 DOI: https://doi.org/10.12345/bdai.v7i3.38175

Edge Computing Based Multi Region Cloud Video Streaming Scheduling and Load Balancing Mechanism

Fei Wang

Beijing Zhongchuang Video Technology Co., Ltd., Beijing, 100096, China

Abstract

With the deep penetration of cloud video and audio/video communication technology into multiple industries, the demand for cross regional deployment is increasing. However, the heterogeneity of resources, network fluctuations, and differentiated demands pose challenges to traditional centralized scheduling, such as low resource utilization, high latency, and poor system stability. This research proposes a dynamic resource scheduling and load balancing mechanism based on edge computing, which reduces the transmission delay through the nearby processing of edge nodes, and realizes dynamic resource allocation across regions with intelligent algorithms, aiming to optimize resource utilization and improve the real-time and stability of the system. Experimental verification shows that this mechanism can significantly reduce end-to-end delay and improve load balancing efficiency. Its innovation lies in the integration of edge computing and industry scenario requirements for differentiated design, which provides both technical and management support for the large-scale implementation of cloud video communication.

Keywords

edge computing; cloud video streaming; load balancing; resource scheduling; industry scenario adaptation

基于边缘计算的多地区云视频流调度与负载均衡机制

王飞

北京中创视讯科技有限公司，中国 · 北京 100000

摘　要

随着云视频与音视频通信技术向多行业深度渗透，其跨地区部署需求日益增长，但资源异构性、网络波动性及需求差异化导致传统中心化调度面临资源利用率低、延迟高、系统稳定性差等挑战。本研究提出基于边缘计算的动态资源调度与负载均衡机制，通过边缘节点就近处理降低传输延迟、结合智能算法实现跨地区资源动态分配，旨在优化资源利用率、提升系统实时性与稳定性。实验验证表明该机制可显著降低端到端延迟并提高负载均衡效率，其创新点在于融合边缘计算与行业场景需求进行差异化设计，为云视频通信规模化落地提供技术与管理双重支撑。

关键词

边缘计算；云视频流；负载均衡；资源调度；行业场景适配

1 引言

云视频与音视频通信技术正加速向教育、医疗、工业等多行业渗透，其跨地区规模化部署需求激增，但资源异构性导致计算能力分布不均、网络波动性引发传输质量不稳定、需求差异化要求系统灵活适配等核心矛盾日益凸显。边缘计算通过分布式部署与本地化处理可显著降低端到端延迟并提升带宽利用率，为解决上述问题提供技术支撑。然而跨地区资源调度与负载均衡的动态适配、技术优化目标与商业成本目标的协同平衡、行业场景差异对系统架构的约束仍是待突破的关键挑战。本研究提出基于边缘计算的动态调度与负载均衡机制，构建技术实现与商业落地双闭环管理体系，并通过典型行业场景实验验证其有效性，为云视频通信技术的规模化应用提供理论与方法支持。

【作者简介】王飞（1975–），男，中国山西大同人，硕士，从事云视频与音视频通信技术及相关企业管理研究。

2 云视频通信系统的多维度管理框架

云视频通信系统的规模化应用需构建覆盖技术实现与商业落地的多维度管理框架，以应对跨部门协作、成本控制及组织能力建设等核心挑战。

2.1 跨部门协同与项目管理闭环

跨部门协同与项目管理闭环聚焦全流程管控：需求管理通过客户节点优先级排序、行业监管合规性前置、硬件生命周期倒排等机制实现需求精准对齐；风险控制采用技术不确定性显式化建模与关键路径动态识别方法降低项目偏差；闭环验证则通过需求 - 研发 - 测试 - 运维 - 售后全链条衔接确保交付质量，例如将客户反馈直接映射至研发迭代环节形成持续优化循环。

2.2 成本治理与商业化适配

成本治理与商业化适配构建量化决策体系：资源消耗模型将带宽占用、转码算力、存储空间等成本要素分解为可计量单元，为动态优化提供数据基础；优化策略通过峰谷时段资源调度、边缘节点与中心云协同处理、智能降级（如分辨率自适应调整）等手段实现成本与体验平衡；商业模型匹配基于单位成本测算、毛利结构分析及规模效应预测，指导产品定价与市场拓展策略制定，例如通过边缘节点部署降低 30% 核心带宽成本的同时提升 20% 区域覆盖率。

2.3 组织能力建设与知识沉淀

组织能力建设与知识沉淀强化可持续交付：技术分层架构明确协议优化、媒体处理、云平台运维等岗位分工边界，避免职责重叠导致的效率损耗；经验转化机制通过故障案例库建设、实时指标看板监控、标准化培训体系等手段将隐性知识显性化，例如将典型网络抖动问题处理流程固化为可复用的操作指南；组织能力复制通过模块化交付流程设计、跨区域团队协同机制及质量门禁控制，确保从个人经验到团队标准化输出的稳定性，支撑系统在多行业场景中的快速复制与规模化落地。

3 基于边缘计算的调度与负载均衡机制设计

3.1 边缘计算赋能的资源调度模型

边缘计算通过分布式部署与本地化处理能力，构建了覆盖地理覆盖、网络拓扑与算力分布的联合优化模型。边缘节点部署策略基于多目标优化算法，结合区域用户密度、网络延迟阈值及算力冗余度，动态确定节点位置与数量，例如在人口密集区部署高算力节点以支撑实时转码需求。动态任务分配采用基于实时负载、网络质量及任务优先级的调度算法，通过实时采集边缘节点的 CPU 占用率、带宽剩余量及任务队列长度，结合任务类型（如音视频流处理、数据存储）的优先级权重，实现任务与节点的精准匹配。中心 - 边缘协同机制通过全局资源池与本地缓存的互补设计，中心云负责全局资源调度与历史数据存储，边缘节点处理实时性要求高的本地任务，并通过缓存热门内容减少跨区域传输，例如在视频会议场景中预加载常用背景素材以降低带宽占用。

3.2 多地区负载均衡的智能决策框架

负载均衡框架以量化评估指标为基础，构建了覆盖 CPU 利用率、带宽占用及请求延迟的多维度评估模型。通过实时采集各节点资源使用数据，结合历史负载趋势分析，生成动态负载指数，例如将 CPU 利用率超过 80%、带宽占用超过 90% 或请求延迟超过 200ms 的节点标记为高负载状态。均衡策略选择基于场景化适配原则，针对低并发场景采用轮询算法实现负载均摊，针对高并发场景采用加权分配算法，根据节点算力、带宽等资源参数分配任务权重，针对突发流量场景采用最小连接数算法优先调度空闲节点。自适应调整机制通过集成机器学习模型，基于历史负载数据预测未来负载变化趋势，动态调整调度参数（如任务分配权重、负载阈值），例如在电商直播场景中提前预判流量高峰并增加边缘节点资源预留，确保系统稳定性。

3.3 行业场景驱动的差异化设计

行业场景差异化需求驱动调度与负载均衡机制的定制化设计。教育场景中，高并发互动需求要求系统支持千人级实时音视频交互，通过边缘节点部署低延迟编码器与 QoS 保障策略，将端到端延迟控制在 150ms 以内，同时采用动态码率调整技术适应不同网络环境。医疗场景中，高清影像传输与合规性约束要求系统具备高可靠性，通过边缘节点部署加密模块与本地存储备份机制，确保数据传输安全，同时采用冗余链路设计避免单点故障。工业场景中，远程协作与工业现场适配要求系统支持毫秒级实时控制，通过边缘节点部署工业协议转换模块与时间敏感网络（TSN）技术，实现设备数据与音视频流的同步传输，例如在机器人远程操控场景中通过边缘计算降低控制指令传输延迟至 10ms 以内，满足工业自动化需求。

4 实验验证与结果分析

4.1 实验环境搭建

实验测试平台采用多地区分布式架构，涵盖北京、上海、广州等 6 个核心区域的边缘节点、阿里云 ECS 云服务器集群及 1000+ 模拟客户端。边缘节点部署于运营商骨干网边缘机房，单节点配置 8 核 CPU、32GB 内存及 10Gbps 带宽；云服务器采用通用型 c6 实例（16 核 64GB 内存）作为中心控制节点；模拟客户端通过 Docker 容器化技术生成，支持自定义视频流参数（分辨率 720P/1080P、码率 1-5Mbps）及并发请求量（100-1000 路 / 节点）。基准数据集覆盖教育、医疗、工业三大典型场景：教育场景采用高并发互动视频流（平均并发 500 路 / 节点、延迟敏感型）；医疗场景包含高清影像传输（单流 4K 分辨率、带宽占用 8Mbps）；工业场景模拟设备监控数据与音视频流混合传输（时延要求 ＜ 50ms）。所有数据通过真实业务系统抓取并脱敏处理，确保实验环境与实际部署高度一致。

4.2 性能对比实验

调度效率实验表明，基于边缘计算的动态调度机制将任务分配延迟从传统中心化调度的 120ms 降至 35ms，资源利用率提升 28.6%。通过对比中心调度与边缘调度在 1000 路并发请求下的 CPU 占用率，边缘节点平均负载降低至 62%（中心节点达 89%），证明分布式调度可有效分散计算压力。负载均衡效果验证显示，跨地区节点负载差异标准差从 0.32 降至 0.09，系统稳定性指标如请求成功率提升至 99.97%。在教育场景高并发测试中，边缘调度机制保障了 98.7% 的会话延迟低于 150ms；医疗场景下 4K 影像传输的卡顿率从 12% 降至 1.5%；工业场景中设备控制指令传输时延稳定在 8ms 以内，满足实时性要求。行业适配验证结果表明，所提机制在三大场景的关键性能指标延迟、吞吐量、丢包率均优于行业基准值 15% 以上。

4.3 成本与商业化分析

资源消耗成本优化显著：带宽成本通过边缘节点本地处理减少 31.2%（中心云带宽需求下降 42%）；存储成本因数据分级存储策略（热数据边缘缓存、冷数据中心归档）降低 27.5%；算力成本通过动态资源调度实现按需分配，单位任务算力消耗下降 19.8%。商业模型验证显示，在年服务 10 万用户规模下，单位用户成本从 12.6 元 / 月降至 8.3 元 / 月，毛利结构从 35% 提升至 52%。规模效应分析表明，当用户规模超过 50 万时，边际成本下降趋势趋缓，但通过边缘节点复用（单节点支持多行业服务）可进一步摊薄成本。敏感性测试证明，在带宽价格波动 ±20% 或算力成本上升 15% 的场景下，系统仍能保持 40% 以上的毛利率，验证了商业模型的可持续性。实验数据表明，所提机制在技术性能与商业价值间实现有效平衡，为云视频通信技术的规模化落地提供可复制的解决方案。

5 结论与展望

5.1 研究成果总结

本研究提出基于边缘计算的调度与负载均衡机制，通过地理覆盖、网络拓扑与算力分布的联合优化模型实现边缘节点动态部署，结合实时负载、网络质量及任务优先级的调度算法提升资源分配效率。构建技术 - 商业双闭环管理体系，技术层实现需求 - 研发 - 测试 - 运维全链条衔接，商业层通过资源消耗模型与动态优化策略降低单位成本。实验验证表明，该机制在教育、医疗、工业场景中显著降低任务分配延迟（降幅达 71%）、提升资源利用率（提升 28.6%），并实现跨地区负载差异标准差从 0.32 降至 0.09 的系统稳定性优化，证明其技术可行性与商业价值。

5.2 实践启示

技术落地需强化企业管理能力与组织协同：跨部门协作机制是保障需求精准对齐与风险可控的关键，例如通过故障复盘标准化流程将隐性知识转化为可复用的经验库；行业理解深度决定场景适配效果，医疗场景需优先满足合规性约束（如数据加密与本地存储），工业场景则需聚焦毫秒级实时控制需求，差异化竞争策略可提升市场渗透率。此外，商业模型设计需平衡规模效应与成本弹性，通过边缘节点复用与动态资源调度实现边际成本递减，为技术规模化应用提供可持续的盈利路径。

5.3 未来研究方向

边缘计算与 AI 的深度融合将成为核心优化方向：通过强化学习算法实现调度策略的自适应进化，结合数字孪生技术构建动态网络环境模拟平台，提升决策鲁棒性；探索跨行业场景的通用化调度框架，设计可配置的任务优先级权重模型与资源分配规则库，降低多场景适配成本；针对 5G/6G 网络波动特性，研究基于时延预测的预调度机制与多路径传输冗余设计，增强系统在动态网络环境下的抗干扰能力。长期来看，边缘计算与区块链技术的结合可实现分布式资源可信调度，为云视频通信生态构建提供新的技术范式。

参考文献

[1] 黄景朝,吴鹏,吴国辉,等.政务信息系统智能运维中的边缘计算技术应用[J].通讯世界,2026,33(02):141-144.

[2] 朱锦辉.基于嵌入式边缘设备的端-边-云视频流分析系统构建[J].舰船电子工程,2022,42(07):110-115.

[3] 庄瑞,程伟强,王瑞雪,等.一种基于报文容器的智算中心网络负载均衡方案[J/OL].南京邮电大学学报(自然科学版),1-10[2026-04-03].

[4] 朱希康.边缘计算环境下的资源优化关键技术研究[D].北京邮电大学,2025.

[5] 张冰雪.边缘计算中协作网络优化与计算卸载调度研究[D].北京科技大学,2025.