****

**本科毕业论文（设计）开题报告**

**题 目：**

**专 业 软件工程**

**学 生 付书煜**

**学 号 2021111824**

**指导教师**

**日 期**

**哈尔滨工业大学教务处制**

**目录**

[1 课题来源及研究的目的和意义 3](#_Toc16239)

[2 国内外在该方向的研究现状及分析 3](#_Toc10033)

[3 主要研究内容 4](#_Toc26839)

[3.1 云-边协同架构的系统建模 4](#_Toc25555)

[3.2 响应时间差异最小化的公平性目标设计 4](#_Toc29441)

[3.3 多约束优化建模 5](#_Toc17729)

[3.4 模型公平性优化效果的评估标准 5](#_Toc3734)

[4 研究方案 5](#_Toc29704)

[4.1 需求分析与系统建模 5](#_Toc14925)

[4.2 目标函数和约束条件的设计和实现 5](#_Toc4833)

[4.3 初始值设定与仿真数据生成 5](#_Toc6509)

[4.4 模型求解与实现 5](#_Toc23962)

[4.5 实验验证与性能分析 5](#_Toc12177)

[5 进度安排，预期达到的目标 6](#_Toc11364)

[5.1 进度安排 6](#_Toc13112)

[5.2 预期目标 6](#_Toc26247)

[6 课题已具备和所需的条件、经费 6](#_Toc30247)

[7 研究过程中可能遇到的困难和问题，解决的措施 6](#_Toc1997)

[8 主要参考文献 6](#_Toc6581)

# 课题来源及研究的目的和意义

随着互联网迅速发展，微服务架构[1]作为一种高效、灵活的架构模式，正在被广泛应用于各种需要高实时性和灵活性的服务场景，如视频流媒体[2]、在线游戏等。微服务将应用程序划分为多个独立、细粒度的服务模块，使其能够独立扩展和部署，并通过轻量级通信协议相互协作，从而提供了高度的灵活性和扩展性[3]。

为了进一步优化微服务的响应速度和服务质量体验（QoE），云-边协同计算模式应运而生。云端拥有丰富的计算资源，能够高效处理大规模数据和复杂的计算任务，但由于其物理位置往往远离用户，可能带来较高的网络传输延迟[4，5]。边缘节点位于用户附近[6]，可以就近处理用户请求，显著降低网络传输延迟[7]。然而，边缘节点的计算资源相对有限，在高并发访问或者复杂任务处理中，其资源限制可能导致性能瓶颈，增加延迟。因此，云-边协同计算通过在云端和边缘节点之间合理分配任务，既能利用云端丰富的资源以及强大的计算能力，又能借助边缘节点的低延迟优势，优化资源利用，提升用户体验[8]。

尽管现有的云-边协同部署方案在整体上提升了服务质量，但其通常聚焦于最大化整体QoE[9]，往往忽略了不同用户间的体验差异，即公平性问题[10]。由于地理位置、网络条件以及节点资源负载等差异，不同用户的响应时间可能有较大波动，部分用户的体验可能显著劣于其他用户[11]。这种差异在对实时性和一致性要求较高的应用场景中尤其突出。例如，在视频流媒体服务中，远离边缘节点的用户可能因更高的网络传输延迟而产生卡顿，导致体验显著下降。

因此，如何在云-边协同环境中缩小不同用户之间的服务质量差异，实现公平性部署，成为一个亟待研究的课题。随着用户对服务一致性要求的提升，公平性逐渐成为影响用户整体满意度的关键因素。在保障高QoE的的前提下，通过优化服务实例的部署和资源分配来减少用户之间的响应时间差异，不仅能够提升用户整体的满意度，也能推动云-边协同计算在更多应用场景中的有效普及。因此，在云-边协同计算中实现微服务的公平性部署，具有重要的研究价值和广阔的应用前景。

# 国内外在该方向的研究现状及分析

近年来，随着云计算和边缘计算的快速发展，微服务在云-边协同环境中的部署及资源管理成为了研究热点。我们围绕整体服务质量体验（QoE）优化以及公平性优化两个方面进行了调研，来分析当前的研究现状。

为了提升整体QoE，一些研究者提出了基于用户位置的任务分配方法，降低了用户请求的响应时间。Liu等人通过边缘计算的低延迟特性，设计了一种分布式任务调度策略，将用户任务在边缘节点进行就近处理，显著提升了系统对高并发请求的响应速度[12]。Alsurdeh等人提出了一种混合工作流调度方法，通过在云端和边缘节点之间协同分配任务，提高了资源利用率，同时显著增强了整体的QoE[13]。这些研究聚焦于系统整体QoE的提升，表明云-边协同框架在增强资源利用、降低延迟方面有良好效果。然而，其却忽略了不同用户因地理位置差异而可能产生的QoE不均衡问题。特别是在远离边缘节点的用户中，QoE可能显著降低，从而影响整体的用户体验公平性。

随着公平性需求的提升，部分学者也逐渐关注不同用户体验的一致性，并致力于优化资源分配与任务调度，以减少不同用户的QoE差异。Zhang等人基于图神经网络提出了一种工作负载迁移方案，通过在边缘节点间动态调整资源分配，有效平衡了不同地理位置用户的服务质量体验（QoE），从而显著降低了用户之间的QoE差异[14]。然而，该方法依赖复杂的迁移机制，需进一步提升其实用性和计算效率，尤其是在大规模分布式环境中。Hao等人提出了一种基于深度强化学习（DRL）的计算卸载方案，通过学习算法在边缘节点之间动态分配计算任务，以减少用户体验差异并提升资源利用率[15]。此外，Zhou等人提出了一种公平性导向的移动边缘缓存策略，通过缓存优化平衡不同用户的响应时间，减少资源分配不均带来的体验差异[16]。这一方法特别适用于资源受限环境，为用户提供了更具公平性的服务。

尽管现有研究在提高整体QoE和优化用户体验一致性方面取得了一定进展，但大多数方法仍面临着计算效率和迁移机制的挑战。此外，现有的研究多集中于单一层次的资源优化（如边缘计算或缓存策略），缺乏对云-边协同环境中微服务部署的整体性考量，未能全面解决不同用户QoE差异问题。在此背景下，如何在云-边协同架构中实现公平性优化，结合云端和边缘资源有效减少不同用户之间的QoE差异，仍是一个亟待解决的问题。因此，未来的研究需要进一步探索云-边协同环境中，如何通过高效的部署策略和公平性优化来全面提升微服务系统的用户体验公平性。

# 主要研究内容

本研究围绕在云-边协同环境中优化视频流媒体服务的公平性微服务部署。随着视频流媒体用户数量的增长，确保用户体验的一致性成为关键问题。现有的云边协同部署方法通常侧重于提升整体用户体验质量（QoE），但在不同用户之间可能存在显著的体验差异，如响应时间的差距。为此，本研究针对视频流服务的特性，提出了以公平性为导向的优化方法，以最小化不同用户之间的响应时间差异为核心目标，从而提升用户体验的公平性。具体研究内容如下：

## 云-边协同架构的系统建模

针对视频流媒体服务的需求，建立云-边协同环境的系统模型，包含用户分布、边缘节点和云节点的资源配置、服务实例部署等因素。明确系统的物理拓扑结构、传输延迟、带宽限制、计算资源配置等对用户响应时间的影响。

## 响应时间差异最小化的公平性目标设计

提出以最小化用户间响应时间差异为优化目标的设计思路，通过数学模型定义用户连接到边缘或云节点的响应时间，包含传输延迟和处理延迟。结合视频流特有的带宽要求，确保优化目标适应低延迟需求。

## 多约束优化建模

综合考虑视频流服务的成本控制和资源利用率，设计包括平均响应时间、部署成本、资源限制等多重约束条件，确保方案的经济可行性和资源高效利用，满足视频流服务在实际场景下的应用需求。

## 模型公平性优化效果的评估标准

设定评估模型公平性优化效果的指标，包括响应时间差异、平均响应时间、部署成本和资源消耗。通过这些指标来衡量模型在视频流服务中实现的公平性表现和实际应用效果，为实验验证提供客观标准。

# 研究方案

本研究将通过以下步骤实现云边协同环境中视频流服务的公平性优化部署方案：

## 需求分析与系统建模

首先收集视频流媒体服务的用户服务质量要求，明确公平性在用户体验中的关键作用。然后构建系统模型，模型中包含用户分布、边缘和云节点的资源配置、服务实例的部署情况等。定义相关模型参数，包括传输延迟、带宽、计算资源等，为优化模型的建立奠定基础。

## 目标函数和约束条件的设计和实现

根据公平性优化目标，设计以最小化用户响应时间差异为核心的目标函数。并结合视频流媒体服务的特性，设定多项约束条件：

* 平均响应时间：设定最大允许响应时间以确保整体用户体验；
* 资源限制：边缘和云节点的资源配置不超过设定最大值；
* 部署成本：边缘和云节点的部署成本控制在预算内；
* 用户唯一连接：每个用户仅连接到一个服务器节点，以便于资源分配和计算优化。

## 初始值设定与仿真数据生成

根据实际视频流媒体服务的特征数据，设定初始参数值，包括用户请求大小、节点物理距离、传输速度、带宽等，并生成仿真数据集。数据集应能模拟不同用户分布、服务请求和带宽条件，确保模型能够准确反映实际场景。初始值的设定参考常见视频流服务的需求，将其数据输入到模型中，模拟真实的用户请求行为。

## 模型求解与实现

为了为求解云边协同环境下的视频流服务公平性部署优化问题，本研究选用遗传算法（GA）来进行求解。遗传算法是一种启发式优化方法，通过模拟生物进化过程中的选择、交叉和变异操作，在多个候选解之间进行迭代优化，以获得接近最优的部署方案。具体求解过程如下：

1. 编码方案：使用遗传算法求解该问题时，将用户连接到服务节点的关系编码为个体的基因序列。每个基因序列中的元素代表一个用户连接到某一服务节点的选择。采用整数编码表示每个用户连接的节点编号。
2. 适应度函数：定义目标函数为适应度函数，以最小化用户响应时间差异的方差为主要优化目标。在计算适应度时，综合考虑了用户到节点的传输和处理延迟。适应度函数的设计确保了用户体验的公平性，响应时间差异越小，适应度越高。
3. 遗传算法参数设定：设定GA的主要参数，包括种群大小、交叉概率、变异概率、最大代数等。
4. 选择、交叉与变异操作：在每一代中，通过选择、交叉和变异操作生成新的候选解：

* 选择：采用锦标赛选择，优先保留适应度较高的个体。
* 交叉：使用合适的交叉方法生成新个体，确保基因序列在父代间的交换。
* 变异：随机改变个体的某些基因（用户连接的节点编号），以保证种群的多样性。

1. 结果提取：迭代至最大代数或收敛条件满足时停止。选择适应度最高的个体作为最终的优化方案，提取用户到服务节点的连接配置。
2. 结果验证：计算最终方案的用户响应时间差异方差，验证是否满足公平性要求和系统约束。

## 实验验证与性能分析

在模拟环境中运行模型，验证不同用户之间的响应时间差异是否缩小，分析部署方案的公平性效果。使用平均响应时间、资源消耗、部署成本等指标，评估系统的整体表现，并分析约束条件的满足情况。针对不同规模的用户负载、资源配置条件进行多次实验，以验证模型的扩展性和适用性。

# 进度安排，预期达到的目标

## 进度安排

|  |  |  |
| --- | --- | --- |
| 工作安排 | 周数 | 起止时间 |
| 需求分析与模型构建 | 3 | 2024.12.02-2024.12.22 |
| 算法设计与实现 | 6 | 2024.12.22-2025.02.02 |
| 实验验证与性能分析 | 6 | 2025.02.03-2025.03.16 |
| 优化完善 | 4 | 2025.03.17-2025.04.13 |
| 结题，论文编写 | 4 | 2025.04.14-2025.05.11 |

## 预期目标

中期目标：完成云边协同环境中视频流服务的公平性优化模型设计和遗传算法的基本 框架，进行初步实验与参数调优。

结题目标：实现云边协同环境下的公平性优化模型，能够基于视频流服务的实际需求， 在满足资源和成本限制的前提下，最大限度地减少用户间响应时间差异。

# 课题已具备和所需的条件、经费

使用实验室提供的测试环境。

# 研究过程中可能遇到的困难和问题，解决的措施

1. 数据真实性不足：仿真数据可能无法完全模拟实际应用场景的数据特征，影响实 验结果的可靠性。

解决措施：在仿真数据的基础上，可以通过公开数据集进行结果验证，提高模型 的适用性。

1. 遗传算法收敛速度较慢：遗传算法在复杂的约束条件下可能出现收敛慢的问题， 影响优化效率。

解决措施：在GA的基础上引入局部搜索算法，提升收敛速度。

# 主要参考文献

1. J. Thönes, "Microservices," in IEEE Software, vol. 32, no. 1, pp. 116-116, Jan.-Feb. 2015, doi: 10.1109/MS.2015.11.
2. Ashwin Rao, Arnaud Legout, Yeon-sup Lim, Don Towsley, Chadi Barakat, and Walid Dabbous. 2011. Network characteristics of video streaming traffic. In Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies (CoNEXT '11). Association for Computing Machinery, New York, NY, USA, Article 25, 1–12. https://doi.org/10.1145/2079296.2079321
3. Dragoni, N. *et al.* (2017). Microservices: Yesterday, Today, and Tomorrow. In: Mazzara, M., Meyer, B. (eds) Present and Ulterior Software Engineering. Springer, Cham. <https://doi.org/10.1007/978-3-319-67425-4_12>
4. M. N. O. Sadiku, S. M. Musa and O. D. Momoh, "Cloud Computing: Opportunities and Challenges," in IEEE Potentials, vol. 33, no. 1, pp. 34-36, Jan.-Feb. 2014, doi: 10.1109/MPOT.2013.2279684.
5. W. Zhang, Y. Hu, Y. Zhang, and D. Raychaudhuri, “SEGUE: Quality of service aware edge cloud service migration,” in Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci., 2016, pp. 344–351.
6. Q. He et al., “A game-theoretical approach for mitigating edge DDoS attack,” IEEE Trans. Dependable Secure Comput., vol. 19, no. 4, pp. 2333–2348, Jul./Aug. 2022.
7. K. Cao, Y. Liu, G. Meng and Q. Sun, "An Overview on Edge Computing Research," in IEEE Access, vol. 8, pp. 85714-85728, 2020, doi: 10.1109/ACCESS.2020.2991734.
8. J. Pan and J. McElhannon, "Future Edge Cloud and Edge Computing for Internet of Things Applications," in IEEE Internet of Things Journal, vol. 5, no. 1, pp. 439-449, Feb. 2018, doi: 10.1109/JIOT.2017.2767608.
9. X. He, H. Xu, X. Xu, Y. Chen and Z. Wang, "An Efficient Algorithm for Microservice Placement in Cloud-Edge Collaborative Computing Environment," in IEEE Transactions on Services Computing, vol. 17, no. 5, pp. 1983-1997, Sept.-Oct. 2024, doi: 10.1109/TSC.2024.3399650.
10. H. SHI, R. V. Prasad, E. Onur and I. G. M. M. Niemegeers, "Fairness in Wireless Networks:Issues, Measures and Challenges," in IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 5-24, First Quarter 2014, doi: 10.1109/SURV.2013.050113.00015.
11. P. Lai et al., “QoE-aware user allocation in edge computing systems with dynamic QoS,” Future Gener. Comput. Syst., vol. 112, pp. 684–694, 2020.
12. F. Liu, G. Tang, Y. Li, Z. Cai, X. Zhang and T. Zhou, "A Survey on Edge Computing Systems and Tools," in Proceedings of the IEEE, vol. 107, no. 8, pp. 1537-1562, Aug. 2019, doi: 10.1109/JPROC.2019.2920341.
13. R. Alsurdeh, R. N. Calheiros, K. M. Matawie and B. Javadi, "Hybrid Workflow Scheduling on Edge Cloud Computing Systems," in IEEE Access, vol. 9, pp. 134783-134799, 2021, doi: 10.1109/ACCESS.2021.3116716.
14. C. Zhang, J. Yin and S. Deng, "Ensuring Fairness in Edge Networks: A GNN-Based Media Workload Migration Scheme With Fairness Guarantee," in IEEE Transactions on Services Computing, vol. 17, no. 3, pp. 934-948, May-June 2024, doi: 10.1109/TSC.2023.3298695.
15. H. Hao, C. Xu, W. Zhang, S. Yang and G. -M. Muntean, "Computing Offloading With Fairness Guarantee: A Deep Reinforcement Learning Method," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 10, pp. 6117-6130, Oct. 2023, doi: 10.1109/TCSVT.2023.3255229.
16. J. Zhou, F. Chen, Q. He, X. Xia, R. Wang and Y. Xiang, "Data Caching Optimization With Fairness in Mobile Edge Computing," in IEEE Transactions on Services Computing, vol. 16, no. 3, pp. 1750-1762, 1 May-June 2023, doi: 10.1109/TSC.2022.3197881.