵活性得到了提高,管理大型异构设备的能力也得到了改善.借助 NFV 技术,网络运营商的经济成本和运营开销可能会减少.NFV 的使用,改变了移动通信行业的格局,并带来了许多益处,如缩短上市时间、实时优化网络配置和拓扑结构、支持多租户共享等[18].

(4) D2D 通信技术

随着移动设备端的功能变得更强大以及更智能,在未来的网络通信中,MD 将发挥更积极的作用.其中一项重要的技术便是 D2D(device to device)通信技术.在许多应用场景中(如游戏和社交网络),距离相近的设备具有共享内容或彼此交互的需求.D2D 通信技术能够在如下几个方面提高网络效率,提高应用的使用体验:首先,它节省了大量信令带宽资源,并减少了传输延迟;其次,与借助基站执行设备间的交互相比,它可以节省大量的能耗;此外,由于路径损耗远低于基站到 MD 之间的通信损耗,D2D 技术可以提高网络信道的频谱效率.

(5) 网络切片技术

网络切片已经成为了实现网络平台灵活化的关键概念,它支持具有不同服务需求的新兴业务在网络平台运行.网络切片技术能够将一个网络整体分割成多个实例,每个实例都针对特定的应用程序和服务进行优化.网络分片的优势在于引入了多租户环境,支持灵活配置网络资源,根据网络功能、无线接入类型(RAT)和应用程序的需求特点的动态分配资源.结合网络切片技术的 MEC 模型,将会在未来的 5G 网络架构中扮演关键的角色.

移动边缘计算研究领域涵盖广泛,涉及移动网络、移动计算、无线网络等多个领域的相关技术.目前,学术界已经发表了数篇与移动边缘计算相关的综述文章.Cloudlet 的作者 Satyanarayanan 等人指出[19]:边缘计算能够为移动计算提供更可靠的云服务,在物联网中提供更高的可扩展性以及隐私保护增强,同时还具备屏蔽云中断的优势.Shi 等人[20]对边缘计算的定义以及典型的案例研究进行了归纳总结,并提出了边缘计算领域的一些新的挑战和机遇.此外,Ahmed 等人[21]通过描述关键属性介绍了 MEC 的分类,对涉及 MEC 的研究工作进行了简要地总结.Roman 等人[22]则展望了各种边缘计算概念中的安全问题.

与上述英文文献的不同之处在于,文本关注近年来移动边缘计算的研究进展,描述了移动边缘计算概念的发展历程,并在第 2 节~第 5 节围绕 MEC 架构、计算迁移、边缘缓存和服务编排这 4 个关键问题的研究成果进行综述.然后,在第 6 节对移动边缘计算中的典型应用场景,如增强现实、虚拟现实、物联网等,进行深入分析.最后,在第 7 节讨论移动边缘计算的发展趋势以及未来的研究挑战,以期对其在国内的研究起到一定的推动作用.

2 MEC 框架及架构

MEC 架构的融入与传统移动网络向 5G 系统的自然演进是同时进行的.为了在网络边缘实现基于软件的移动边缘应用和云计算服务,MEC 实现了模块化、开放式的解决方案,提供了可编程的生态系统,同时允许应用程序提供商获取有关用户的更多信息,以期达到提升用户体验的目的.本节概述了 ETSI 目前为止在 MEC 架构研究方面所做的努力,在第 2.1 节和第 2.2 节介绍了 ETSI 的 MEC 框架及架构.除了 ETSI 的架构之外,还有一些其他团队的研究成果,同样比较新颖和有趣,本文将在第 2.3 节对它们进行综述.

2.1 ETSI MEC框架

ETSI MEC 框架[23]描述了一种生态系统架构,包括所涉及的实体和功能.该架构可分为 MEC 系统层、MEC 服务器管理层和网络层.MEC 服务器管理层是 ETSI MEC 框架的基本组成部分,它包括两个主要部分:MEC 服务器和 MEC 服务器级管理.MEC 服务器提供虚拟化基础设施和移动边缘平台,便于移动边缘应用程序的执行.最底层的网络层实体提供与各种访问的连接,包括 3GPP 移动网络、本地网络和其他外部网络(例如 Internet)访问.在最顶层,MEC 系统级管理提供了基础 MEC 系统的抽象,便于移动设备和第三方的访问.

2.2 ETSI MEC参考架构

ETSI 描述的参考架构由功能元件和允许它们之间相互作用的功能模块组成[23].架构中的功能模块不一定代表移动网络中的物理节点,而是代表在虚拟化基础设施之上运行的软件实体.虚拟化基础架构是指运行虚拟机的物理数据中心,每一个虚拟机表示一个独立的功能元件.在这方面,ETSI NFV 组(与 ETSI MEC 相似)的一些架构特征将在 MEC 参考架构中扮演重要的角色,因为 NFV 的基本思想是将所有网络节点功能虚拟化.

MEC 可以直接由 MD 中的应用程序使用,或者由第三方客户通过面向客户的服务(CFS)端口使用.MD 和 CFS 端口都通过 MEC 系统级管理与 MEC 系统交互.MEC 系统级管理由用户应用程序生命周期管理(LCM)代理、运行支持系统(OSS)和移动边缘编排器组成.LCM 代理是使 MD 能够请求 APP 相关服务的功能,例如实例化服务、终止服务、MEC 平台之间的重定位.OSS 负责做出是否授权用户请求的决策.授权的请求将转发给移动边缘编排器.移动边缘编排器是 MEC 系统级管理的核心功能,它负责维护可用计算/存储/网络资源和 MEC 服务的整体情况.在这方面,移动边缘编排器根据应用程序需求(例如等待时间)将虚拟化 MEC 资源分配给即将启动的应用程序.此外,编排器还可以灵活地缩小/增加正在运行中应用程序的可用资源.

MEC 系统级管理与 MEC 服务器级管理互连.服务器级管理由移动边缘平台和虚拟化平台组成.前者负责管理应用程序的生命周期、应用程序规则、服务授权和流量规则等;后者负责分配、管理和释放位于 MEC 服务器内的虚拟化基础设施提供的虚拟化计算/存储资源.此外,MEC 服务器是参考架构不可或缺的一部分,因为它代表虚拟化资源,并在虚拟化基础架构之上托管作为虚拟机运行的 MEC 应用程序.

除了架构的制定,ETSI MEC 行业支持组织还专注于如何评估 MEC 性能增益的关键指标、最佳实践和指南(与没有 MEC 服务器的传统系统相比)[24].这项工作将为未来的测试和性能测量活动铺平道路.MEC 平台旨在实现对 3GPP 移动网络架构透明,在不影响 MD、RAN 和核心网功能的前提下,减少网络流量的合法拦截.MEC 管理程序应考虑并补充 3GPP 网络管理,确保应用和服务的可移植性.从监管的角度来看,MEC 服务应该是透明的,不得歧视专业服务,即符合网络中立框架,同时允许用户购买受网络中立性影响的专业服务.

2.3 MEC架构的研究现状

除了 ETSI MEC 之外,另一个旨在推动边缘/雾计算架构发展的行业机构 OpenFog,同样为 MEC 架构的落地做出了相应的努力.OpenFog 于 2015 年 11 月推出,其目标是“根据开放标准技术定义分布式计算的通用框架”.其主要目的是借助网络的边缘节点实现开放式计算、控制和数据存储,以充分利用本地分布式云集群的性能优势.OpenFog 使用多个边缘云平台,从信息处理的角度引入逻辑层次结构,同时考虑云平台的联合,定义了雾和云之间以及雾和雾之间的接口[25].

欧盟资助的 SESAME 项目[26]提出了小型蜂窝云(cloud-enabled small cell,简称 CESC)的概念.SESAME 旨在使 CESC 成为多运营商(多租户)实体,这意味着多个网络运营商将能够同时使用 SESAME 平台,每个平台都使用自己的“切片”网络,包括其中的 MEC 功能.该项目提出了一份白皮书[27],其中包含参考模型和架构设计,除了其特征描述和原型的实现外,大多与 ETSI 的 MEC 框架保持一致.

网络功能虚拟化(NFV)被认为是实现 MEC 基础设施的支撑技术之一.Cziva 等人[28]为 MEC 的 NFV 平台提出了格拉斯哥网络功能(Glasgow network functions,简称 GNF).GNF 是基于容器技术的轻量级模块封装,它能够提供快速的实例化建立时间和低资源开销.作者在文中提出了潜在的 MEC 服务器实现规范,将 GNF 与现有的虚拟化网络功能进行了比较,并给出了一个用例.

Rimal 等人[29]考虑了 MEC 架构与现有光纤无线网络的结合.在文章中,作者设置了一些具备无线光纤接入能力的 AP(access point)作为放置 MEC 服务器的候选地点.然后,它提出了一种基于 TDMA 的统一资源管理方案,该方案允许传统的非 MEC 流量和 MEC 相关流量共存.由于 TDMA 的调度性质,该方案允许 MEC 辅助的用户设备在其他设备运行的时间片期间进入睡眠模式.性能评估的结果表明,文中所提出的架构实现了良好的时间响应效率,同时延长了 MEC 辅助设备的电池寿命.

3GPP 正在探索上下文感知服务与 MEC 架构的融合.这项研究工作在 RAN 3 工作组中进行了讨论,以确保

高数据传输速率和低时间延迟[30].3GPP 的工作集中于无线电接入和 MEC 平台之间的接入机制、协议和接口方面.该研究工作的重点是 RAN 的跨层优化与视频感知调度,并提出了移动视频传送优化和本地内容缓存之类的应用.其价值在于探索能够有效利用资源且提升用户体验的用例,分析该类用例对网络协议和相关信令的潜在影响.

3 计算迁移

3.1 计算迁移简介

(1) 计算迁移的一般模型

从节省移动设备能耗、拓展移动设备计算能力的角度来看,关于 MEC 的关键技术是计算迁移.因为计算迁移可以将应用程序中全部/部分的计算任务迁移到MEC服务器去执行,减少了移动设备用于执行计算任务的能耗.同时,由于 MEC 服务器的计算能力远远胜过移动设备,这使得应用程序能够获得更好的性能表现.那么,如何在实践中使用和管理计算迁移?通常,在移动设备端运行的应用程序需要包含代码分析器、系统分析器和决策引擎这 3 个组件:代码分析器的职责是确定应用程序/计算任务是否可以迁移以及哪些部分支持迁移迁移(这取决于应用程序类型和计算任务的特征);然后,系统分析器负责监控各种参数,例如可用带宽、待迁移的数据量或在 MD 本地执行计算任务所消耗的电池能量;最后,决策引擎基于迁移决策算法决定是否执行计算迁移.图 3 所示为移动边缘计算环境下计算迁移的一般模型.

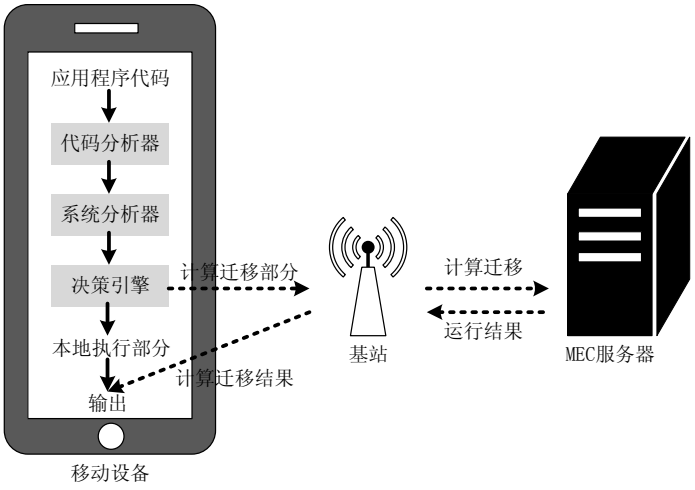

Fig.3 General model for computation offloading in mobile edge computing environment

图 3 移动边缘计算环境下计算迁移的一般模型

(2) 计算迁移的方案

对于移动设备端,计算迁移的关键部分是做出计算迁移决策.如图 4 所示,通常,计算迁移的决策会有以下 3 种方案:(1) 本地执行,所有计算任务在 MD 本地处理;(2) 完全迁移,将所有的计算任务迁移到 MEC,由 MEC 处理计算任务;(3) 部分迁移,一部分计算任务在 MD 本地处理,其余部分则迁移到 MEC 处理.由于 MEC 的计算能力强于移动设备,因此,同样的计算任务在 MEC 上的处理时间更短.图 4 中,我们用灰色和黑色表示同样的计算任务在 MEC 和 MD 上不同的处理时间,灰色表示处理时间较短,黑色表示处理时间较长.

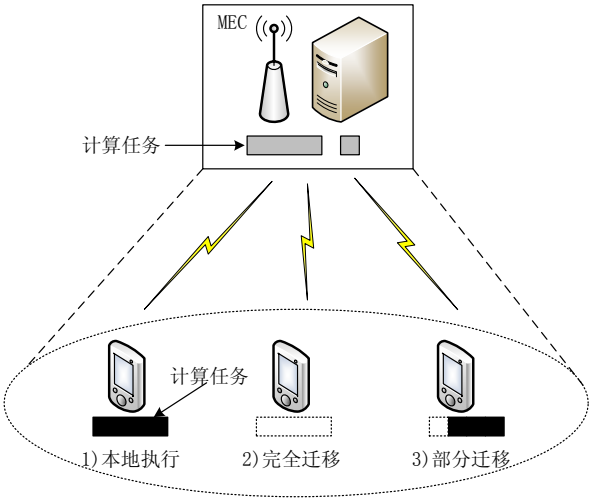

Fig.4 Three schemes of computation offloading decision

图 4 计算迁移决策的 3 种方案

(3) 影响计算迁移决策的因素

计算迁移决策是一个非常复杂的过程,会受到用户偏好、网络连接质量、移动设备性能等因素的影响[4].在这些因素当中,一个重要因素是应用程序的类型,它决定了待迁移的任务是否可以被分割,哪些任务支持迁移到远程去执行以及如何迁移.我们可以按照以下 3 个标准对运行在移动设备上的应用程序进行分类.

- 根据应用的可迁移性.支持迁移的应用可以分为两种类型.第 1 种类型的应用可以被分为多个可迁移的部分,所有的这些部分都可以迁移到远程的服务端去运行.由于每个可迁移部分所需的计算能力和数据量可能不同,因此有必要决定哪些部分应该迁移到 MEC.第 2 种类型的应用则包含多个可迁移的部分和一个不可迁移的部分,不可迁移的部分必须在 MD 本地执行.
- 根据应用执行的连续性.一种是非连续性执行的应用,如人脸识别、病毒扫描等,预先知道待处理的数据量;另一种是连续性执行的应用,如在线交互式游戏,由于无法估计待处理的数据量,更不可能预测该类应用的运行时间.显然,对于连续性执行的应用,计算迁移的决策可能相当棘手.
- 根据迁移任务的并行性.同一应用的各个计算任务之间的关系可以是并行的也可以是串行的:在前一种情况下,迁移到远程执行的各个任务可以同时迁移以及并行处理;在后一种情况下,计算任务之间的关系是相互依赖的,后一个任务的执行必须要等待前一个任务的结果,不适合执行并行迁移.各个任务之间的关系可以用任务依赖图来表示.

(4) 服务质量的衡量

计算资源是移动网络的重要资源.近些年出现了许多计算密集型的应用,如增强现实、高清视频流传输和交互式游戏等.但是,移动设备的计算能力是有限的.尽管计算迁移能够为移动用户的使用体验带来多种多样的有益影响,但不同的应用程序可能有不同的性能需求,如下所示为计算迁移的研究工作中常见的服务质量衡量指标.

- 时延.时延是影响用户体验的重要性能指标.下一代 5G 网络对于时延的需求是 1ms 的往返时间,这比 4G 网络的 10ms 往返时间缩短了近 10 倍[31].对于实时应用程序,将任务/应用迁移到云中心所造成的时延是不可接受的.将计算能力赋予移动网络的边缘设备是一种更可行的方法.
- 能耗.在现有的文献中,已有很多评估移动边缘计算能耗效率的研究工作,提出了多种优化方案以最小化网络侧和移动设备侧的能量消耗.对于未来 5G 网络环境下的计算迁移,应该同时考虑用于计算和传输任务的能量开销.

- 时延和能耗之间的权衡.时延和能耗这两个性能指标,单纯地优化时延忽略移动端的能量消耗,会导致电池电量迅速下降,进而导致 CPU 降频运行,降低用户的使用体验;同理,单纯地优化能耗也会降低用户的使用体验.因此,需要恰当地解决能耗和时延之间权衡的问题.

(5) 研究场景

计算迁移是边缘计算的关键研究问题之一,它能够打破移动设备的资源限制,拓展移动设备的计算能力、电池电量和存储能力等.根据当前的研究现状,从研究场景的角度来看,目前已提出的各种迁移决策方法可以划分为单移动设备场景和多移动设备场景.

- 单移动设备场景.在单移动设备场景下,影响计算迁移决策的因素主要有计算任务队列长度、MD 本地计算单元的执行状态、传输单元的状态等.MD 端的决策引擎需要收集与这些因素相关的数据,并且基于这些数据对应用的执行时延和能量消耗做出预测,最终决定是否执行计算迁移.我们将在第 3.2.1 节重点综述计算迁移问题在单移动设备场景下的研究工作.
- 多移动设备场景.多移动设备场景下的计算迁移问题要比单用户场景下更加复杂.在多移动设备场景下,影响迁移决策的因素要更加复杂一些,因为网络带宽资源、MEC 计算资源、移动设备数量都是在动态变化的.我们将在第 3.2.2 节重点综述计算迁移问题在多移动设备场景下的研究工作.

3.2 用户端的迁移决策研究

从移动用户的角度出发,关于计算迁移研究的核心是如何做出合适的迁移决策.我们将在下面的第 3.2.1 节和第 3.2.2 节总结不同场景(单移动设备场景和多移动设备场景)下计算迁移决策的研究现状.

3.2.1 计算迁移技术在单移动设备场景下的研究

将计算任务迁移到 MEC 所带来的一个优势是可以减少应用的运行时延或节省 MD 的能量消耗.在 MD 不执行计算迁移的情况下,运行时延仅包含应用程序在 MD 本地处理所花费的时间.在 MD 执行计算迁移的情况下,应用的运行时延包括以下 3 个部分:1) 迁移任务到 MEC 的传输时间;2) 任务在 MEC 上的计算时间;3) 接收任务计算结果的时间.同样地,对于应用程序运行期间所消耗的能量:如果 MD 不执行计算迁移,应用运行期间的能耗就仅包括 MD 端的本地执行能耗;如果 MD 执行计算迁移,应用运行期间的能耗就要包括本地执行能耗和传输能耗.

本节总结了单移动设备场景下计算迁移决策问题的研究.我们从迁移方案的角度出发,将单移动设备场景下已有的研究工作分为两类:一类是完全迁移方案,另一类是部分迁移方案.

(1) 完全迁移方案

早期的完全迁移方案大多以优化时延为主要目标.Liu 等人[32]提出了一种基于一维搜索算法实现的迁移决策方法,该算法根据应用缓冲区的队列状态、MD 和 MEC 服务器处可用的计算能力以及 MD 和 MEC 服务器之间的信道特征来找到最优的迁移决策.是否执行计算迁移是由 MD 的计算迁移决策模块完成的,该模块在每个时间片的开始决定在缓冲器中等待的任务是在 MD 本地处理还是迁移到 MEC 处理.作者在文中将该算法的性能与本地执行策略(计算任务默认在本地运行)、MEC 端执行策略(计算任务默认在 MEC 服务器运行)、贪婪迁移策略做了对比实验.仿真结果表明,与本地执行策略相比,文中所提出的最优迁移策略能够减少高达 80%的执行时延;与 MEC 端执行策略相比,能够减少大约 44%的执行时延.该方法的缺点是 MD 需要基于 MEC 服务器的反馈来做出迁移决策,但是文中并没有讨论 MEC 服务器发送反馈到 MD 所产生的通信开销.

Mao 等人[33]提出了另一种基于 Lyapunov 优化的动态计算迁移算法(LODCO),旨在优化应用的执行时延.与之前的研究相比,该文献中假设 MD 使用能量收集技术[53]来最小化本地执行期间的能量消耗,并且使用功率控制的方法优化用于数据传输的能量开销.LODCO 在每个时间片内做出计算迁移决策,如果任务在 MD 本地执行,则 MD 要为计算任务分配 CPU 周期;如果任务被迁移到 MEC 服务器执行,则 MD 要为计算任务分配传输功率.文中所提出的 LODCO 算法能够通过迁移到 MEC 服务器的方式减少高达 64%的执行时延.此外,该算法能够完全防止计算迁移过程中出现数据包丢失的情况.

除了时延之外,能耗也是影响 MD 端做出计算迁移决策的重要因素.Kamoun 等人[34]提出了一种在满足应

用执行时延的同时最小化 MD 端能量消耗的计算迁移决策算法.作者在文中将该优化问题形式化为一种受约束的马尔可夫决策过程.为了解决该优化问题,文中引入了两种资源分配策略:第 1 种是基于在线学习的策略,网络资源针对 MD 端运行的应用动态地进行适配;第 2 种是基于预先计算的离线策略,它利用了与应用相关的一些数据(如任务到达速率、无线信道状态等).实验表明,基于预先计算的离线策略在低等和中等任务到达速率的情况下优于在线学习策略高达 50%.Kamoun 等人的算法显示了离线资源分配策略的优点,Labidi 等人[35]则提出了另外两种动态的离线迁移策略:确定离线策略和随机离线策略.结果表明,与计算任务在 MD 端本地执行或计算任务完全在 MEC 端执行相比,离线迁移策略可以节省更多的能耗.

综合考虑时延和能耗这两个性能指标,Chen 等人[36]提出了一种权衡 MD 端时延和能耗的计算迁移策略.文中假定:如果 MEC 服务器的计算资源无法满足计算任务的需求,则计算任务也可以被迁移到远程云中心(CC).计算迁移决策以顺序方式完成:在第 1 步中,MD 决定是否将计算任务迁移到 MEC 服务器,如果计算任务被迁移到 MEC 服务器,则 MEC 在第 2 步中评估自己的剩余计算资源是否能够满足任务的需求.作者将该问题表述为一个非凸二次约束二次规划问题.这是一个 NP 问题,因此文中提出了一种半自定义的随机启发式算法.与计算任务总是在 MD 端执行或总是在 MEC/CC 端执行相比,该算法能够显著降低系统的总体开销(能耗和时延的加权和).

(2) 部分迁移方案

上述文献的研究工作都属于完全迁移方案的范畴,没有考虑计算任务/应用程序的可分割性.本节综述了单移动设备场景下部分迁移方案的研究工作.根据待迁移的计算任务/应用程序的可分割性,部分迁移方案一般可以分为两种类型:如图 5(a)所示为第 1 种类型的应用程序,它可以分为多个可迁移部分,所有的这些部分都可以迁移到 MEC 服务器执行;如图 5(b)所示为第 2 种类型的应用程序,它由一个不可迁移部分和多个可迁移部分组成.

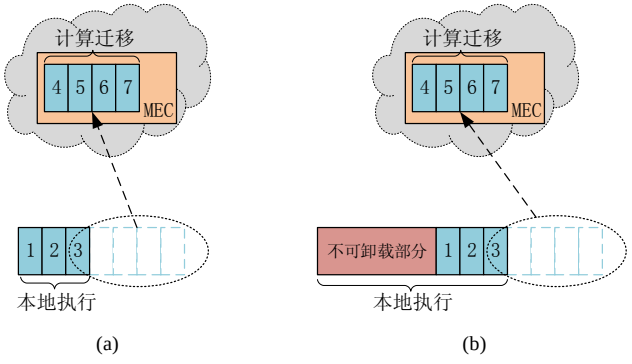

Fig.5 Examples of partial offloading

图 5 部分迁移方案示例

Cao 等人[37]考虑了计算任务/应用程序包含 1 个不可迁移部分和多个可迁移部分的情况(如图 5 所示).作者在文中提出了一种基于组合优化方案的最优自适应算法.为了降低最优算法的复杂度,作者在文中还提出了一种次优算法.实验证明:最优算法能够节省高达 48%的能耗;而次优算法虽然效果略微低于最优算法,但依然能够节省 47%的能耗.此外,作者在文中表明,随着 MD 和基站之间信噪比的增加,能耗的降低会更加明显.

如果计算任务是可分割的,那么当任务之间有依赖关系的时候,部分迁移决策的过程会更加复杂一些.Deng 等人[38]假设在 MD 上运行的应用由若干个有依赖关系的部分组成,即某一部分只有在等待其他部分的执行结果之后才能开始运行.作者在文中将该迁移问题定义为 0-1 模型,并提出了一种二进制粒子群优化算法(BPSO)以解决该问题.实验结果显示:与完全迁移决策相比,部分迁移决策可以为移动设备节省更多的能耗.

Munoz 等人[39]针对计算迁移过程中能耗和时延之间的权衡问题提出了一种部分迁移方案.迁移决策会受

到以下参数的影响:(1) 待处理任务的数据量;(2) MD 和 MEC 的计算能力;(3) MD 和基站之间的信道状态;(4) MD 的能量消耗.在文中,计算迁移决策被形式化为通信和计算资源分配的联合优化.仿真结果表明,MD 端的能耗会随着总执行时延的增加而下降.然而,随着时延的增加,MD 端节省的能耗会变得无关紧要,因为应用程序往往是时延敏感型的.此外,作者指出:如果通信信道的质量很差,是不适合执行计算迁移的,因为大量的能量会消耗在传输计算任务上.在这种情况下,应用程序应优先在 MD 本地处理.对于中等通信质量的信道,可以迁移一部分计算任务到 MEC,因为这可以节省 MD 端的能量消耗.最后,如果信道质量较高,则计算迁移可以大大节省应用运行期间的能量消耗,所以优先选择迁移计算任务到 MEC.Sehati 等人[40]对文献[39]中的能耗和延迟之间的权衡研究提供了更深入的理论分析.作者进一步表明,对于良好的信道质量,执行计算迁移的收益 更高.

在上述所有关于迁移决策的论文中,MD 迁移到 MEC 的数据传输请求都是被单独看待的.与上述论文的不同之处在于,Sehati 等人[41]考虑在应用运行期间在 MD 端将相同的迁移请求聚合,然后一并发给 MEC 服务器.尽管聚合迁移请求会产生额外的时延,但是这样可以节省传输重复数据的能耗.文中将迁移请求聚合问题形式化为一个成本最小化问题,其中,功耗和时延之间的权衡由成本函数衡量.之后,作者提出了一个用于最小化聚合成本的在线算法,并表明,该算法与已知请求顺序的最优离线算法相比,在节省能耗的同时还不会影响到用户的 QoE.

3.2.2 计算迁移技术在多移动设备场景下的研究

本节总结了多移动设备场景下迁移决策问题的研究.第 3.2.1 节所述的单移动设备场景是一种偏向理想化的场景,因为在真实的世界中,MEC 服务器能够同时为大量的移动设备提供类似于云计算的服务.本节总结了多移动设备场景下计算迁移决策问题的研究.我们依然从迁移方案的角度出发,将多移动设备场景下已有的研究工作分为两类:一类是完全迁移方案,另一类是部分迁移方案.

(1) 完全迁移方案

针对多移动设备场景下的迁移决策问题,Labidi 等人[42]提出了两种基于动态规划的算法:一种是离线的,一种是在线的.其研究工作的主要目的是在保证 MD 之间迁移决策公平性的前提下,优化资源调度策略和计算迁移策略.实验结果表明,离线策略在降低能耗方面明显优于在线策略.此外,作者在文中表明,单个 MD 的能耗会受到运行在其他 MD 上的应用需求的影响.

Barbarossa 等人[43]提出了另一种多 MD 场景下的迁移决策策略,旨在满足应用的执行时延的同时最小化 MD 处的能量消耗.作者在文中将时间划分为一个个均匀的时间片.在每个时间片期间,所有的 MD 被分成两组,分别做出计算迁移决策.其中,当第一组的 MD 将计算任务迁移到 MEC 时,第二组中的 MD 必须执行本地计算.对于执行计算迁移的第一组 MD,通过寻找单个 MD 的最佳传输功率并且将 MEC 的计算资源分配给组内所有的 MD 来执行通信和计算资源的联合分配.最后,基于计算任务的平均队列长度以及 MD 和基站之间可用天线数量来评估文中提出策略的性能.实验结果表明,使用的天线数量越多,MD 处的传输能耗越小.

上述文献只讨论了单个基站的场景,连接到不同基站之间的 MD 不存在干扰.因此,Sardellitti 等人[44]考虑了具有多个基站的多蜂窝小区场景,以反映真实的网络部署.该问题在多蜂窝小区场景下不再是一个凸优化问题,因此,作者提出了一种基于连续凸逼近方法的分布式迭代优化算法.实验结果显示,文中所提出的无线电资源和计算资源联合优化方法显著优于单独优化无线电资源或计算资源的方法.此外,作者在文中表明,对于数据量小同时又需要大量 CPU 周期来处理的应用程序,更适合执行计算迁移.这样,用于迁移计算任务所花费的能耗开销才会低于执行计算迁移所产生的能耗收益.

除了上述方法之外,基于分类的方法也具有很好的性能表现.Zhang 等人[45]提出了一种基于分类的节能计算迁移算法(EECO).该算法可以分为 3 个阶段:在第 1 阶段中,根据用于计算的时间成本和能量成本将 MD 分为应当执行计算迁移的 MD、应当在本地计算的 MD 和可以执行计算迁移或本地计算的 MD;在第 2 阶段,通过它们对通信信道和计算能力的需求确定第 1 类和第 3 类 MD 的迁移优先级;在第 3 阶段,基站根据给定的优先级为 MD 分配无线电资源.根据实验的数值结果,与不执行计算迁移相比,EECO 能够节省约 15%的能耗.此外,可以证明:随着 MEC 计算能力的增加,决定执行计算迁移的 MD 数量也会增加.

Chen 等人[46]提出了一种多用户多信道环境下能耗和时延之间权衡的计算迁移决策.该算法由权重参数决定最终的迁移决策倾向于优化时延还是优化能耗.其主要目的有两个:(1) 根据权重参数决定 MD 是否应该将计算任务迁移到 MEC;(2) 在 MD 做出迁移决策的情况下,选择最合适的无线信道用于数据传输.为此,作者提出了一个最佳的集中式解决方案.但是多用户多信道环境下的计算迁移决策问题是一个 NP 难问题.因此,作者又提出了一种实现纳什均衡的分布式计算迁移算法.作者对集中式解决方案和分布式解决方案在如下两个性能指标方面进行了比较:(1) 在保持系统整体利益的前提下,执行计算迁移的 MD 数量;(2) 通过能耗和时延的加权和表示的计算开销.实验数据表明,在上述两个性能指标中,分布式算法的性能比集中式算法略差.此外,在所有的 MD 都选择本地执行或所有的 MD 都选择计算迁移的情况下,分布式算法的性能大大优于集中式算法.

针对多个 MD 连接到同一个基站时,MD 端计算迁移决策和服务端计算和通信资源分配的联合优化问题,Chen 等人[47]提出了一种 3 步算法,包括半定松弛(SDR)、交替优化(AO)和顺序调整(ST).该算法能够计算一个局部最优的解决方案,并且在广泛的参数设置下给出接近最优的性能.通过评估 SDR-AO-ST 算法中 3 个组成部分不同组合的性能表现,作者得出它们在整体解决方案中的作用和贡献.数值结果证明了所提出的算法在计算资源和通信资源的联合优化中是有效的.

(2) 部分迁移方案

针对多移动设备场景下的部分迁移决策问题,Zhao 等人[48]进行了详细的讨论.作者在文中假设 MD 可以分割计算任务,并且能够决定将多少子任务迁移到 MEC.该问题被形式化为一个高复杂度的非线性约束优化问题.然后,作者将该问题简化为一个可解决的线性规划问题.如果采用基于穷举搜索的最佳解决方案,与不执行计算迁移的情况相比,可节省 40%的能耗.如果采用启发式算法,则可以节省 MD 端 30%的能耗.该研究工作的缺点在于,作者假设系统中的 MD 具有相同的信道质量,并且具有相同的计算能力.这些假设对于真实的网络来说显然是不现实的.

You 等人[49]在基于 TDMA 的系统中对部分迁移问题进行了进一步的讨论,系统中的时间被划分为持续时间为 T 秒的均匀时间片.在每个时间片期间,MD 可以根据其信道质量、本地计算的能量消耗以及 MD 之间的公平竞争性,将其计算任务的一部分迁移到 MEC.文中定义了一个基于阈值的最佳资源分配策略,该策略为每个 MD 做出二元化的迁移决策:如果 MD 具有高于阈值的优先级,则 MD 执行完全计算迁移(将所有的计算任务迁移到 MEC);如果 MD 具有比阈值更低的优先级,则它仅迁移很小一部分的计算任务以满足应用程序的时延需求.由于通信和计算资源的联合最优分配具有很高的复杂度,作者在文中又提出了一种次优的资源分配策略.仿真结果表明,与最优分配策略相比,次优的方案给 MD 增加的能量消耗可以忽略不计.You 等人[50]扩展了自己的研究工作,使用 OFDMA 接入代替 TDMA 系统.作者表明,与 TDMA 系统相比,由于无线电资源的粒度变粗,OFDMA 接入能够节省大约 10 倍的能耗.

Munoz 等人[51]探讨了部分迁移决策问题中时延和能耗之间的权衡.在多 MD 的场景下,由于 MEC 所提供的通信和计算资源被多个 MD 共享,所以影响部分迁移决策的因素都是在动态变化的.作者在文中证明:系统中 MD 数量越多,用于传输计算任务的时间也就越长;同时,在 MEC 中处理计算任务也需要更长的时间.产生这种现象的原因是显而易见的,因为每个 MD 分配到的通信资源和计算资源都较少.但是,作者所提出的部分迁移决策算法依然可以节省多达 90%的能耗.

针对部分迁移决策中时延和功耗之间权衡的问题,Mao 等人[52]使用了缓冲区稳定性约束来制定能耗最小化问题.作者在文中提出了一种基于 Lyapunov 优化的在线算法,以确定执行本地运算时最优的 CPU 频率,并且将传输功率和通信带宽优先分配给执行计算迁移决策的 MD.文中所提出的算法能够根据优先级来控制功耗和时延之间的权衡.文献的模拟实验表明,借助 MEC 执行计算迁移可以降低 MD 端大约 90%的能耗,应用的运行时延大约减少 98%.

3.2.3 小 结

表 2 对上述计算迁移方案进行了对比分析.大多数计算迁移决策算法的目的都是在满足被迁移应用程序执行时延的同时优化 MD 处的能耗,或者是在能耗和时延之间找到适当的权衡.已有的文献表明,通过将计算任

务/应用程序迁移到 MEC 可以节省高达 90%的能耗,同时可能将执行时延降低 98%.此外,几乎所有的论文都是通过仿真实验的方式来评估所提出的计算迁移方案.

Table 2 Comparison of existing papers addressing computation offloading decisions

表 2 计算迁移决策方法的对比分析

文献	迁移方案	目标	所提出的方法	场景	实验方法	效果
[32]	完全迁移	优化时延	一维搜索算法	单 MD	模拟	时延缩短 80%
[33]	完全迁移	优化时延	基于 Lyapunov 优化的算法	单 MD	模拟	时延缩短 64%
[34]	完全迁移	优化能耗	基于在线学习和离线预测的算法	单 MD	模拟	能耗节省 78%
[35]	完全迁移	优化能耗	确定性和随机性的离线决策算法	单 MD	模拟	能耗节省 78%
[36]	完全迁移	权衡时延和能耗	基于半定松弛随机映射的启发式算法	单 MD	模拟	总开销降低 70%
[37]	部分迁移	优化能耗	基于组合优化方法的自适应算法	单 MD	模拟	能耗节省 47%
[38]	部分迁移	优化能耗	基于粒子群优化的算法	单 MD	模拟	能耗节省 25%
[39]	部分迁移	权衡时延和能耗	通信和计算资源的联合优化	单 MD	模拟	N/A
[42]	完全迁移	优化能耗	基于决策后学习的确定性在线算法	多 MD	模拟	N/A
[43]	完全迁移	优化能耗	通信和计算资源的联合优化	多 MD	模拟	N/A
[44]	完全迁移	优化能耗	基于凸优化的分布式迭代算法	多 MD	模拟	N/A
[45]	完全迁移	优化能耗	EECO 算法	多 MD	模拟	能耗节省 15%
[46]	完全迁移	权衡时延和能耗	基于博弈论的计算迁移决策算法	多 MD	模拟&真实	能耗节省 40%
[47]	完全迁移	权衡时延和能耗	基于半定松弛随机映射的启发式算法	多 MD	模拟	总开销降低 45%
[48]	部分迁移	优化能耗	基于任务调度的算法	多 MD	模拟	能耗节省 40%
[49]	部分迁移	优化能耗	基于阈值的资源优化分配策略	多 MD	模拟	N/A
[51]	部分迁移	权衡时延和能耗	通信和计算资源的联合优化	多 MD	模拟	能耗节省 90%
[52]	部分迁移	权衡时延和能耗	基于 Lyapunov 优化的算法	多 MD	模拟	时延缩短 98%

3.3 服务端的资源分配研究

如果移动设备端已经做出了计算迁移决策,那么服务端(即 MEC 服务器)必须对 MEC 资源进行适当的分配.与计算迁移决策相似的地方在于,MEC 资源的分配同样会受到任务并行性的影响.如果被迁移的计算任务/应用程序是不支持并行计算的,则只能分配一个物理节点执行计算任务.而如果被迁移的计算任务/应用程序是支持并行执行的,那么可以通过多个 MEC 节点合作的方式来处理迁移的任务/应用.我们将在第 3.3.1 节和第 3.3.2 节分别综述有/无云中心参与场景下 MEC 资源分配问题的研究现状.

3.3.1 有云中心场景下 MEC 资源的分配

在有云中心参与的场景下,若 MEC 节点的计算资源不足,可以通过 MEC 和云中心协作的方式来提高移动边缘计算的服务质量.Zhao 等人[54]设计了一种基于阈值的协同调度策略,使得 MEC 服务器在满足应用程序时延需求的同时,最大化 MEC 中正在运行应用程序的数量.在 MD 做出计算迁移决策之后,由应用程序的优先级和 MEC 服务器中计算资源的可用性决定被迁移的应用程序应该放置在哪里(云中心或 MEC).服务端资源分配的基本示例如图 6 所示.

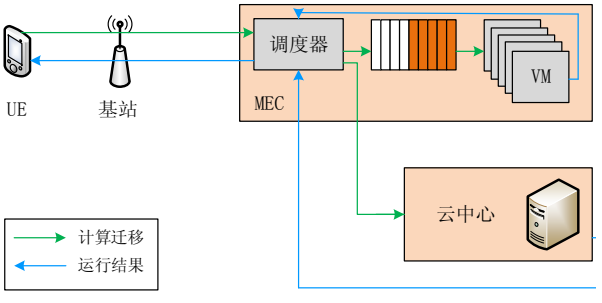

Fig.6 An example of MEC resource allocation

图 6 MEC 资源分配示例

迁移的应用程序首先交付给 MEC 内的本地调度器.调度器检查本地 MEC 节点是否有足够的计算资源:如

果 MEC 节点具有足够的可用资源,则将运行着该应用程序的虚拟机分配给该 MEC 节点;但是,如果 MEC 服务器提供的计算能力不足,则调度器将该应用程序委托给远程的云中心.为了最大化在 MEC 中处理的应用程序的数量同时满足其延迟要求,作者提出了基于优先级的合作策略,为每个优先级级别定义了几个缓冲区阈值.因此,如果缓冲区已满,应用程序将被发送到云中心.

Zhao 等人的研究工作是在单个 MEC 场景下分配计算资源.Tanzil 等人[55]则针对多个 MEC 场景下的资源分配问题开展了研究.作者在文中通过构建 MEC 集群的方式完成资源分配,以达到缩短执行时延的目的.MEC 集群的构建通过合作博弈的方法完成,并在固定的时间周期之内进行 MEC 集群的重组.当集群内的某一 MEC 无法单独处理计算任务时,计算任务才被转发到同一集群中的所有 MEC.数值结果表明,与单一 MEC 服务器提供计算资源相比,文中所提出的 MEC 集群方案能够将执行延迟减少高达 50%.但是,文中并没有解决新的集群如何重组的问题.

3.3.2 无云中心场景下的资源分配

在无云中心的场景下,服务端的资源分配只能在 MEC 服务器之间进行.一方面,从移动设备的角度出发,资源分配方法中最重要的问题是选择合适的 MEC 服务器执行计算迁移;另一方面,站在 MEC 服务器的立场上,如何合理地构建 MEC 集群才能达到系统整体性能的优化,是服务端需要重点考虑的问题.

Guo 等人[56]假设存在一个覆盖大量 MD 的热点区域,区域内的每一个 MD 都能够通过附近的基站访问多个 MEC 服务器.作者在文中提出了一个基于离散马尔科夫决策过程(MDP)框架的最优任务调度策略.然而,随着 MEC 服务器数量的增加,这种方法会引发高通信开销和高计算复杂度.因此,又在文中提出了一个基于索引的分配策略以降低算法复杂度和通信开销.每一个 MEC 服务器根据其计算资源的剩余状态生成自己的索引.然后在网络中广播该索引,这样 MD 就能够选择合适的 MEC 服务器,以期达到最小化执行延迟和能耗的目的.实验结果显示:在系统成本方面,基于索引的策略比最优调度策略的成本高了 7%;在系统性能方面,延迟和能耗之间的权衡可以通过系统成本中的权重参数灵活调整.

无云中心的场景下,MEC 服务器需要通过协同工作的方式完成资源的合理分配.在构建 MEC 集群时,节点的选择不仅会显着影响服务时延,还会影响计算节点的能耗.因此,Oueis 等人[57]的主要目标是分析集群规模(即执行计算任务的 MEC 服务器数量)对服务时延和 MEC 服务器能耗的影响.文中针对不同的回程网络拓扑(环形、树形和全网状)和传输媒介(光纤、LTE)进行了分析.实验表明,使用光纤作为传输媒介的全网状拓扑在执行时延方面是最有优势的(执行时延减少了 90%);使用光纤作为媒介的环形拓扑在能耗方面性能最佳.此外,论文表明,增加 MEC 服务器的数量并不总是能缩短执行延迟;恰恰相反,如果传输时延比 MEC 服务器端的计算时延更长,整体服务时延可能会增加.此外,随着 MEC 服务器数量的增加,系统整体能耗也会增加.因此,适当的 MEC 集群构建方法和 MEC 节点选择方法在系统性能中起着至关重要的作用.

针对 MEC 集群的构建方法,Oueis 等人[58]提出了多种解决方案,旨在验证不同的聚类策略对 MEC 集群特性(规模、时延、能耗)的影响.文中提出了 3 种不同的聚类策略.

- 第 1 种聚类策略以优化服务时延为目标选择 MEC 服务器.由于系统模型中的所有 MEC 服务器被假设为相互连接(即全网络拓扑),所以基本上所有 MEC 服务器都可以提供计算资源.由于计算增益远大于传输开销,这导致系统总体的服务时延降低了 22%.
- 第 2 种聚类策略的目标是优化集群的总体能耗.在这种情况下,每一个 MD 迁移的计算任务仅仅选择一个最优的 MEC 服务器执行计算,因此,减少了 MEC 服务器之间用于数据传输的开销(能耗降低 61%).然而,第 2 种聚类策略会增加系统整体的服务时延,同时会导致 MEC 服务器之间负载失衡.
- 最后一种聚类策略则是将优化集群中每个 MEC 服务器的能耗作为目标,以解决第 2 种聚类策略中 MEC 服务器之间负载失衡的问题.

上述研究工作的缺陷在于,构建 MEC 集群的策略仅考虑了服务端,忽略了移动设备端的需求.Oueis 等人[59]假设存在一个多 MD 的场景.集群的构建与 MD 端的任务调度一起完成.文中提出的资源分配过程包括两个步骤:第 1 步是本地计算资源分配,每个 MEC 服务器根据特定的调度规则(如时延约束、计算负载、任务需求等)

将其计算资源分配给附近的 MD;第 2 步是 MEC 集群的建立,为第 1 步中没有分配到计算资源的 MD 创建 MEC 集群.基于应用程序的优先级和性能目标,作者设计了 3 种不同的算法.仿真结果表明,文中所提出的方法能够满足 95%用户的需求;同时,所有 MEC 服务器的能耗都处于可接受的范围内.

3.3.3 小 结

表 3 对上述 MEC 资源分配方案进行了比较.对 MEC 资源进行合理分配的目的主要是优化被迁移应用程序的执行时延.换句话说,资源分配的目标是保障 MD 的 QoE,以充分利用 MEC 服务器靠近移动网络边缘的优势.此外,一些研究工作还着眼于优化 MEC 节点的能耗,或致力于平衡服务端计算和通信的负载,以满足 MD 对服务质量的需求.所有已提出方案的共同缺点是缺乏真实实验场景的验证,同时忽略了用户的移动性.如果 MD 正在朝着远离 MEC 服务器的方向移动,那么由于传输时延变长可能导致用户的 QoE 下降.

Table 3 Comparison of existing papers addressing allocation of MEC resources

表 3 MEC 资源分配方法的对比分析

文献	目标	所提出的方法	场景	实验方法	效果
[54]	最大化提供服务的应用程序数量	基于优先级的合作策略	有云中心	模拟	时延缩短 25%
[55]	优化时延&避免使用云中心	基于激励措施的合作策略	有云中心	模拟	时延缩短 50%
[56]	优化时延&优化 MEC 能耗	基于等效离散 MDP 框架的策略	无云中心	模拟	N/A
[57]	优化时延&优化 MEC 能耗	3 种聚类策略	无云中心	模拟	时延缩短 22%
[58]	优化时延&优化 MEC 能耗	为所有活动用户构建群集	无云中心	模拟	提高了用户满意率
[59]	优化时延&优化 MEC 能耗	同时进行集群构建与用户调度	无云中心	模拟	提高了用户满意率

4 边缘缓存

计算迁移和资源分配的研究工作主要是为了解决如何高效使用 MEC 计算资源的问题.而如何有效使用 MEC 服务器的存储资源,存在着很多难点有待进一步攻克.传统的以内容为中心的网络缓存方案没有考虑流量负载的动态变化等移动网络特性.在本节中,我们将综述移动边缘计算环境下针对边缘缓存问题的研究工作,主要包括边缘缓存的性能目标、缓存内容流行度的衡量方法、缓存策略.

4.1 边缘缓存的性能目标

边缘缓存能够带来的有益之处是多种多样的.不同的应用程序或系统可能有着不同的性能需求,如下所示为边缘缓存的常见性能目标.

(1) 系统整体容量.现有的边缘缓存方面的工作已经证明:在网络边缘缓存热门的内容可以显著提高系统整体容量.例如,Ahlehagh 等人[60]提出的解决方案与不具备缓存能力的移动网络相比,可以将系统整体容量提高 3 倍.

(2) 时间延迟.由于边缘节点与移动设备之间的距离很近,利用网络边缘的 MEC 服务器执行内容缓存可以显著减少内容传输延迟.Tandon 等人[61]通过联合调度 RAN(无线电接入网络)回程和无线信道资源来减少视频会话的初始延迟和卡顿,提升了视频流实时传输型应用的 QoE.

(3) 能耗效率.能耗效率是未来 5G 网络的另一个重要性能指标.Liu 等人[62]分析了边缘缓存对下行链路网络能耗效率的影响.结果表明,当文件目录规模较小时,能耗效率将得到提高,并且借助多个小容量的边缘节点执行边缘缓存,比在单个大容量的边缘节点上执行边缘缓存更节能.Zhang 等人[63]研究了软件定义的异构网络中能耗效率和小区密度之间的权衡,实验结果表明,支持边缘缓存的异构网络远高于当前 LTE 网络的能耗效率.

4.2 内容流行度的衡量

MEC 具备在网络的边缘提供存储资源的能力.为了决定在 MEC 服务器中要缓存什么内容,应该考虑内容的流行度,尽量最大化边缘缓存的命中率,即用户的内容请求在 MEC 服务器的缓存里命中的概率.我们将现有文献中衡量内容流行度的模型分为静态模型和动态模型两种.

(1) 静态模型.目前,大多数关于边缘缓存的研究工作都假设内容流行度是静态的,并采用独立参考模型:内容请求是基于独立的泊松过程产生的,该过程的速率与基于二八法则的内容流行度相关[64].常用的流行度模型是在 Web 缓存中观察得到的 Zipf 分布[65].

(2) 动态模型.静态模型无法反映随着时间的流逝而发生变化的真实内容流行度.Traverso 等人[66]提出了被称为散粒噪声模型(SNM)的动态流行度衡量模型.该模型使用具有两个参数的脉冲来模拟每个内容,脉冲持续时间反映了内容的流行周期,脉冲高度反映了内容的瞬时流行度.Cha 等人[67]分析了用户生成内容(UGC)流行度分布的统计特征,并讨论了利用“长尾”视频需求的机会.

4.3 缓存策略

在现有的边缘缓存研究中,已经提出了多种不同的缓存策略和算法.大多数的研究工作都是根据移动网络的特性,对传统的有线网络中的一些缓存策略进行了修订.此外,还提出了一些新的缓存方案,如基于用户偏好、增强学习或多节点合作的缓存策略.

传统的缓存替换策略如最近最少使用(LRU)和最近最少访问频次(LFU),已经被大量的研究工作采用[68,69].对于相同规模的内容,这两种策略简单而且高效.但是,它们会忽略内容的下载时延以及内容的数据量.内容交付网络中常用的另一种主动式缓存策略是 MPV 策略,该策略根据全球视频流行度的分布来缓存最受欢迎的视频内容[60].但是与内容交付网络相比,移动网络的高速缓存大小非常有限,这导致 MPV 策略实现的缓存命中率在移动网络环境下太低.

Ahlehagh 等人[60]提出了一种基于用户偏好的缓存策略.根据观察显示:区域视频流行度与全国视频流行度是显著不同的,不同区域的用户群体可能对特定的视频类别表现出强烈的偏好.作者在文中将用户偏好定义为每个用户请求特定视频类型的概率.仿真结果表明,基于用户偏好的缓存策略能够有效提高视频请求初次命中的概率.即使是在无线信道带宽受限的网络中,文中所提出的缓存策略依然有着非常不错的性能表现.

事实上,内容流行度是随时间变化的,并且是无法事先预知的.因此,对于内容流行度的及时跟踪和估计是一个重要问题.基于机器学习技术,Sengupta 等人[70]提出了一种基于增强学习的边缘缓存策略.该策略从增强学习的角度出发,解决了移动边缘网络中的分布式缓存问题.通过编码缓存的方法,将缓存问题简化为与网络连接性相关的线性规划问题.仿真结果表明,采用编码的缓存方案比未编码的方案执行得更好.

早期的边缘缓存策略研究通常都是基于非合作方式的策略.Ahlehagh 等人[60]提出的方案基于特定小区中活动用户的偏好配置文件执行缓存决策,不考虑其他小区中缓存内容的影响.Gu 等人[71]提出了一种基于学习的算法以解决 MEC 节点的缓存替换问题.作者在文中将该问题形式化为马尔科夫决策过程,以分布式的方式解决了缓存替换的问题,无需在 MEC 节点之间交换与缓存数据相关的额外信息.与传统的 LFU 和 LRU 策略相比,该策略的性能更佳.

随着边缘缓存策略研究的深入,研究工作者们开始考虑通过 MEC 服务器协作的方式来提高算法的性能表现.Borst 等人[72]提出了一种轻量级的协作式缓存管理算法,以最大化缓存服务的流量,同时最小化带宽成本.Jiang 等人[73]研究了 MEC 服务器之间的协作方案,以优化 MEC 和移动设备之间内容缓存和交付的性能表现.协作缓存问题被形式化为整数线性规划问题,并通过次梯度优化的方法来解决;内容传输策略被形式化为不平衡分配问题,并使用 Hungrian 算法来解决.Yu 等人[74]探索了可伸缩视频编码(SVC)技术在协作视频缓存和小区间调度中的应用,以进一步提高系统整体的缓存容量和 QoE.Wang 等人[75]研究了移动网络中边缘缓存节点之间的协作,并获得了在每个边缘节点中高速缓存内容的最佳冗余比.Poularakis 等人[76]共同设计了缓存和路由方案,以在边缘网络带宽限制下,最大化边缘服务器的缓存命中率.该问题被简化为位置放置问题,并使用有界逼近算法来解决.Ren 等人[77,78]提出了基于分组的缓存策略,并考虑存储资源分配,以降低获取内容的平均时延和总能耗.

4.4 小 结

针对边缘缓存问题进行的研究工作,主要解决了移动设备如何访问并使用 MEC 存储资源的问题.表 4 对上

述边缘缓存策略进行了对比分析.边缘缓存策略的目标主要是提升系统整体容量、提高缓存命中率、减小总传输开销、优化服务时延,提升移动用户内容交付服务的 QoE.所有已提出方案的共同缺点是仅有仿真时延证明其有效性,缺乏真实场景下的实验数据.

Table 4 Comparison of existing papers addressing edge caching problems

表 4 边缘缓存策略的对比分析

文献	缓存方案	目标	所提出的方法	实验方法	效果
[60]	非合作方案	提高系统整体容量	基于用户偏好	仿真	系统整体容量提升 300%
[70]	非合作方案	提高缓存命中率	基于增强学习	仿真	N/A
[71]	非合作方案	减少总传输开销	基于马尔科夫决策过程	仿真	总传输开销减少 70%
[72]	合作方案	提高缓存命中率	基于线性规划	仿真	N/A
[73]	合作方案	优化服务时延	基于整数线性规划	仿真	时延缩短 32%
[74]	合作方案	提高系统整体容量	二阶多项式时间算法	仿真	N/A
[75]	合作方案	减少总传输开销	自适应粒子群优化算法	仿真	总传输开销减少 54%
[76]	合作方案	提高缓存命中率	缓存和路由联合设计	仿真	N/A
[77]	合作方案	降低时延和能耗	基于分组的策略	仿真	N/A

5 MEC 服务编排与移动性管理

MEC 的服务质量依赖于服务编排功能以及 MD 与网络架构的交互.由于 MEC 是基于虚拟化平台的,那么 MEC 中管理和编排功能够尽可能地复用基于 NFV 的基础架构,将虚拟化网络功能(VNF)和移动边缘应用程序托管到相同或相似的基础设施.目前,ETSI MEC 正在研究如何在 NFV 环境中实现 MEC,并在未来的 5G 网络中定义管理和编排框架[79].本节对 MEC 服务编排中的研究工作进行了总结,并探索了 MEC 服务的移动性管理.

5.1 MEC服务编排研究

将 MEC 平台集成到移动网络环境中带来了与服务编排相关的许多挑战,主要原因是动态变化的无线信道状态以及由用户移动性引发的资源波动.MEC 系统应支持应用程序生命周期管理,即根据第三方的需求实例化或终止应用程序.当一个 MEC 平台执行服务编排时,资源管理、服务放置以及边缘节点的选择,对于提高网络资源利用率、提升用户体验质量和服务可靠性是至关重要的.本节从以下 4 个方面对 MEC 服务编排的研究现状进行了总结.

(1) 资源管理

灵活的资源可用性在 MEC 服务的性能表现中起着至关重要的作用.Taleb 等人[80]分析了有关 VM 资源管理的各种影响因素,同时考虑了资源竞争时 CPU、内存、外存、网络带宽的可用性.Liang 等人[81]研究了在云环境中分配资源时的应用概念及其特定的服务特性.Liu 等人[82]详细阐述了基于半马尔可夫决策过程的多资源分配策略,该策略确定了用于实现最优系统效益的确切无限带宽资源和计算资源的管理.

(2) 服务部署

将 MEC 服务部署在多个边缘云平台上,对于用户 QoE 来说是至关重要的.在服务部署的过程中,应该考虑人口密集的地区,例如吸引大量用户的购物中心.此外,服务的部署还需要考虑潜在的用户移动模型,以确保相关用户始终在时延、计算能力等方面获得所需的性能.Jia 等人[83,84]在给定移动模型和动态资源需求的前提下,详细阐述了最佳边缘云部署的优化研究.Volley[85]则是另一种解决方案,它专注于解决许多固定位置的分布式边缘云平台上的 VM 布局问题,在考虑用户位置的同时,执行服务部署的动态迁移.此外,MEC 服务的放置问题可以参考虚拟移动网络研究中用于 VNF 放置的解决方案[86].

(3) 边缘节点选择

在边缘节点的选择研究工作中,通常从性能的角度(例如时延)出发,将 MD 分配到距离最近的 MEC 平台.但是,这种策略可能会导致效率低下,引发性能瓶颈,尤其是在不考虑 MEC 服务器负载时.对于移动设备来说,由于用户移动和无线信道状态的不确定性,边缘服务器节点的选择变得至关重要.

(4) 可靠性

编排MEC服务的另一个重要问题是可靠性.传统的方案中,通常使用检查点算法解决计算系统中的容错问题,它负责维护应用程序状态的常规快照,可用于在出错时恢复应用程序,保障服务的可靠性.然而,由于移动环境下信道状态的动态变化特征,可能需要频繁地使用检查点算法,从而提高系统的可靠性.解决可靠性问题的另一种可行方案是复制MEC服务实例.与传统的检查点算法相比,该方案能够节省更多的时间.

5.2 MEC服务的移动性管理

在用户使用MEC服务期间,如何在用户发生移动的过程中保障MD和MEC之间会话连接的稳定,是服务编排之外另一个需要重视的问题.在用户发生移动的情况下,MD的位置会频繁地改变(例如从一个MEC节点到另一个MEC节点).在这种场景下,保障MEC服务处于最佳QoE状态变得极具挑战性,尤其是对于时延敏感型的应用.分布式移动性管理(DMM)[87]是管理用户移动性的解决方案之一,也克服了集中式移动性管理方案在可扩展性和可靠性方面的缺陷.然而,考虑到与MD单跳连接的MEC服务器需要频繁地迁移MEC服务并将它们放置在靠近移动用户的位置,那么,管理用户移动性并将用户的MEC服务请求重定向到托管服务的远程MEC服务器可能不是最佳的解决方案.此外,在跨边缘部署环境中,服务迁移需要将虚拟应用实例(VM/容器)传输到目标位置,可能导致该方案的传输代价过高.

虽然学术界在VM迁移方面已经做了许多的研究工作,但从服务质量的角度来看,VM迁移的过程会带来重大的技术挑战[88],例如VM迁移期间的服务连续性问题.为了在广域网上执行具有IP连续性的迁移,Watanabe等人[89]提出了IP移动解决方案.但是对基于IP的服务而言,IP地址的更改将导致会话中断,从而需要重新建立新的IP地址连接,进而会对QoE造成影响.

当网络中两个节点里的任何一个在移动期间发生IP地址变化时,引入DNS和NAT等技术可应对两节点之间IP对话崩溃的问题[90],但这两个技术本身不支持终端移动性或实时位置更新.为了解决这个基于IP的服务移动性管理问题,Ksentini等人[91]提出了Follow-Me-Cloud的概念.作者在文中介绍了一个能够确保云服务跟随用户移动的框架.该方案通过服务/数据识别的方式转换IP寻址.此外,通过将服务移动性与框架的第2层和第3层分离,该框架确保了无缝迁移和服务连续性,详细阐述了迁移服务的决策逻辑.Nadembega等人[92]融合了支持QoS/QoE的移动预测模型,进一步完善了Follow-Me-Cloud的概念.

此外,为了补充Follow-Me-Cloud的解决方案,Aissioui等人[93]引入了基于SDN的方法,作者在架构中融合了分布式弹性控制器.Secci等人[94]的研究工作则同时考虑了VM迁移以及用户的移动性,引入了基于位置/ID分离协议(LISP)的方法,以避免三角路由并减少VM迁移的中断时间.对于边缘计算平台,借助SDN控制器,可以通过监控服务的方式做出策略调整,以确保灵活的连接和有效的QoE/QoS管理[95].一般来说,根据应用程序类型的不同,MD的服务需求(如低时延)可能会有所不同.因此,SDN控制器或协调器应该在服务迁移的情况下考虑到这一点.例如,对于时延敏感型的应用,批量迁移的性能表现将优于实时迁移[96].

6 移动边缘计算环境中的典型应用

由于新兴应用在数据传输速率、服务时延等方面的需求越来越严格,因此,新兴应用的普及是网络架构发生演变的主要推动力量.本节将总结一些基于移动边缘计算的新兴应用和使用案例.

6.1 增强现实/虚拟现实

增强现实技术(AR)和虚拟现实技术(VR)被认为是改变我们生活方式的最有前途的应用.AR/VR应用程序需要使用一些实时的用户状态信息,例如用户所面对的位置和方向.此外,为了保障AR/VR应用的高服务质量,该类型应用程序通常需要较强的计算能力、较低的时延以及较高的带宽来完成任务迁移.MEC服务器能够利用本地上下文信息,同时具有较强的处理能力,非常适合AR/VR应用.Simone等人[97]提出了一种移动边缘计算环境下的视场(FOV)渲染解决方案,优化了传输360°VR视频流所需的带宽和时延.Dastjerdi等人[98]介绍了一种通过检测人脑电波工作的“大脑-计算机”交互AR应用.脑电波数据由脑电生物传感器实时接收,同时,借助MEC

和云计算平台处理大型的计算任务.Schneider 等人[99]设计了一种基于边缘计算的 AR 应用架构,克服了智能手机、平板电脑等移动 AR 设备在性能方面存在的困难,同时,将移动 AR 应用的端到端延迟减少到了 50ms 以内.

6.2 动态内容交付

据观察,视频流量占到当前网络中移动数据总流量的一半以上,而且这一比例仍在上升,这导致移动网络回程链路面临着传统集中式网络架构中的拥塞问题.在网络的边缘执行内容缓存,可以根据链路状态信息和用户的情景感知信息提供动态的内容交付服务[100],这避免了许多冗余视频流通过移动核心网传输到互联网 CDN.此外,借助 MEC 服务器,用户可以在功能强大的边缘计算平台上执行一些视频分析的操作,而不是在视频的源头[101].由于内容被放置在贴近用户的位置,视频分析服务的 QoE 能够得到显著提高.

除了多媒体内容交付之外,移动边缘计算技术在网站性能优化中也起着关键作用,例如缓存 HTML 内容、重新组织网页布局以及调整页面组件大小.移动用户发出的 HTTP 请求会经过 MEC 服务器;然后,MEC 服务器通过执行多种任务来加载用户移动设备界面上的网页或基于 Web 的应用程序.由于 MEC 服务器部署在边缘设备附近,因此 Web 应用的请求和响应非常节省时间.

6.3 车联网

移动边缘计算可以在车辆连接、V2X 通信和汽车安全服务中发挥重要作用,例如实时警告高速公路的路面是否结冰以及协调机动车更改车道[102].运行在 MEC 服务器上的应用程序与车辆的距离非常近,可以提供低延迟的路边功能[103].此外,借助于 MEC 服务还能够实现交通控制和智能停车,因为边缘计算平台能够收集和分析来自传感器设备的实时数据[21].

互联网的接入,使得车辆能够与道路上的其他车辆互相连接.在路边部署 MEC 平台,可以实现移动车辆之间的双向通信.一辆车可以与其他近距离的车辆通信,并告知他们预期的风险事故或交通堵塞.此外,MEC 可以实现与局域传感器同步的分布式网络环境[104].

6.4 物联网

随着智能传感器、互联网协议和通信技术的进步,物联网(IoT)正在逐步走进人们的日常生活[105].边缘计算架构在支持物联网应用程序方面有着天然的优势.Gazis 等人[106]提出了一个自动适配的操作平台,能够在工业物联网环境中应用边缘计算组件.基于虚拟机的 Cloudlet,能够根据物联网中的众包视频内容实现边缘分析[107].MEC 将提供之前不可行的新 IoT 服务[108].物联网中 MEC 的具体应用和使用案例将介绍如下.

6.4.1 智慧医疗

实时处理和事件响应对于医疗保健类的应用非常重要.像其他行业一样,医疗行业也可以利用边缘计算得到帮助,例如避免患有中风的病人跌倒.为了检测和防止跌倒,学术界已经进行了大量的研究,例如通过引入智能手机、智能手表和谷歌眼镜等人机交互设备.Cao 等人[109]提出了一种名为 U-Fall 的智能医疗基础设施,它利用边缘计算技术开放了智能手机的能力.U-Fall 设计了基于加速度数值和非线性时间序列分析的跌倒检测算法,借助智能设备传感器(如陀螺仪和加速度计)感知运动检测.U-Fall 智能地保持智能手机和 MEC 服务器之间的持续性交互,以确保实时检测.

6.4.2 智能电网

由于智能电网的基础设施是由多种组件构成的,因此,对智能电网环境中生成的数据进行分析是一项非常具有挑战性的任务.移动边缘计算的使用,可以提高数据吞吐量、缩短响应时间和传输延迟[110].智能电网是一种典型的需要 MEC 和云中心协同工作的用例[111].部署在网络边缘的 MEC 服务器收集并处理由电网传感器和设备生成的本地数据,云中心则负责提供全局覆盖并保存所有的数据备份.

6.4.3 智能家居

智能家居系统已成为未来家庭生态系统的新趋势.智能家居是一种占地空间和本地化通信受限的小型物联网系统.将 MEC 服务器部署在靠近智能对象的物联网网关,将使未来网络中的 M2M 直接交互成为可能[112].MEC 节点支持部署在家庭路由器、机顶盒或智能手机上,这一优势能够为智能家居带来低时延、本地化和即

插即用的服务.

7 发展趋势与开放式研究挑战

移动边缘计算是未来5G移动网络架构下的重要组成部分.与现有的3G/4G移动蜂窝网络相比,它具有许多新功能,拥有更好的 QoE 和灵活性.因此在未来的研究工作中,移动边缘计算会面临着各种各样的机遇和挑战.在本节中,我们总结了移动边缘计算中的开放式研究挑战,并阐述了未来可能的研究方向.

7.1 服务质量保障

(1) 异构性

在未来的移动网络环境下,随着物联网和新兴应用的发展,网络、信道和基础设施的异构性将会成为一个关键问题.在当前的研究现状中,网络环境通常被假设为是同构的(MEC 服务器配备有相同的计算资源、存储资源).虽然现有的模型对于仿真实验和分析来说很简单,但它并没有充分反映移动网络异构性的特点.因此,应该假设在异构网络的环境下模拟与移动边缘计算相关的一些实验.

(2) 用户移动性

用户移动性是移动边缘计算中的关键挑战,它对计算迁移决策有着不可忽视的影响.最近的一些研究成果[113,114]多采用虚拟机迁移或路径选择的方式实现用户移动过程中服务连续性的保障.然而这种机制却无形中引发了 MEC 服务器之间频繁的数据交换,增加了网络的负担.如何在用户的移动过程中实现 MD 与 MEC 服务器之间的无缝衔接,也是移动边缘计算必须解决的问题.

(3) 可扩展性

与传统的计算范例相比,可扩展性是移动边缘计算的一项重要特征.随着 5G 通信技术和 IoT 技术的发展,越来越多的终端设备(如物联网中的传感器设备)需要服务的可扩展性.近年来,边缘设备的数量不断增加,如果大量设备同时访问 MEC 服务器,将会形成带宽瓶颈,最终导致服务中断.参考移动云计算中的做法,移动边缘计算也许可以引入编排器来灵活管理 MEC 服务器,提高网络的可扩展性[115].

7.2 安全可用性

(1) 安全

在移动边缘计算环境下,MEC 服务器需要使用移动设备的一些信息,这将会引发一些新的安全挑战[116].尽管安全方案的发展速度也在不断地加快,但是依然无法跟上新出现安全威胁的步伐.许多现有的安全协议都将网络链路假设为完全连接状态[117],这在移动边缘计算环境下并不现实,因为许多链路在默认情况下都是间歇性的.Mtibaa 等人[116]提出了一种名为 HoneyBot 的防御技术,HoneyBot 节点可以检测、跟踪和隔离 D2D(设备到设备)内部攻击,该技术的速度和准确性受 HoneyBot 节点的位置和数量的影响.

用于MCC环境的安全解决方案可能不适合移动边缘计算,因为云是集中式的计算架构,而边缘计算则是分散式的,MEC 的工作环境将面临许多不同的新威胁.比如,不同级别的网关和智能设备的认证就是移动边缘计算环境中的一个重要的安全问题.针对 MEC 环境下的认证问题,学术界也提出了一些新的解决方案,如基于公钥基础设施(PKI)的解决方案[118]、基于 DiffieHellman 密钥交换的解决方案[119].

(2) 隐私

如何在不侵犯用户隐私的前提下通过 MEC 服务器访问并使用移动设备的资源,是 MEC 环境下的一个全新挑战.已存在的机制可以为移动边缘计算中各个MEC服务器之间隐私机制的设计指明方向.例如在智能电网中,由MEC服务器执行智能电表数据的加密以及汇总结果确保了数据的隐私安全,原始数据只能在操作中心执行解密[120].此外,MCC 中已经制定了许多数据隐私保护机制,以便在移动设备之间执行隐私保护策略并达到隐藏客户端位置数据的效果.边缘计算范例还有助于增强某些服务的隐私保护功能.在基于位置的服务中,可以通过可信 MEC 服务器部署一个众包平台来完成用户的匿名化操作[121].

7.3 功能增强

(1) 计费策略

在移动边缘计算中,存储、计算和通信资源会根据用户的需求动态分配.因此,MEC 的最优计费策略与传统计算模型有很大的不同.移动边缘计算环境会涉及到多种服务提供者,这些服务提供者的报价是不同的.每一种服务提供者有着不同的支付方式、不同的客户管理方式和不同的服务政策.例如,用户设备上的游戏应用必须使用边缘计算资源、移动网络和游戏服务.用户必须为游戏支付费用,该费用应该平等地分配给游戏服务中参与设计的所有实体.已有研究已证明:当用户关心服务费用时,MEC 服务器的盈利受到计费策略的显著影响[122].

(2) Web 接口

在当前的网络环境下,移动用户访问 MEC 或云中心的 Web 接口是无法满足新兴的时延敏感型应用需求的.当前的 Web 接口通常不是为该类应用设计的,因此存在兼容性问题.用户、MEC 和云之间的通信需要新的标准协议.最新版本的 HTML5 专为新出现的移动设备而设计,如平板电脑或智能手机.但是该技术依然需要进行性能调教和功能测试方面的研究,以便为未来的移动边缘计算场景做好准备.

8 结 论

移动边缘计算能够将传统核心网络的计算和存储能力纳入到无线电接入网络的范围之内.在这种新兴架构中,传统的基站不仅可以执行流量控制,还可以部署具备轻量级资源的 MEC 服务器,为移动用户提供具有上下文感知、位置感知特点的服务.移动边缘计算的主要目标是为移动端的应用程序提供更小的时延和更高的带宽利用率.随着 5G 通信技术的发展和移动互联网的广泛应用,移动边缘计算受到了学术界的广泛关注,并在 MEC 架构、计算迁移、边缘缓存以及服务编排等方面开展了深入研究,取得了一系列重要研究成果.本文对这些成果进行了系统的归纳和总结,并进一步指出了未来的一些研究发展方向.然而,随着虚拟/增强现实、动态内容交付、物联网等新兴应用的不断涌现,以及移动应用向医疗、教育、公共服务等领域的进一步渗透,移动边缘计算在可靠性、高效性和安全性方面还面临着许多新的技术挑战,也为研究者提供了一系列新的研究方向.

References:

[1] Ahmed E, Gani A, Khan MK, *et al*. Seamless application execution in mobile cloud computing: Motivation, taxonomy, and open challenges. Journal of Network & Computer Applications, 2015,52:154−172.

[2] Dinh HT, Lee C, Niyato D, *et al*. A survey of mobile cloud computing: Architecture, applications, and approaches. Wireless Communications & Mobile Computing, 2013,13(18):1587−1611.

[3] Barbarossa S, Sardellitti S, Lorenzo PD. Communicating while computing: distributed mobile cloud computing over 5G heterogeneous networks. IEEE Signal Processing Magazine, 2014,31(6):45−55.

[4] Khan AUR, Othman M, Madani SA, *et al*. A survey of mobile cloud computing application models. IEEE Communications Surveys & Tutorials, 2014,16(1):393−413.

[5] Satyanarayanan M, Bahl P, Davies N. The case for VM-based cloudlets in mobile computing. IEEE Pervasive Computing, 2009, 8(4):14−23.

[6] Bonomi F, Milito R, Zhu J, *et al*. Fog computing and its role in the internet of things. In: Proc. of the Workshop on Mobile Cloud Computing (MCC). ACM Press, 2012. 13−16.

[7] Zhu J, Chan D S, Prabhu M S, *et al*. Improving Web sites performance using edge servers in fog computing architecture. In: Proc. of the IEEE 7th Int'l Symp. on Service-oriented System Engineering. IEEE Computer Society, 2013. 320−323.

[8] Stojmenovic I, Wen S. The fog computing paradigm: Scenarios and security issues. In: Proc. of the 2014 Federated Conf. on Computer Science and Information Systems. IEEE, 2014. 1−8.

[9] Stojmenovic I. Fog computing: A cloud to the ground support for smart things and machine-to-machine networks. In: Proc. of the 2014 Australasian Telecommunication Networks and Applications Conf. (ATNAC). IEEE, 2014. 117−122.

[10] Liu Y, Fieldsend JE, Min G. A framework of fog computing: Architecture, challenges, and optimization. IEEE Access, 2017,5: 25445−25454.

[11] Mach P, Becvar Z. Mobile edge computing: A survey on architecture and computation offloading. IEEE Communications Surveys & Tutorials, 2017,19(3):1628−1656.

[12] European Telecommunications Standards Institute. Mobile-edge computing—Introductory technical white paper. ETSI ISG MEC, 2014. https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge computing—Introductory technical white paper v1%2018-09-14.pdf

[13] Tanaka H, Yoshida M, Mori K, *et al*. Multi-access edge computing: A survey. Journal of Information Processing, 2018,26:87–97.

[14] Zhang WL, Guo B, Shen Y, *et al*. Computation offloading on intelligent mobile terminal. Chinese Journal of Computers, 2016, 38(5):1021−1038 (in Chinese with English abstract).

[15] Botta A, Donato WD, Persico V. On the integration of cloud computing and Internet of things. In: Proc. of the Int'l Conf. on Future Internet of Things and Cloud. IEEE, 2014. 23–30.

[16] Jararweh Y, Doulat A, Darabseh A, *et al*. SDMEC: Software defined system for mobile edge computing. In: Proc. of the IEEE Int'l Conf. on Cloud Engineering Workshop. IEEE. 88–93.

[17] Salman O, Elhajj I, Kayssi A, *et al*. Edge computing enabling the Internet of things. In: Proc. of the 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT). IEEE, 2015. 603–608.

[18] Cui C, Deng H, Telekom D, Mihel U, Damker H. Network function virtualisation: Network operator perspectives on industry progress. Updated White Paper, 2013. https://www.researchgate.net/publication/275037832_Network_Functions_Virtualisation_NFV_Network_Operator_Perspectives_on_Industry_Progress

[19] Satyanarayanan M. The emergence of edge computing. Computer, 2017,50(1):30–39.

[20] Shi W, Cao J, Zhang Q, *et al*. Edge computing: vision and challenges. IEEE Internet of Things Journal, 2016,3(5):637–646.

[21] Ahmed A, Ahmed E. A survey on mobile edge computing. In: Proc. of the Int'l Conf. on Intelligent Systems and Control. IEEE, 2016. 1–8.

[22] Roman R, Lopez J, Mambo M. Mobile edge computing, Fog *et al*.: A survey and analysis of security threats and challenges. Future Generation Computer Systems, 2016,78:680–698.

[23] Mobile Edge Computing (MEC). Framework and reference architecture. V1.1.1, ETSI GS MEC Standard 003, 2016.

[24] Mobile Edge Computing (MEC). MEC metrics best pratice and guidelines. V0.1.0, ETSI GS MEC-IEG Standard 006, 2016.

[25] OpenFog Consortium. OpenFog architecture overview. White Paper, 2016. https://www.openfogconsortium.org/wp-content/uploads/OpenFog-Architecture-Overview-WP-2-2016.pdf

[26] SESAME project: Small cEllS coordination for multi-tenancy and edge services (SESAME). SESAME Project (Web page), 2018. http://www.sesame-h2020-5g-ppp.eu/

[27] Ahmad S, Ahed A, Eshraq A, *et al*. SESAME project: SESAME: An innovative multi-operator enabled Small Cell based infrastructure that integrates a virtualised execution platform for deploying virtual network functions. SESAME Project 2nd White Paper, 2017. http://http://sesame.org.jo/ sesame_2018/machine-and-beamlines/sesame-publications/sesame-white-book

[28] Cziva R, Pezaros DP. Container network functions: Bringing NFV to the network edge. IEEE Communications Magazine, 2017, 55(6):24–31.

[29] Rimal BP, Van DP, Maier M. Mobile edge computing empowered fiber-wireless access networks in the 5G era. IEEE Communications Magazine, 2017,55(2):192–200.

[30] CMCC. Study on context aware service delivery for LTE, 3GPP TSG RAN meeting#71, 3GPP. Technical Report, RP-160633, Gothenburg, 2016.

[31] Agiwal M, Roy A, Saxena N. Next generation 5G wireless networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2017,18(3):1617–1655.

[32] Liu J, Mao Y, Zhang J, *et al*. Delay-optimal computation task scheduling for mobile-edge computing systems. In: Proc. of the IEEE International Symposium on Information Theory (ISIT). IEEE, 2016. 1451–1455.

[33] Mao Y, Zhang J, Letaief KB. Dynamic computation offloading for mobile-edge computing with energy harvesting devices. IEEE Journal on Selected Areas in Communications, 2016,34(12):3590–3605.

[34] Ulukus S, Yener A, Erkip E, *et al*. Energy harvesting wireless communications: A review of recent advances. IEEE Journal on Selected Areas in Communications, 2015,33(3):360–381.

[35] Kamoun M, Labidi W, Sarkiss M. Joint resource allocation and offloading strategies in cloud enabled cellular networks. In: Proc. of the IEEE Int'l Conf. on Communications. IEEE, 2015. 5529–5534.

[36] Labidi W, Sarkiss M, Kamoun M. Energy-optimal resource scheduling and computation offloading in small cell networks. In: Proc. of the Int'l Conf. on Telecommunications. IEEE, 2015. 313–318.

[37] Chen MH, Liang B, Dong M. A semidefinite relaxation approach to mobile cloud offloading with computing access point. In: Proc. of the IEEE Int'l Workshop on Signal Processing Advances in Wireless Communications. IEEE, 2015. 186–190.

[38] Cao S, Tao X, Hou Y, *et al*. An energy-optimal offloading algorithm of mobile computing based on HetNets. In: Proc. of the Int'l Conf. on Connected Vehicles and Expo. IEEE, 2016. 254−258.

[39] Deng M, Tian H, Fan B. Fine-granularity based application offloading policy in small cell cloud-enhanced networks. In: Proc. of the IEEE ICC. IEEE, 2016. 638−643.

[40] Muñoz O, Pascual-Iserte A, Vidal J. Joint allocation of radio and computational resources in wireless application offloading. In: Proc. of the Future Network and Mobile Summit. IEEE, 2014. 1−10.

[41] Muñoz O, Pascual-Iserte A, Vidal J. Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading. IEEE Trans. on Vehicular Technology, 2015,64(10):4738−4755.

[42] Sehati A, Ghaderi M. Energy-delay tradeoff for request bundling on smartphones. In: Proc. of the IEEE Int'l Conf. on Computer Communications. IEEE, 2017. 1−7.

[43] Labidi W, Sarkiss M, Kamoun M. Joint multi-user resource scheduling and computation offloading in small cell networks. In: Proc. of the IEEE Int'l Conf. on Wireless and Mobile Computing, Networking and Communications. IEEE, 2015. 794−801.

[44] Barbarossa S, Sardellitti S, Lorenzo PD. Joint allocation of computation and communication resources in multiuser mobile cloud computing. 2013,395(6):26−30.

[45] Sardellitti S, Scutari G, Barbarossa S. Joint optimization of radio and computational resources for multicell mobile cloud computing. In: Proc. of the IEEE Int'l Workshop on Signal Processing Advances in Wireless Communications. IEEE, 2014. 89−103.

[46] Zhang K, Mao Y, Leng S, *et al*. Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks. IEEE Access, 2017,4(99):5896−5907.

[47] Chen X, Jiao L, Li W, *et al*. Efficient multi-user computation offloading for mobile-edge cloud computing. IEEE/ACM Trans. on Networking, 2016,24(5):2795−2808.

[48] Chen MH, Liang B, Dong M. Joint offloading and resource allocation for computation and communication in mobile cloud with computing access point. In: Proc. of the IEEE Conf. on Computer Communications (IEEE INFOCOM 2017). IEEE, 2017. 1−9.

[49] Zhao Y, Zhou S, Zhao T, *et al*. Energy-efficient task offloading for multiuser mobile cloud computing. In: Proc. of the IEEE/CIC Int'l Conf. on Communications in China. IEEE, 2016. 1−5.

[50] You C, Huang K. Multiuser resource allocation for mobile-edge computation offloading. In: Proc. of the 2016 IEEE Global Communications Conf. (GLOBECOM). IEEE, 2016. 1−6.

[51] You C, Huang K, Chae H, *et al*. Energy-efficient resource allocation for mobile-edge computation offloading. IEEE Trans. on Wireless Communications, 2017,16(3):1397−1411.

[52] Muñoz O, Iserte AP, Vidal J, *et al*. Energy-latency trade-off for multiuser wireless computation offloading. In: Proc. of the Wireless Communications and Networking Conf. Workshops. IEEE, 2014. 29−33.

[53] Mao Y, Zhang J, Song SH, *et al*. Power-delay tradeoff in multi-user mobile-edge computing systems. In: Proc. of the 2016 IEEE Global Communications Conf. (GLOBECOM). IEEE, 2016. 1−6.

[54] Zhao T, Zhou S, Guo X, *et al*. A cooperative scheduling scheme of local cloud and internet cloud for delay-aware mobile cloud computing. In: Proc. of the IEEE GLOBECOM Workshops. IEEE, 2015. 1−6.

[55] Tanzil SMS, Gharehshiran ON, Krishnamurthy V. Femto-cloud formation: A coalitional game-theoretic approach. In: Proc. of the IEEE Global Communications Conf. IEEE, 2015. 1−6.

[56] Guo X, Singh R, Zhao T, *et al*. An index based task assignment policy for achieving optimal power-delay tradeoff in edge cloud systems. In: Proc. of the IEEE Int'l Conf. on Communications. IEEE, 2016. 1−7.

[57] Oueis J, Calvanese-Strinati E, De Domenico A, *et al*. On the impact of backhaul network on distributed cloud computing. In: Proc. of the Wireless Communications and Networking Conf. Workshops. IEEE, 2014. 12−17.

[58] Oueis J, Strinati EC, Barbarossa S. Small cell clustering for efficient distributed cloud computing. In: Proc. of the IEEE Int'l Symp. on Personal, Indoor, and Mobile Radio Communication. IEEE, 2015. 1474−1479.

[59] Oueis J, Strinati EC, Barbarossa S. The fog balancing: load distribution for small cell cloud computing. In: Proc. of the Vehicular Technology Conf. IEEE, 2015. 1−6.

[60] Ahlehagh H, Dey S. Video-aware scheduling and caching in the radio access network. IEEE/ACM Trans. on Networking, 2014, 22(5):1444−1462.

[61] Tandon R, Simeone O. Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks. In: Proc. of the IEEE Int'l Symp. on Information Theory. IEEE, 2016. 2029−2033.

[62] Liu D, Yang C. Will caching at base station improve energy efficiency of downlink transmission? In: Proc. of the Signal and Information Processing. IEEE, 2015. 173−177.

[63] Zhang J, Zhang X, Wang W. Cache-enabled software defined heterogeneous networks for green and flexible 5G networks. IEEE Access, 2016,4(4):3591−3604.

[64] Paschos G, Bastug E, Land I, *et al*. Wireless caching: technical misconceptions and business barriers. IEEE Communications Magazine, 2016,54(8):16−22.

[65] Breslau L, Cao P, Fan L, *et al*. Web caching and Zipf-like distributions: Evidence and implications. In: Proc. of the IEEE Conf. on Computer Communications, Vol.1. 1999. 126−134.

[66] Traverso S, Ahmed M, Garetto M, *et al*. Temporal locality in today's content caching: Why it matters and how to model it? ACM SIGCOMM Computer Communication Review, 2013,43(5):5−12.

[67] Cha M, Kwak H, Rodriguez P, *et al*. Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Trans. on Networking, 2009,17(5):1357−1370.

[68] Ioannou A, Weber S. A survey of caching policies and forwarding mechanisms in information-centric networking. IEEE Communications Surveys & Tutorials, 2016,18(4):2847−2886.

[69] Ahlehagh H, Dey S. Video caching in radio access network: Impact on delay and capacity. In: Proc. of the 2012 IEEE Wireless Communications and Networking Conf. (WCNC). IEEE, 2012. 2276−2281.

[70] Sengupta A, Amuru SD, Tandon R, *et al*. Learning distributed caching strategies in small cell networks. In: Proc. of the Int'l Symp. on Wireless Communications Systems. IEEE, 2014. 917−921.

[71] Gu J, Wang W, Huang A, *et al*. Distributed cache replacement for caching-enable base stations in cellular networks. In: Proc. of the IEEE Int'l Conf. on Communications. IEEE, 2014. 2648−2653.

[72] Borst S, Gupta V, Walid A. Distributed caching algorithms for content distribution networks. In: Proc. of the Conf. on Information Communications. IEEE Press, 2010. 1478−1486.

[73] Jiang W, Feng G, Qin S. Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. IEEE Trans. on Mobile Computing, 2017,16(5):1382−1393.

[74] Yu R, Qin S, Bennis M, *et al*. Enhancing software-defined RAN with collaborative caching and scalable video coding. In: Proc. of the 2016 IEEE Int'l Conf. on Communications (ICC 2016). IEEE, 2016. 1−6.

[75] Wang S, Zhang X, Yang K, *et al*. Distributed edge caching scheme considering the tradeoff between the diversity and redundancy of cached content. In: Proc. of the IEEE/CIC Int'l Conf. on Communications in China. IEEE, 2016. 1−5.

[76] Poularakis K, Iosifidis G, Tassiulas L. Approximation caching and routing algorithms for massive mobile data delivery. In: Proc. of the Global Communications Conf. IEEE, 2014. 3534−3539.

[77] Ren D, Gui X, Lu W, *et al*. GHCC: Grouping-based and hierarchical collaborative caching for mobile edge computing. In: Proc. of the 2018 16th Int'l Symp. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt). IEEE, 2018. 1−6.

[78] Ren D, Gui X, Dai H, *et al*. Hierarchical resource distribution network based on mobile edge computing. In: Proc. of the 2017 IEEE 23rd Int'l Conf. on Parallel and Distributed Systems (ICPADS). IEEE, 2017. 57−64.

[79] Deployment of mobile edge computing in an NFV environment. ETSI MEC work item DGS/MEC-0017MECinNFV. Sophia Antipolis, 2017. https://www.etsi.org/deliver/etsi_gr/MEC/001_099/017/01.01.01_60/gr_MEC017v010101p.pdf

[80] Taleb T, Bagaa M, Ksentini A. User mobility-aware virtual network function placement for virtual 5G network infrastructure. In: Proc. of the 2015 IEEE Int'l Conf. on Communications (ICC). IEEE, 2015. 3879−3884.

[81] Liang H, Cai LX, Huang D, *et al*. An SMDP-based service model for interdomain resource allocation in mobile cloud networks. IEEE Trans. on Vehicular Technology, 2012,61(5):2222−2232.

[82] Liu Y, Lee MJ, Zheng Y. Adaptive multi-resource allocation for cloudlet-based mobile cloud computing system. IEEE Trans. on Mobile Computing, 2016,15(10):2398−2410.

[83] Jia M, Cao J, Liang W. Optimal cloudlet placement and user to cloudlet allocation in wireless metropolitan area networks. IEEE Trans. on Cloud Computing, 2017,5(4):725−737.

[84] Bagaa M, Taleb T, Ksentini A. Service-aware network function placement for efficient traffic handling in carrier cloud. In: Proc. of the 2014 IEEE Wireless Communications and Networking Conf. (WCNC). IEEE, 2014. 2402−2407.

[85] Agarwal S, Dunagan J, Jain N, *et al*. Volley: Automated data placement for geo-distributed cloud services. 2010. https://www.usenix.org/legacy/events/nsdi10/tech/full_papers/agarwal.pdf

[86] Ksentini A, Bagaa M, Taleb T, *et al*. On using bargaining game for optimal placement of SDN controllers. In: Proc. of the 2016 IEEE Int'l Conf. on Communications (ICC). IEEE, 2016. 1–6.

[87] Giust F, Cominardi L, Bernardos CJ. Distributed mobility management for future 5G networks: Overview and analysis of existing approaches. IEEE Communications Magazine, 2015,53(1):142–149.

[88] Bittencourt LF, Lopes MM, Petri I, *et al*. Towards virtual machine migration in fog computing. In: Proc. of the 2015 10th Int'l Conf. on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC). IEEE, 2015. 1–8.

[89] Watanabe H, Ohigashi T, Kondo T, *et al*. A performance improvement method for the global live migration of virtual machine with IP mobility. In: Proc. of the 5th Int'l Conf. on Mobile Computing and Ubiquitous Networking (ICMU 2010), Vol.94. 2010. 1–6.

[90] Samdanis K, Taleb T, Schmid S. Traffic offload enhancements for eUTRAN. IEEE Communications Surveys & Tutorials, 2012, 14(3):884–896.

[91] Ksentini A, Taleb T, Chen M. A Markov decision process-based service migration procedure for follow me cloud. In: Proc. of the 2014 IEEE Int'l Conf. on Communications (ICC). IEEE, 2014. 1350−1354.

[92] Nadembega A, Hafid AS, Brisebois R. Mobility prediction model-based service migration procedure for follow me cloud to support QoS and QoE. In: Proc. of the 2016 IEEE Int'l Conf. on Communications (ICC). IEEE, 2016. 1–6.

[93] Aissioui A, Ksentini A, Gueroui AM, *et al*. Toward elastic distributed SDN/NFV controller for 5G mobile cloud management systems. IEEE Access, 2015,3:2055–2064.

[94] Secci S, Raad P, Gallard P. Linking virtual machine mobility to user mobility. IEEE Trans. on Network and Service Management, 2016,13(4):927–940.

[95] Gramaglia M, Digon I, Friderikos V, *et al*. Flexible connectivity and QoE/QoS management for 5G networks: The 5G NORMA view. In: Proc. of the 2016 IEEE Int'l Conf. on Communications Workshops (ICC). IEEE, 2016. 373–379.

[96] Ceselli A, Premoli M, Secci S. Cloudlet network design optimization. In: Proc. of the IFIP Networking Conf. (IFIP Networking 2015). IEEE, 2015. 1–9.

[97] Mangiante S, Klas G, Navon A, *et al*. VR is on the edge: How to deliver 360° videos in mobile networks. In: Proc. of the Workshop on Virtual Reality and Augmented Reality Network. ACM Press, 2017. 30–35.

[98] Dastjerdi AV, Gupta H, Calheiros RN, *et al*. Fog Computing: Principles, Architectures, and Applications. Morgan Kaufmann Publishers, 2016. 61–75.

[99] Schneider M, Rambach J, Stricker D. Augmented reality based on edge computing using the example of remote live support. In: Proc. of the IEEE Int'l Conf. on Industrial Technology. IEEE, 2017. 1277–1282.

[100] Zhu J, Chan DS, Prabhu MS, *et al*. Improving web sites performance using edge servers in fog computing architecture. In: Proc. of the 2013 IEEE 7th Int'l Symp. on Service-oriented System Engineering. IEEE, 2013. 320–323.

[101] Mäkinen O. Streaming at the edge: Local service concepts utilizing mobile edge computing. In: Proc. of the Int'l Conf. on Next Generation Mobile Applications, Services and Technologies. IEEE, 2016. 1–6.

[102] Klas GI. Fog computing and mobile edge cloud gain momentum open fog consortium ETSI MEC and cloudlets. 2015. http://yucianga.info/?p=938

[103] Sabella D, Vaillant A, Kuure P, *et al*. Mobile-edge computing architecture: The role of MEC in the Internet of things. IEEE Consumer Electronics Magazine, 2016,5(4):84–91.

[104] Datta SK, Bonnet C, Haerri J. Fog computing architecture to enable consumer centric Internet of things services. In: Proc. of the IEEE Int'l Symp. on Consumer Electronics. IEEE, 2015. 1–2.

[105] Al-Fuqaha A, Guizani M, Mohammadi M, *et al*. Internet of things: A survey on enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials, 2015,17(4):2347–2376.

[106] Gazis V, Leonardi A, Mathioudakis K, *et al*. Components of fog computing in an industrial Internet of things context. In: Proc. of the 2015 12th Annual IEEE Int'l Conf. on Sensing, Communication, and Networking-Workshops (SECON Workshops). IEEE, 2015. 1–6.

[107] Satyanarayanan M, Simoens P, Xiao Y, *et al*. Edge analytics in the Internet of things. IEEE Pervasive Computing, 2015,14(2): 24–31.

[108] Corcoran P, Datta SK. Mobile-edge computing and the Internet of things for consumers: Extending cloud computing and services to the edge of the network. IEEE Consumer Electronics Magazine, 2016,5(4):73–74.

[109] Cao Y, Chen S, Hou P, *et al*. FAST: A fog computing assisted distributed analytics system to monitor fall for stroke mitigation. In: Proc. of the IEEE Int'l Conf. on Networking, Architecture and Storage. IEEE, 2015. 2−11.

[110] Kumar N, Zeadally S, Rodrigues JJPC. Vehicular delay-tolerant networks for smart grid data management using mobile edge computing. IEEE Communications Magazine, 2016,54(10):60−66.

[111] Bonomi F, Milito R, Zhu J, *et al*. Fog computing and its role in the Internet of things. In: Proc. of the Edition of the MCC Workshop on Mobile Cloud Computing. ACM Press, 2012. 13−16.

[112] Vallati C, Virdis A, Mingozzi E, *et al*. Mobile-edge computing come home connecting things in future smart homes using LTE device-to-device communications. IEEE Consumer Electronics Magazine, 2016,5(4):77−83.

[113] Secci S, Raad P, Gallard P. Linking virtual machine mobility to user mobility. IEEE Trans. on Network & Service Management, 2017,13(4):927−940.

[114] Plachy J, Becvar Z, Mach P. Path selection enabling user mobility and efficient distribution of data for computation at the edge of mobile network. Computer Networks, 2016,108:357−370.

[115] Cau E, Corici M, Bellavista P, *et al*. Efficient exploitation of mobile edge computing for virtualized 5G in EPC architectures. In: Proc. of the IEEE Int'l Conf. on Mobile Cloud Computing, Services, and Engineering. IEEE, 2016. 100−109.

[116] Mtibaa A, Harras K, Alnuweiri H. Friend or foe? Detecting and isolating malicious nodes in mobile edge computing platforms. In: Proc. of the IEEE Int'l Conf. on Cloud Computing Technology and Science. IEEE, 2016. 42−49.

[117] Deng H, Li W, Agrawal DP. Routing security in wireless ad hoc networks. IEEE Communications Magazine, 2002,40(10):70−75.

[118] Law YW, Palaniswami M, Kounga G, *et al*. WAKE: Key management scheme for wide-area measurement systems in smart grid. IEEE Communications Magazine, 2013,51(1):34−41.

[119] Fadlullah ZM, Fouda MM, Kato N, *et al*. Toward intelligent machine-to-machine communications in smart grid. IEEE Communications Magazine, 2011,49(4):60−65.

[120] Lu R, Liang X, Li X, *et al*. EPPA: An efficient and privacy-preserving aggregation scheme for secure smart grid communications. IEEE Trans. on Parallel & Distributed Systems, 2012,23(9):1621−1631.

[121] Abdo JB, Demerjian J, Chaouchi H, *et al*. Privacy using mobile cloud computing. In: Proc. of the Int'l Conf. on Digital Information & Communication Technology & Its Applications. IEEE, 2015. 178−182.

[122] Zhao T, Zhou S, Guo X, *et al*. Pricing policy and computational resource provisioning for delay-aware mobile edge computing. In: Proc. of the IEEE/CIC Int'l Conf. on Communications in China. IEEE, 2016. 1−6.

附中文参考文献:

[14] 张文丽,郭兵,沈艳,等.智能移动终端计算迁移研究.计算机学报,2016,38(5):1021−1038.

张开元(1992－),男,河南扶沟人,学士,CCF 学生会员,主要研究领域为移动边缘计算,计算迁移,资源分配.

李敬(1992－),男,硕士,主要研究领域为移动边缘计算,服务部署.

桂小林(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为云计算隐私保护,网络安全,可信计算.

吴杰(1995－),男,学士,主要研究领域为物联网,移动边缘计算.

任德旺(1989－),男,硕士,主要研究领域为移动边缘网络,边缘缓存,移动性管理.

任东胜(1994－),男,学士,主要研究领域为移动边缘计算,物联网.