6G 异构边缘计算

王鹏飞[1,2]，邸博雅[1,2]，宋令阳[1,2]，韩竹[3]
（1. 北京大学信息与通信研究所，北京 100871；2. 北京大学大数据研究所，北京 100871；
3. 休斯顿大学电子与计算机工程部，德克萨斯 休斯顿 77004）

摘 要：随着物联网的发展，在未来 6G 通信中，将会产生众多实时性应用场景。在低时延的数据处理需求的驱动下，移动边缘计算将成为提升用户体验和降低网络成本的重要技术。然而，单一的边缘服务器计算能力有限，很难满足计算密集型应用的低时延数据处理需求。设计了一种异构的多层边缘计算（HetMEC, heterogeneous multi-layer mobile edge computing）网络架构，综合利用云计算中心和多层边缘服务器的计算和传输资源，通过合理分割卸载计算任务，共同为边缘应用提供可靠、高效的计算服务。实验证明，HetMEC 网络架构可以有效降低处理时延，提升网络处理速率和稳健性。

关键词：物联网；异构网络；移动边缘计算；任务卸载；资源分配
中图分类号：TP391
文献标识码：A
doi: 10.11959/j.issn.2096−3750.2020.00147

Heterogeneous multi-layer mobile edge computing for 6G

WANG Pengfei[1,2], DI Boya[1,2], SONG Lingyang[1,2], HAN Zhu[3]
1. Institute of Information and Communication Technology, Peking University, Beijing 100871, China
2. National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871, China
3. Electrical and Computer Engineering Department, University of Houston, Houston 77004, U.S.A

Abstract: With the development of the Internet of things, multiple real-time applications will emerge in future 6G communication network. Driven by demands on low-latency services of multiple real-time applications, mobile edge computing (MEC) will become an important technology to improve the user experience and reduce the network cost. However, the computing capacity of a single MEC server is limited, which induces that it is difficult to meet the low-latency requirement of data processing for the computation-intensive applications. A heterogeneous multi-layer mobile edge computing (HetMEC) network architecture was proposed to jointly utilize the computing and transmission resources of both the cloud computing center and multi-layer MEC servers. By reasonably dividing and offloading computing tasks, reliable and efficient computing services were provided for edge applications. Simulation results show that the proposed algorithm could effectively reduce the processing latency, improve the network processing rate and robustness.

Key words: Internet of things, heterogeneous network, mobile edge computing, task assignment, resource allocation

1 引言

随着科技的发展，人们身边的智能化设备和传感器数量爆发式增长，未来向着万物互联互通的 6G 物联网时代稳步迈进。所谓物联网，是互联网、传统电信网等的资讯承载体，是让所有能行使独立功能的普通物体实现互联互通的网络[1]。5G 移动通信技术有 3 类重要的应用场景，包括增强型移动宽带（eMBB, enhanced mobile broadband）、大规模机器类通信（mMTC, massive machine type communica-

收稿日期：2020−01−13；**修回日期**：2020−03−04
通信作者：宋令阳，lingyang.song@pku.edu.cn
基金项目：国家自然科学基金资助项目（No.61625101，No.61941101）
Foundation Item: The National Natural Science Foundation of China (NSFC) (No.61625101, No.61941101)

tion）和高可靠低时延通信（uRLLC, ultra-reliable and low latency communication）。5G 的演进过程以及将来的 6G 时代，会对传输速率、覆盖范围、时延、通信可靠性、用户接入量等指标提出更高、更极致的要求，进而出现融合多种需求的新应用场景，如完全意义上的自动驾驶、针对超大用户量且融合多重感官信息的虚拟现实/增强现实服务等。同时，数据生产边缘化使得众多终端设备兼具网络数据的消费者和生产者身份。在 6G 时代，海量的原始数据在无线网络边缘产生并汇入通信网络，不仅占用大量的带宽资源，还对快速、可靠的传输和计算提出了巨大的挑战。然而无线通信带宽有限，有线网络带宽也不可能无限增加，并且由于远离计算中心，5G 通信网络的边缘设备计算能力十分有限，当前的计算网络架构难以满足广大终端日益增长的任务处理需求[2]。

云计算提出了一种有效应对数据爆发的方案，即终端设备没有能力处理的计算任务可以上传到云计算中心处理[3]。云计算中心拥有大规模的计算设备，能够提供强大的计算能力和集中化的管理保障，为终端设备提供丰富的网络和计算服务。然而，云计算的瓶颈与限制也显而易见。首先，大规模的云计算中心通常部署在远离终端用户的地方，卸载到云计算中心的任务需要经历较长的传输时延，难以适用于实时性较强的计算任务和应用，如智慧交通场景中，车辆的碰撞检测和预警所要求的毫秒级时延很难在云计算中实现。其次，大规模的原始数据上传到云计算中心需要占用大量的带宽资源，不仅为有限的传输带宽带来了巨大压力，还制造了巨大的计算成本。以视频监控为例，数以百万计的高清摄像头遍布全国，要实现如此大规模视频数据的云端处理，需要耗费大量的带宽资源来进行数据传输。

为了解决云计算所面临的困境，移动边缘计算（MEC, mobile edge computing）应运而生。移动边缘计算是指将云端的计算能力和网络服务下放到通信网络边缘，即无线接入网中，使用户可以在更邻近的无线接入点（AP, access point）获取计算服务[4]。因此，当无线接入网中的边缘设备（ED, edge device）有待处理的计算任务时，可以将原始数据卸载到邻近的边缘服务器上进行数据处理，而不必传到遥远的云服务器，这大大缩短了响应时间、提升了处理效率。然而，边缘服务器的计算能力有限，远没有云计算中心的计算能力强大，因此，面对大量计算密集型任务时仍然难以为继。故而大多数工作中只考虑直接与边缘设备相连的边缘服务器是远远不够的[5]。

本文考虑了云计算和多层边缘计算相结合的异构多层边缘计算（HetMEC, heterogeneous multi-layer mobile edge computing）架构，使得在下层边缘服务器无法处理的计算任务可以继续卸载到上层边缘服务器，直至云计算中心，从而可以有效避免网络阻塞和数据堆积，充分利用多层边缘服务器和云计算中心的算力，有效减少系统时延。在运行 HetMEC 架构的网络中，边缘节点到云计算中心的上行链路中包含多层不同功能的网络节点，自下而上包括 AP、交换机、网关、小型数据中心等[6]，涵盖了局域网、城域网、广域网等多个层级。具备计算能力的功能节点根据位置不同作为不同层级的边缘服务器，计算能力不同、覆盖范围不同，越向上层则计算能力越强、覆盖范围越大、距离边缘设备也越远。

HetMEC 架构面临着诸多挑战：1) 边缘设备的计算任务可以分割成若干部分，然而各层的任务分配受限于相应服务器可调用的计算资源，因此，各层间的任务分配互相耦合，不能单一考虑；2) 有线通信和无线通信网络中的传输资源分配与各层所卸载的任务量密切相关、互相耦合，具体来说，传输资源直接限制了相邻两层之间的传输速率，即最多可以卸载的任务量；3) 针对不同的数据产生速率，由于各个计算设备（包括边缘设备、不同层的边缘服务器以及云计算中心）所拥有的计算和传输资源有限，系统的稳健性需要纳入考虑，即如何才能让 HetMEC 网络在更大的数据产生速率范围内避免网络阻塞。

本文贡献如下：1) 考虑了由边缘设备、多层边缘服务器和云计算中心所构成的 HetMEC 架构，考虑典型的上行场景，即数据从边缘设备产生，经过多层计算和传输直至所有计算结果汇总到云计算中心，研究了上传过程中任务在各层的卸载比例以及各个计算设备计算和传输资源分配策略的制定；2) 为了降低 HetMEC 网络的任务处理时延，综合考虑了多层耦合的任务分配、计算资源分配和传输资源分配，设计了多层联合任务卸载和资源分配（MTR）算法最优化系统时延，即所有设备的计算和传输总时延；3) 分析并研究了系统稳健性与网络

层数、节点资源多少的关系，从网络结构优化的角度给出了增强网络稳健性的策略，即基于当前的资源配置，在何处增加边缘服务器才能增强网络的稳健性。

2　系统模型

针对前文所描述的 6G 应用对传输速率、覆盖范围、时延、通信可靠性、用户接入量等指标更高、更极致的需求以及数据生产边缘化的特点，单一的云计算中心和计算力受限的单层边缘服务器很难在保证海量用户接入和通信可靠性的前提下，满足时延和处理速率极高的要求。因此，在异构边缘计算网络中综合考虑了云计算中心和多层边缘服务器，这样不仅解决了边缘服务器计算力不足的问题，还能够根据边缘场景中不同的应用需求更合理地调配高层服务器的空闲计算资源，进而在保证可靠性和用户接入的前提下，明显提高处理能力，缩短处理时延，解决 6G 所面临的诸多挑战。

异构边缘计算网络包括最上层的云计算中心、中间多层边缘服务器和底层的边缘设备，以应对众多物联网应用场景，如健康监测、智慧家庭、安全监控等，异构边缘计算网络架构如图 1 所示。其中，边缘设备负责采集原始数据，经由本地计算将部分计算结果和剩余的原始数据通过无线链路上传至 AP，再经过逐层的有线传输和计算后，将结果汇总到云计算中心。考虑一个包含 N 层边缘服务器的 HetMEC 网络，从上至下设备所在层的编号从 1 到 N，第 n 层上有 M_n 个计算设备，该层上的设备 i 与第 n+1 层所连接的子设备数量为 Q_n^i。云计算中心和每个边缘服务器都具备一定的传输资源（其中，第 N 层边缘服务器即 AP，为无线传输资源，其余均为有线传输资源），连接到同一节点的多个下层设备通过时分多址（TDMA, time division multiple access）技术接入无线或有线信道，而不同 AP 的无线传输频段是相互正交的，不同小区间不考虑干扰。不同计算设备的计算过程和传输过程可以并行，待上传处理的原始数据不需要等待该层计算结束后和结果一起传输，可以先上传至上层计算，以减少不必要的等待时间。

以准静态的角度考虑一个时隙内所产生的数据，计算设备的计算资源可以用该设备在单个时隙内所能处理的数据量表示。在进行 TDMA 接入传输的过程中，上层设备根据各个子设备传输的数据量占其总接收数据量的比例，来分配时间资源。因此，传输资源的多少可以表示为单个时隙内的数据传输量。

接下来分别从边缘设备、边缘服务器和云计算中心 3 个部分详细介绍不同设备的任务卸载模型、计算模型和传输模型。

1) 边缘设备

边缘设备位于 HetMEC 网络底层，包括智能手表、智能摄像机、智能手机等物联网设备，负责收集和产生原始数据。为了方便表示，将边缘设备视作第 N+1 层，每个边缘设备可以处理一部分数

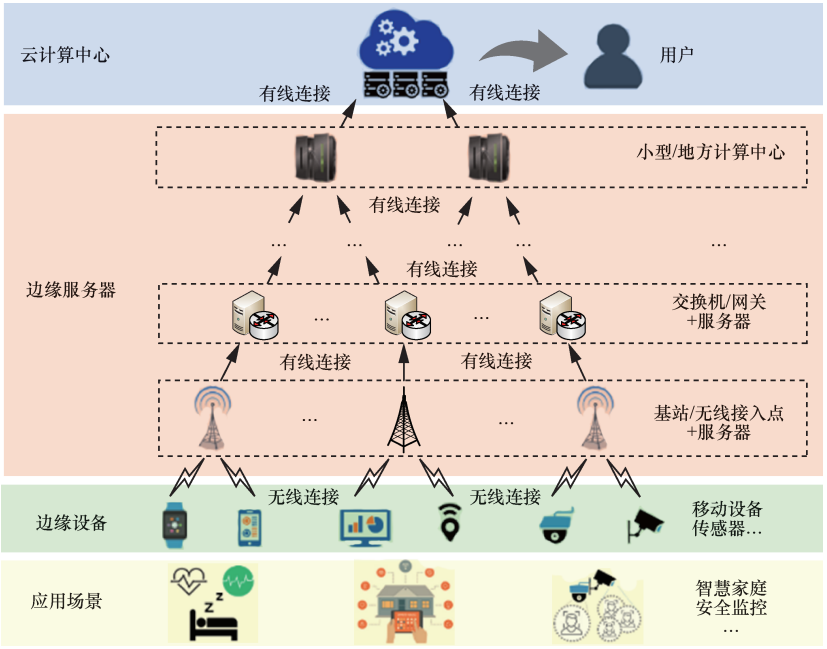

图 1　异构边缘计算网络架构

据，并将计算结果和剩余的原始数据上传到边缘服务器。

用 λ_{N+1}^{i} 表示边缘设备 i 处的数据产生速率，b_{N+1}^{i} 表示所需要的 CPU 执行循环数，针对同一类型任务，b_{N+1}^{i} 与 λ_{N+1}^{i} 是成比例的。边缘设备 i 的计算资源量和从第 N 层的父节点 j 处所分到的传输资源量分别用 θ_{N+1}^{i} 和 $\phi_{N+1}^{j,i}$ 表示。

2) 边缘服务器

边缘服务器从下到上分别位于第 N 层，…，第 1 层。相邻层之间的数据传输包括 3 个部分，以第 $n+1$ 层节点 i 到第 n 层节点 j 为例，分别是第 $n+1$ 层的计算结果、第 $n+1$ 层上传的原始数据和第 $n+1$ 层收到的下层的计算结果。而第 $n+1$ 层节点 i 到第 n 层节点 j 的数据到达速率为 $\lambda_n^{j,i}$，取决于所有传输数据中原始数据量的比例。

类似地，用 λ_n^j 表示第 n 层的边缘服务器 j 处的数据到达速率，θ_n^j 和 $\phi_n^{k,j}$ 分别表示第 n 层的边缘服务器 j 的计算资源量和从第 $n-1$ 层的父节点 k 处所分到的传输资源量。

3) 云计算中心

云计算中心位于整个 HetMEC 网络的最上层，汇集所有计算结果并反馈给用户，所有传给云计算中心的原始数据都由其负责处理，为了方便表示，将云计算中心定义为第 0 层。

3 时延优化和算法设计

本文旨在优化 HetMEC 网络在没有网络阻塞情况下的系统时延，下面分别讨论网络非阻塞约束、系统时延的定义与优化问题的描述，设计了最优的多层联合任务卸载和资源分配算法来最小化系统时延。

3.1 网络约束

网络阻塞发生的情况包括以下两种：1) 节点的计算能力不足以处理其所分到的计算任务；2) 节点的传输能力无法将该时隙内所需要传输的数据都上传到上层。因此，非阻塞的网络约束包括以下 3 种。

1) 所有任务都被完整分割处理。即每个边缘节点的任务被分割成若干部分，每一部分都在相应的节点处（边缘节点、某层边缘服务器或云计算中心）被处理完成，将计算结果汇总到云计算中心。

2) 对于每个网络节点，其所分到的计算任务量不超过本身的计算能力。

3) 对于每个网络节点，其所分到的传输资源不小于需要传输的数据量（包括计算结果和原始数据）。

3.2 系统时延

从数据处理的角度看，系统时延是指 HetMEC 网络中所有层上的计算设备的计算时间和传输时间的总和。

3.3 优化问题描述

给出系统时延最小化的数学形式描述，定义在第 n 层的节点 i 处的任务卸载比例为 s_n^i，原始数据 λ 经过处理后，数据量变小为 $\rho\lambda$，压缩比例为 ρ，那么 HetMEC 网络中第 n 层所有节点的计算时间和传输时间之和 L_n 可以表示为

$$L_n=\sum_{j=1}^{M_{n-1}}\sum_{i\in Q_{n-1}^j}\left[\frac{s_n^i b_n^i}{\theta_n^i}+\frac{\rho s_n^i\lambda_n^i+(1-s_n^i)\lambda_n^i+\beta_n^i}{\phi_n^{j,i}}\right] \tag{1}$$

其中，$\frac{s_n^i b_n^i}{\theta_n^i}$ 表示第 n 层节点 i 的计算时间，β_n^i 表示第 n 层节点 i 收到的下层上传的计算结果的数据量，$\rho s_n^i\lambda_n^i$ 表示该节点的计算结果数据量，$(1-s_n^i)\lambda_n^i$ 表示该节点处待上传的原始数据量，因此，$\frac{\rho s_n^i\lambda_n^i+(1-s_n^i)\lambda_n^i+\beta_n^i}{\phi_n^{j,i}}$ 为第 n 层节点 i 的传输时间。

用 L_0 表示云计算中心的计算时间，那么，包含 N 层边缘服务器的 HetMEC 网络的系统时延可以表示为

$$L=L_0+\sum_{n=1}^{N+1}L_n \tag{2}$$

3.4 MTR 算法设计

针对 HetMEC 的系统时延最小化，设计了多层联合任务卸载和资源分配算法，简称 MTR 算法，该算法的核心在于利用柯西不等式[7]将相互耦合的任务卸载比例和计算、传输资源分配进行解耦，利用等号成立条件构建任务分配比例和计算、传输资源分配之间的关系，将困难的联合优化问题简化成只含一个变量的任务卸载问题。

目标函数是一个耦合了任务卸载比例 s、计算资源分配 θ 和传输资源分配 ϕ 这 3 个变量的函数，为了最小化系统时延 $L(s,\theta,\phi)$，利用柯西不等式可以得到

$$L_n\geqslant\sum_{j=1}^{M_{n-1}}\sum_{i\in Q_{n-1}^j}\frac{s_n^i b_n^i}{\theta_n^{i,\mathrm{MAX}}}+\sum_{j=1}^{M_{n-1}}\frac{\left(\sum\limits_{i\in Q_{n-1}^j}\sqrt{\rho s_n^i\lambda_n^i+(1-s_n^i)\lambda_n^i+\beta_n^i}\right)^2}{\phi_{n-1}^j} \tag{3}$$

其中，$\theta_n^{i,\mathrm{MAX}}$ 为第 n 层节点 i 的最大计算能力，ϕ_{n-1}^{j} 为第 $n-1$ 层节点 j 的所有传输资源，这两个量均为系统参数。在满足等式成立的条件时，原时延最小化问题便完全等价于任务卸载问题 $L_{\min}(s)$，而等号成立条件则给出了任务卸载策略对应的网络计算资源和传输资源分配方案，即只要按照等号成立条件所给出的资源分配方案配置网络资源，任务卸载问题求得的最优解就是原联合时延优化问题的最优解。

接下来给出任务卸载问题的目标函数性质和最优化算法。

命题　任务卸载问题的目标函数 $L_{\min}(s)$ 为凹函数，有最大值。

证明　$L_{\min}(s)$ 的 M 阶 Hessian 矩阵为 $\boldsymbol{H}_M$，另外考虑向量 $\boldsymbol{x}=[x_1,x_2,\cdots,x_M]^{\mathrm{T}}$，利用数学归纳法可以证明多项式 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{H}_M\boldsymbol{x}\leqslant 0$ 对于任意 M 恒成立，则 $\boldsymbol{H}_M$ 是半负定的，因此，转化后的目标函数 $L_{\min}(s)$ 为凹函数，有最大值。

因此，在当前线性限制条件下，系统时延最小值一定在限制条件的某一个交点上，因此，最优解可以通过遍历所有限制条件的交点得到。算法步骤如下：

1) 依据式(3)，利用柯西不等式，增加等号成立条件的约束后，原系统时延最小化问题完全等价于任务卸载问题 $\min\limits_{s} L_{\min}(s)$。

2) 根据所有线性限制条件，利用线性方程组求解可以计算得到所有限制条件交点，将这些交点的集合定义为 $\mathcal{S}$。对集合中的所有交点 $s\in\mathcal{S}$，求解 $L_{\min}(s)$，并将结果与当前记录的最小值进行比较，如果结果小于记录的最小值，那么更新最小值并记录当前的解，直到遍历所有交点后找到使得 $L_{\min}(s)$ 最小的一组解 s^*，即最优的任务卸载比例。

3) 根据柯西不等式的等号成立条件，利用最优解 s^* 得到最优的计算资源分配 θ^* 和传输资源分配 ϕ^*。

4　网络稳健性分析

稳健性是指 HetMEC 网络在不阻塞的情况下，边缘设备的最大数据产生速率反映了运行 MTR 算法时网络所能处理的数据产生速率的动态范围，衡量了整个 HetMEC 网络在变化的边缘设备处理需求的情况下的稳定性和处理能力。本文通过讨论网络非阻塞的临界情况来分析网络层数对 HetMEC 网络稳健性的影响，分为两种情况：计算资源限制情况和通信资源限制情况。

4.1　计算资源限制情况

若网络阻塞的发生是由计算资源不足导致，则称为计算资源限制情况，即所有计算设备均使用了最大的计算力进行数据计算，但是仍有节点分到的数据量超过了其计算能力。在这种情况下，引入新一层的边缘服务器时，只要满足新引入的边缘服务器通信资源不受限，增加的计算资源便可使 HetMEC 网络在单位时间内处理更多的数据，进而提升网络的稳健性。下面通过一个示例来分析说明计算资源限制下 HetMEC 网络的稳健性，计算资源限制下 HetMEC 网络的稳健性分析如图 2 所示。

当计算资源限制下的 N 层 HetMEC 网络处于非阻塞的临界状态时，所有节点分配到的数据计算量刚好等于其最大的计算能力，如果数据产生速率继续增加，那么新增的数据无论分到哪层，都会出现网络阻塞。当引入一层新的边缘服务器 MEC* 时，可以分担其余所有层的计算压力，通过执行 MTR 算法，计算负载被分散，其余层分配的数据计算量减少，整个网络可以处理更多的数据，也可以应对更大的数据产生速率，稳健性增强。

4.2　通信资源限制情况

若发生网络阻塞是由某层的传输资源不足导致的，则称为传输资源限制情况，即当前层已经使用了所有的传输资源来进行原始数据和计算结果的传输，但是待传输的数据量仍然超过了其传输能力。传输资源限制下 HetMEC 网络的稳健性分析如图 3 所示。N 层 HetMEC 网络中，待传输的数据量随着计算比例的增加而逐渐减少，因此这种类型的网络阻塞一定发生在该层及其下面各层的计算力已经用满的情况下，即当前传输的数据量已经是非阻塞情况下网络传输的最小数据量，新增的数据无法再成功传到网络上层。

这种情况下，有两种思路提升网络的稳健性。

1) 减少当前阻塞层的数据传输量。当阻塞层或阻塞层之下引入新一层通信资源不受限的边缘服务器时，新引入的计算资源可以进一步减少原阻塞层的待传输数据量，进而增强 HetMEC 网络的稳健性。如图 3 所示，当引入 MEC* 后，原来传输资源受限的

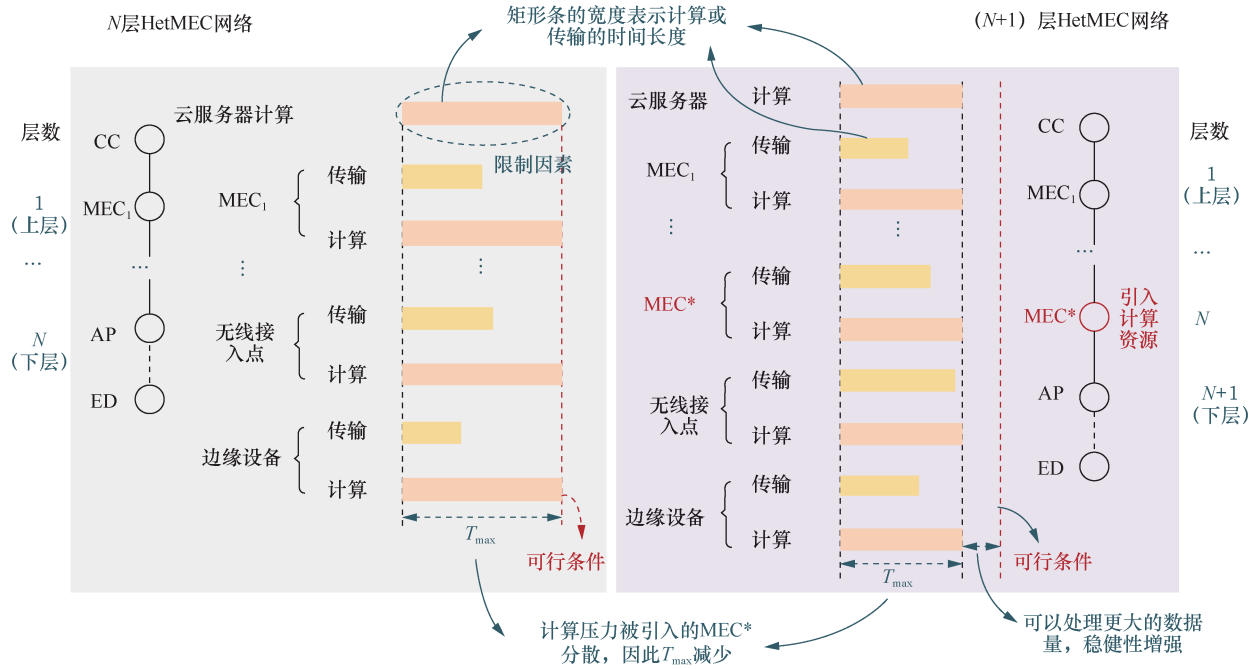

图 2　计算资源限制下 HetMEC 网络的稳健性分析

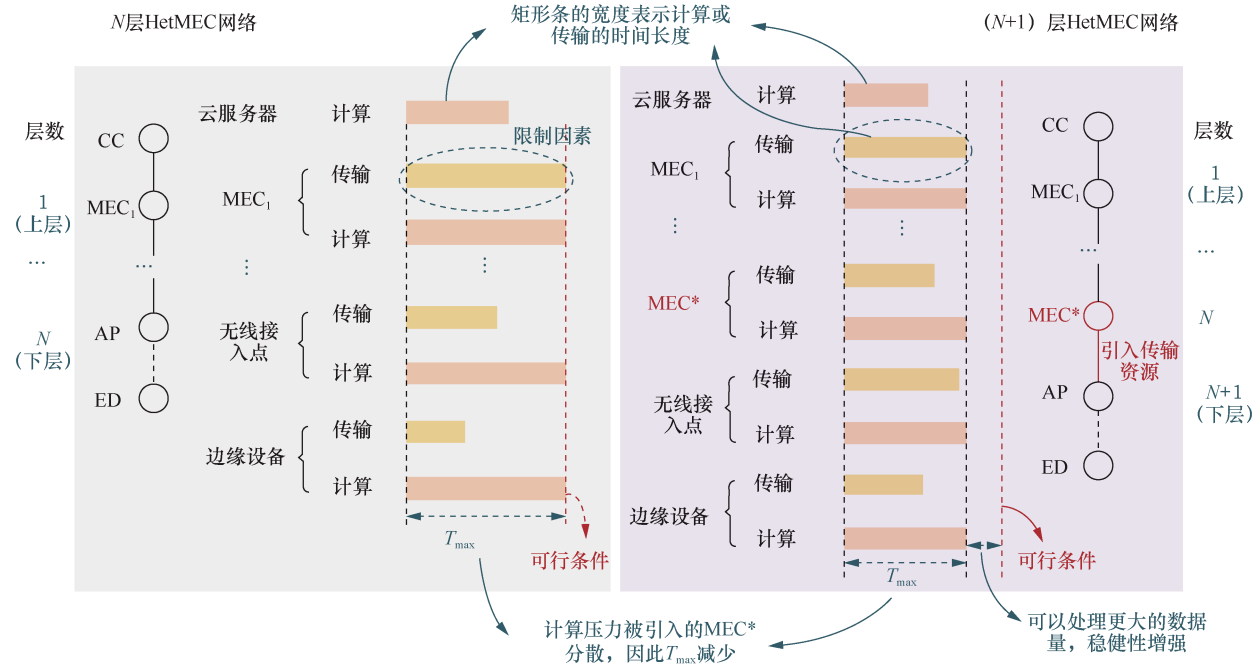

图 3　传输资源限制下 HetMEC 网络的稳健性分析

MEC_1 层及下面各层的计算负载可以分散到新加入的 MEC^* 层，因此，整个网络可以处理更多的数据，也可以应对更大的数据产生速率，稳健性增强。

2) 增强阻塞层的数据传输能力。在当前的阻塞层上面相邻的一层引入新的边缘服务器（因为阻塞层的传输资源是上层分配的），如果这些边缘服务器所拥有的传输资源比原有的传输资源多，那么新引入的边缘服务器可以分担下面各层的计算压力，进而减少原阻塞层的传输数据量，可以有效增强 HetMEC 网络的稳健性。

根据以上分析可知，在通信资源限制情况下，假设阻塞层为第 n 层，那么在阻塞层之上两层，即 $n-2$ 层及以上引入新的边缘服务器，无法提升 HetMEC 网络的稳健性。

5　实验过程

采用 MATLAB 仿真实验平台对算法和网络性能进行测试和验证。考虑一个包含云计算中心、3 层边缘服务器节点和一层边缘节点的网络，不同情况下的 HetMEC 计算架构如表 1 所示，原始数据经过处理后，计算结果的压缩比 ρ 平均为 10%，计算能力表示为 CPU 每秒最大的处理次数，传输能力表示为传输带宽的大小，考虑 4 种情况分别是：云计算网络、单层 HetMEC 网络（即仅有一层边缘服务器参与计算）、双层 HetMEC 网络（即底层和中层的边缘服务器参与计算）和 3 层 HetMEC 网络，HetMEC 网络中各节点的计算资源和传输资源配置如表 2 所示。不同的情况，计算架构不同，但是基于相同的网络结构、节点数量和计算资源量。

实验流程如图 4 所示，以双层 HetMEC 网络的运行为例，整个过程分为以下 6 步。1) 任务发布：用户向云计算中心发起任务，云计算中心利用广播的方式向 HetMEC 网络中所有相连的计算设备逐层发布计算任务；2) 节点注册：参与计算任务的节点，

表 1　**不同情况下的 HetMEC 计算架构**

	云计算网络	单层 HetMEC 网络	双层 HetMEC 网络	3 层 HetMEC 网络
边缘节点	√	√	√	√
参与计算的边缘服务器层数	0	1（底层）	2（底层+中层）	3（底层+中层+高层）
云计算中心	√	√	√	√

表 2　**HetMEC 网络中各节点的计算资源和传输资源配置**

节点	计算资源/Mcps	传输资源/(Mbit·s^{-1})
边缘节点	0.12	—
底层边缘服务器	0.4	1.2
中层边缘服务器	1.5	3
高层边缘服务器	4.2	4.8
云计算中心	12	12

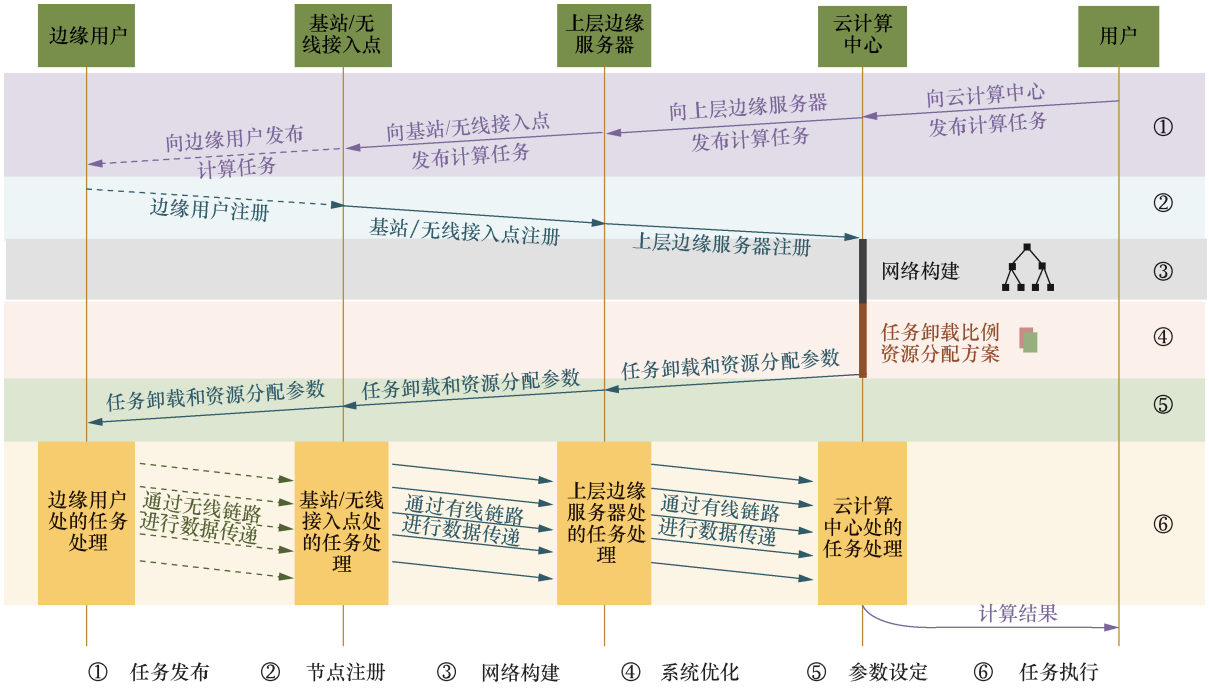

图 4　实验流程

包括边缘设备和各层的边缘服务器，逐层向上注册节点信息，包括节点所在层、IP 地址、端口号、计算资源和传输资源量；3) 网络构建：云计算中心根据 HetMEC 网络中活跃节点的注册信息，构建网络连接关系和资源配置情况；4) 系统优化：根据设计的 MTR 算法优化系统的任务分配比例和资源分配方案；5) 参数设定：将优化结果，即最优的任务分配比例和计算、传输资源分配方案发放给各个活跃的 HetMEC 网络节点；6) 任务执行：各个计算设备根据收到的参数配置，开始进行任务处理和上传。

6 实验结果

本文从 3 个方面衡量 HetMEC 网络的性能，分别是系统时延性能、处理速率性能和网络稳健性。另外，选取了 4 种对比策略与 MTR 算法进行比较，分别是云计算策略（即所有数据都卸载到云计算中心处理）、本地计算策略（即所有数据都在边缘设备中处理）、传统 MEC 策略（即所有数据都卸载到最下层边缘服务器处理）和 JCORA 算法[8]（即最下层边缘服务器和边缘设备交替调整计算卸载策略和资源分配方案）。

6.1 系统时延性能

在不同数据产生速率 λ 的情况下，单层和双层 HetMEC 网络运行不同的任务卸载算法的系统时延如图 5 所示。从仿真实验结果可以看出，采用设计的 MTR 算法可以有效降低整个 HetMEC 网络的系统时延，算法性能在计算压力较大的高数据产生速率情况下的优势尤其明显，例如在 $\lambda > 6$ 时，其他任务卸载策略都导致了明显的网络阻塞，使系统时延明显增加，而 MTR 算法有效利用了多层边缘服务器和云计算中心的计算和传输资源，分散了网络的计算压力，使系统时延大大降低。可以注意到，相较于单层 HetMEC 网络，双层 HetMEC 网络中 λ 增加到了 11，这也体现了边缘服务器层数的增加对网络稳健性的提升。

6.2 处理速率性能

在不同数据产生速率 λ 的情况下，单层和双层 HetMEC 网络运行不同的任务卸载算法的系统处理速率如图 6 所示。系统处理速率定义为整个 HetMEC 网络平均每单位时间内所处理的数据量，由于不同计算设备的计算和传输过程可以并行，而且数据处理可能存在空窗阶段，因此，系统处理速率为处理的总数据量与实际执行时间的商。可以看出，无论在单层还是双层 HetMEC 网络中，在非阻塞的网络状态下，系统处理速率都会随着边缘设备数据产生速率的增加而增加，因为这种情况下更密集的数据到达率可以有效减少设备计算资源和传输资源的空窗阶段，从而提高了处理速率。但是相较于 MTR 算法，其余卸载策略和算法都比较早地达到了速率饱和，这是因为某个节点处已经没有空闲资源而造成了网络阻塞。对比单层和双层 HetMEC 网络中 MTR 算法的性能，可以看出双层 HetMEC 网络的系统处理速率更高，稳健性也更强。增加的一层边缘服务器可以有效分担其他计算设备的计算压力，以更高的并行性缩短实际的执行时间，因此可以获得更高的系统处理速率。

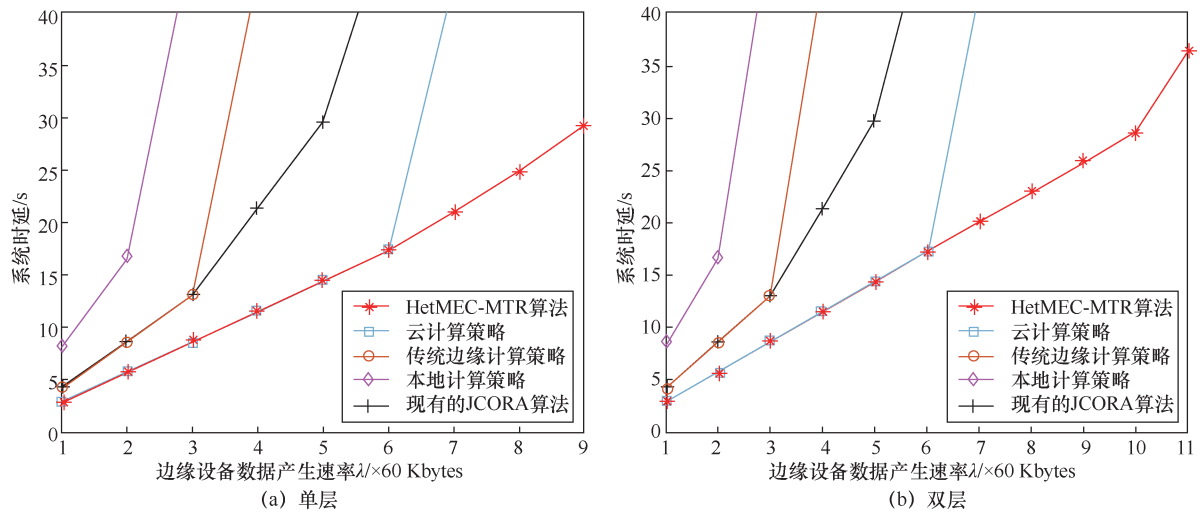

图 5　单层和双层 HetMEC 网络运行不同的任务卸载算法的系统时延

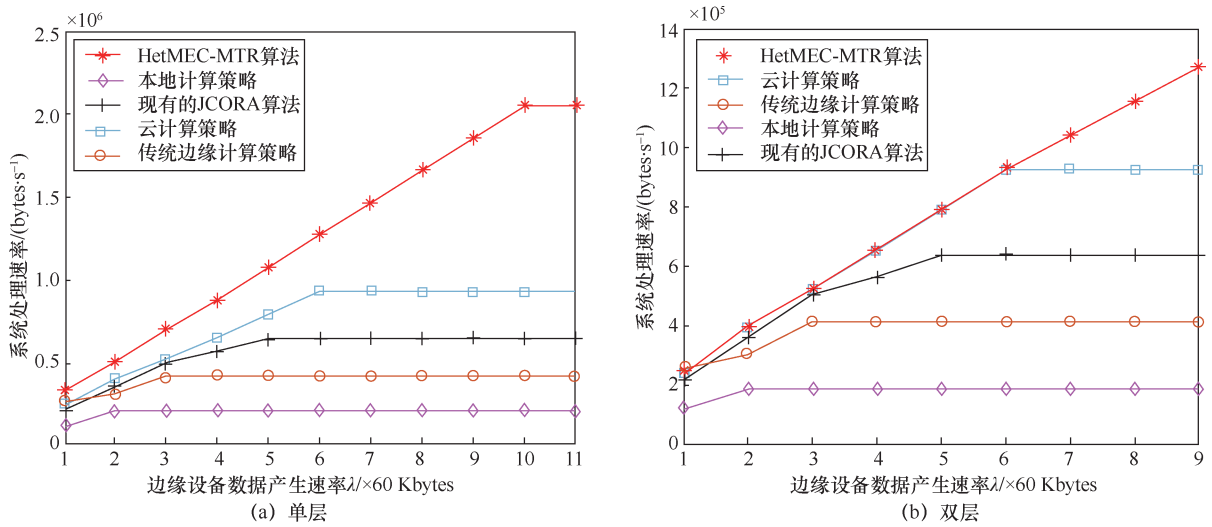

图 6　单层和双层 HetMEC 网络运行不同的任务卸载算法的系统处理速率

6.3　网络稳健性

不同资源配置情况下不同层的 HetMEC 网络稳健性比较如图 7 所示。单考虑云计算网络和单层到 3 层的 HetMEC 网络，分为 3 种情况，HetMEC 网络不同情况下的资源配置情况如表 3 所示。

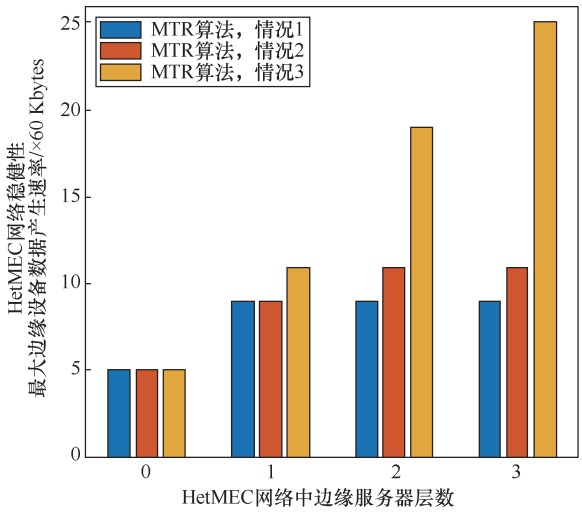

图 7　不同资源配置情况下不同层的 HetMEC 网络稳健性比较

在情况 1 中，边缘服务器有效增强了网络的稳健性，但是增加边缘服务器的层数并没有继续提升 HetMEC 网络的稳健性。因为最初没有边缘服务器时，网络性能受限于计算资源，因此引入边缘服务器的计算资源后，网络的稳健性增强；而引入底层的边缘服务器后，网络性能主要受限于底层边缘服务器的传输资源，根据在上一部分的分析可知，在底层边缘服务器之上引入新的边缘服务器并不能增加网络的稳健性。

在情况 2 中，当边缘服务器少于两层时，HetMEC 网络稳健性随着边缘服务器层数的增加而增强，当边缘服务器少于两层时，HetMEC 网络稳健性不再随着边缘服务器层数的增加而变化。这是因为在边缘服务器层数 $N<2$ 时，网络性能受限于计算资源，而边缘服务器层数 $N\geqslant 2$ 时，网络性能受限于中层边缘服务器的传输资源，因此，之后网络稳健性不再随边缘服务器层数的增加而增强。

在情况 3 中，所有网络架构下，均为计算资源限制情况，因此，增加传输资源不受限的边缘服务

表 3　HetMEC 网络不同情况下的资源配置情况

计算设备	情况 1		情况 2		情况 3	
	计算资源/Mcps	传输资源/(Mbit·s^{-1})	计算资源/Mcps	传输资源/(Mbit·s^{-1})	计算资源/Mcps	传输资源/(Mbit·s^{-1})
边缘设备	0.12	—	0.12	—	0.12	—
底层边缘服务器	0.4	0.9	0.4	1.2	0.4	3
中层边缘服务器	1.5	3	1.5	3	1.5	6
高层边缘服务器	4.2	4.8	4.2	4.8	4.2	12
云计算中心	12	12	12	12	12	15

器能够有效提升 HetMEC 网络的稳健性。

7 结束语

本文提出了云计算与 MEC 结合的 HetMEC 架构，以提供低时延的数据处理服务，研究了典型的任务卸载上传场景，即原始数据在边缘设备处产生采集，并被分割成多份，经由多层边缘服务器的传输和计算，最终计算结果汇聚在云计算中心。设计了 MTR 算法联合优化了 HetMEC 网络中的任务卸载比例、计算资源分配和传输资源分配，有效分散了计算压力、降低了系统时延并提升了网络的稳健性。经分析，网络的稳健性与网络架构和计算、传输资源配置情况密切相关。HetMEC 架构为云边融合的万物互联网络和众多低时延的物联网应用场景提供了新的思路和理论基础，能够让广大用户更方便、快捷地享受计算服务。

参考文献：

[1] ATZORI L, IERA A, MORABITO G. The Internet of things: a survey[J]. Computer Networks, 2010, 54(15): 2787-2805.

[2] PAPAGEORGIOU A, CHENG B, KOVACS E. Real-time data reduction at the network edge of Internet-of-things systems[C]//The 11th International Conference on Network and Service Management (CNSM). IEEE, 2015: 284-291.

[3] ARMBRUST M, FOX A, GRIFFITH R, et al. A view of cloud computing[J]. Communications of the ACM, 2010, 53(4): 50-58.

[4] ETSI. Mobile edge computing: a key technology towards 5G[S]. 2015.

[5] MAO Y, YOU C, ZHANG J, et al. A survey on mobile edge computing: the communication perspective[J]. IEEE Communications Surveys and Tutorials, 2017, 19(4): 2322-2358.

[6] TANENBAUM A S, WETHERALL D. Computer networks[M]. Englewood: Prentice Hall, 1996.

[7] STEELEJ M. The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities[M]. Cambridge: Cambridge University Press, 2004.

[8] WANG C, YU F R, LIANG C, et al. Joint computation offloading and interference management in wireless cellular networks with mobile edge computing[J]. IEEE Transactions on Vehicular Technology, 2017, 66(8): 7432-7445.

[作者简介]

王鹏飞（1994− ），男，河北保定人，北京大学信息与通信研究所硕士生，主要研究方向为无线通信、边缘计算和车载通信等。

邸博雅（1992− ），女，黑龙江大庆人，北京大学信息与通信研究所博士生，主要研究方向为无线通信、边缘计算、车载网络、智能反射面和非正交多址接入等。

宋令阳（1979− ），男，辽宁抚顺人，北京大学信息与通信研究所教授，主要研究方向为无线通信、编码技术、MIMO、OFDMA、认知无线电以及协同通信等。

韩竹（1974— ），男，北京人，休斯顿大学讲席教授，主要研究方向为博弈论、无线通信、网络安全、数据科学以及智能电网等。