

边缘计算环境下智能管理任务的实时性优化策略探究

钱子阳

(广东海洋大学管理学院, 广东 湛江 524088)

摘要: 伴随着物联网感知设备、智慧管理终端大规模普及, 海量感知数据与智能管控任务逐步下沉至边缘计算节点。网络带宽随机波动、链路通信时延、边缘节点算力异构、数据传输阻塞等问题, 破坏感知-运算-反馈的数据闭环, 造成各类智能管理任务响应超时, 制约智慧安防、设施管控、环境监测、园区运维等管理业务落地。本文首先剖析带宽波动、传输延迟对于智能管理任务数据闭环时效性的多层结构性影响; 从任务优先级动态管控、分层式任务卸载部署、深度学习推理模型轻量化压缩、多节点协同缓存-数据预取四个维度, 系统性探究实时性优化方案。研究表明, 组合使用多维度调度优化手段能够削减排队时延、传输时延以及模型推理耗时, 改善边缘节点资源利用率; 相关策略可为智慧城市、海洋监测、海上平台管控等管理类边缘智能业务提供技术参考, 保障管控指令的低延迟下发。

关键词: 边缘计算; 智能管理任务; 时延优化; 任务卸载; 协同缓存; 模型轻量化

中图分类号: TP393.07 **文献标识码:** A **文章编号:** 3134-9986(2026)04-0013-03

一、网络带宽波动与通信延迟对数据闭环时效性的结构性影响

边缘计算架构之下整套智能管理业务形成“终端感知-数据上传-边缘运算-管控决策-指令下发-设备执行”的完整数据闭环。和传统云计算模式相比, 边缘服务器部署于距离感知终端较近的网络侧, 原生具备降低传输时延的优势, 适合处理园区设备管控、近海环境监测、公共设施巡检、安防预警等时间敏感型的管理任务。但现实场景当中无线网络环境复杂多变, 带宽扰动、链路阻塞、信号衰减、终端大规模并发接入等因素持续干扰闭环传输效率, 从传输层、计算层、调度层产生多层结构性时延影响。

首先是上行感知数据传输时延扰动。绝大多数智能管理终端依靠无线局域网、蜂窝网络、低-power广域网完成传感数据上报。环境遮挡、信道干扰、多终端抢占带宽会造成网络带宽动态起伏。当监测节点大批量同步上传视频、传感器时序数据、图像信息, 上行链路容易出现拥塞, 感知数据包排队等待时间大幅拉长, 原始监测数据抵达边缘节点时刻发生滞后。对于应急预警类智能管理任务, 传感信息滞后会直接降低环境异常识别速度, 耽误应急处置时机。

其次为下行管控指令的交付延迟。智能管理任务运算结束以后, 边缘节点需要下发启停、告警、参数调节、设备闭锁等控制指令。带宽不稳定极易引发指令数据包丢包、重传, 管控命令到达受控设备时间延后。部分闭环管控业务拥有严苛硬截止时间, 一旦指令超出

时延阈值, 整套管理闭环将会失效, 严重时引发设施故障、海洋监测设备失控、安防预警失灵等风险。

再次, 端-边-云多层架构之下链式时延叠加。复杂的综合性管理任务无法单纯依靠单台边缘节点完成运算, 需要终端预处理、边缘初步解析、云端深度研判之间多次交互。带宽波动使得每一轮数据交互均附带不确定的通信耗时, 多次数据往返造成时延累积。并且边缘节点算力资源有限, 当大量异构智能管理任务同时抵达服务器, 任务排队、资源抢占、内存拥堵催生排队时延, 通信延迟与计算时延互相耦合, 进一步放大闭环时延波动幅度。

最后是数据流量突发性引发的周期性阻塞。在突发事件触发之下, 例如海域灾害预警、园区警报触发, 大量传感设备瞬间激活并且上报监测数据, 形成突发性流量冲击。常规静态带宽分配模式难以适配流量峰值, 网络吞吐能力到达上限之后数据包长时间积压。不同于普通互联网业务, 智能管理任务带有较强的公共管控属性, 时延抖动、响应时间不可控相比平均时延过高更容易造成管理业务出错。

综合上述结构性问题, 实时性优化不能仅依靠硬件带宽扩容, 需要从任务调度、计算卸载、AI推理优化、缓存预取等软件层面搭建完整的时延管控体系。

二、实时性优化的策略

(一) 任务分类与优先级动态管理

边缘侧同时运行种类繁杂的智能管理任务, 各类任务的时延阈值、数据体量、安全等级、运算开销差

异悬殊。按照业务属性可以划分应急预警类、常规巡检类、数据统计类、后台运维管理任务。应急告警、危险设备关停属于硬实时任务,拥有严格最晚响应时限;日常环境数据采集、设备状态巡检属于软实时任务,可以承受小幅时延波动;报表统计、历史数据归档属于非实时任务,时延约束宽松。

首先搭建多等级优先级划分机制,基于截止时间、任务紧急程度、业务权重进行静态分级;引入动态优先级调度算法,依据边缘服务器当前负载、网络带宽状态、剩余排队时长实时调整任务优先权。当监测到突发事件流量冲击时,自动抬高预警管控任务优先级,将归档、数据分析类后置处理。

其次,执行任务抢占式调度机制,高优先级的应急管理任务抵达边缘节点之后,可以抢占低优先级任务的CPU、内存资源,保障管控指令优先运算。

最后,设置任务超时监测队列,定时筛查已经逼近时延阈值的任务;对于已经出现严重阻塞的非实时任务,可以执行任务休眠或者迁移至空闲次级边缘节点,防止低价值任务挤占紧急管控业务的计算资源。借助动态优先级管控实现边缘节点算力资源向时间敏感型智能管理任务倾斜,从任务调度源头降低排队时延。

(二) 计算任务的动态卸载与分层部署

端-边-云三层分层部署、自适应任务卸载是削减传输耗时、均衡节点负载的核心手段。智能管理任务可以拆解为终端预处理、边缘实时运算、云端深度分析三大层级,按照任务计算量、时延门槛、带宽质量动态确定任务执行位置。

首先,终端本地轻量化预处理。传感终端率先完成数据降噪、无效数据过滤、图像裁剪、特征提取,舍弃冗余监测数据之后再向上传输,减小数据包大小,降低无线网络的数据传输压力。简单开关判断、阈值比对类极简管理任务直接在本地终端完成运算,不需要上传边缘服务器。

其次,边缘节点之间的动态卸载调度。当主边缘服务器CPU占用率、内存负载到达阈值,系统实时侦测周边空闲的相邻边缘节点,将普通巡检、数据分析任务卸载至负载较低的节点;应急管控任务依旧留存本地处理,规避跨节点传输带来的额外通信时延。可采用强化学习算法实时感知带宽波动、节点负载,输出最优卸载决策,实现负载均衡。

最后,端-边-云的任务分层分工。所有具备硬实时约束的智能管控业务全部交由边缘计算层处理;复杂长期趋势研判、大规模历史数据集训练、多区域综合管理任务上传远端云计算中心。云端只负责非实时

深度业务,管控闭环不再依赖远距离云端交互,由此规避广域网传输带来的不稳定延迟。

针对海洋平台、野外监测点这类网络条件较差的偏远管理区域,还可以采用就近边缘下沉部署,缩短信号传输距离,削弱带宽波动造成的负面影响。

(三) 推理模型的轻量化压缩与自适应精度调节策略

当下绝大多数智能管理任务依托深度学习模型完成图像识别、异常检测、环境参数预判,神经网络推理耗时是边缘节点计算时延的主要组成部分。边缘服务器算力有限,原版大型神经网络推理速度较慢,需要开展模型轻量化改造以及动态精度调控。

首先,执行多途径模型压缩工作。依次进行模型剪枝、参数量化、知识蒸馏、网络稀疏化处理。剪枝操作去除神经网络当中权重较低的冗余神经元;量化将32位浮点参数压缩至16位或者8位整型参数,减少内存占用、加快矩阵运算;借助知识蒸馏,由大型教师模型训练小型轻量化学生模型,在损耗少量识别准确率的前提下大幅缩短推理时长。面向安防识别、水质监测、设备故障检测等不同管理业务制作专属轻量化推理模型。

其次,搭建自适应精度调节机制。系统实时读取当前网络时延、服务器负载、任务紧急等级。当应急管理任务到来,自动启用低精度快速推理模式,优先保证响应速度;网络通畅、节点负载宽松的时候切换至高精度模式,保障异常识别准确率。针对多阶段推理任务,可以采取分层推理,基础特征检测在边缘执行,精细识别任务按需上传云端。

最后,实行模型分片部署。将大型检测模型拆分成多个子模块,不同模块部署在相邻多个边缘节点之上,实现分布式并行推理,分摊单一边缘节点的运算压力,缩短模型推理耗时。在时延指标和识别精度之间建立动态平衡,适配多变的边缘运行环境。

(四) 边缘节点间的协同缓存与数据预取机制

频繁重复传输监测特征数据、管控模型、设备参数文件会消耗大量带宽资源,加剧网络拥堵。搭建多边缘节点协同缓存与数据预取机制,可以减少重复的数据上传行为,优化闭环传输效率。

首先,边缘节点协同缓存架构。同一管控区域内所有边缘服务器组建缓存集群,存储高频访问的设备配置参数、轻量化推理模型、历史监测基准数据、常规管控脚本。当终端需要调取公共资源时优先读取本地缓存;本地缓存缺失之后向邻近边缘节点调取资源,不用每次从远端云端下载文件,节约广域网络传输开销。设置缓存淘汰策略,淘汰长时间没有调用的

数据, 优先保存智能管理业务高频资源。

其次, 基于业务规律的数据预取策略。智能管理任务存在显著周期性, 例如全天候定时环境采样、早晚高峰安防巡检。边缘控制器依托历史任务运行规律, 预判下一阶段终端需要的模型文件、参考数据集, 在网络带宽空闲时段提前完成资源预下载、本地缓存。等到管控任务触发时直接读取本地缓存资源, 避开带宽高峰期的数据传输阻塞。

最后, 管控指令缓存与热点任务分流。对于周期性重复下发的设备调控指令, 在对应区域边缘节点缓存指令模板; 当触发条件达成之后立刻下发指令, 省去云端调度流程。针对人流量大、设备数量较多的管控片区, 利用协同缓存分担主节点的数据访问压力, 规避缓存访问拥堵。依靠缓存-预取机制降低重复传输频次, 缓和带宽波动对于整套管理闭环的冲击。

三、结束语

在边缘计算环境之下开展各类智能管控业务, 数据闭环全过程容易受到无线网络带宽扰动、数据包重传、节点负载失衡、深度学习推理耗时过长等要素制约, 引发管控任务响应延迟。本文针对时延的结构性成因, 从任务优先级动态管控、分层式任务卸载部署、AI推理模型轻量化压缩、多节点协同缓存和数据预取四项路径, 构建完整的实时性优化框架。在实际的智慧管理系统搭建过程当中, 各项优化策略之间可以协同运作: 依靠优先级调度保障紧急管控任务优先获取

算力资源; 分层任务卸载均衡各个边缘节点负载; 轻量化推理模型削减计算开销; 缓存预取机制降低重复数据传输压力。后续优化工作还可以结合海洋环境、园区设施、城市公共管控等不同业务场景, 借助人工智能调度算法动态适配网络环境变化, 持续降低智能管理闭环的时延抖动, 增强边缘管控系统的稳定性、即时性, 助力各类线下设施的智慧化管理建设。

参考文献:

- [1] 谢人强, 陈永红. 边缘计算原理、架构与关键技术 [M]. 北京: 电子工业出版社, 2023.
- [2] 毛玉欣, 赵弘. 移动边缘计算环境下任务卸载优化综述 [J]. 通信学报, 2022, 43 (06): 186-204.
- [3] 刘军, 张冬松, 吴彤. 边缘智能中深度学习推理加速技术研究进展 [J]. 计算机研究与发展, 2023, 60 (08): 1721-1746.
- [4] 杨守义, 韩浩锦. 面向边缘计算的协同缓存和任务迁移优化算法 [J]. 电子与信息学报, 2024, 46 (12): 3891-3899.
- [5] 董裕民, 何志敏. 云-边-端架构下工业智能任务时延优化策略 [J]. 计算机应用研究, 2025, 42 (02): 567-572.
- [6] 王浩, 许子豪. 基于动态优先级的物联网边缘任务调度算法 [J]. 传感器与微系统, 2024, 43 (05): 154-157.
- [7] 李慧, 李秀华. 移动边缘计算网络下深度学习模型分片推理优化 [J]. 计算机工程, 2025, 51 (03): 132-139.
- [8] 陈兴, 黎明. 细粒度功能拆分的边缘计算任务卸载方案 [J]. 小型微型计算机系统, 2024, 45 (07): 1480-1487.
- [9] 张宇真, 赵旭辉. 基于深度强化学习的工业互联网云边资源调度 [J]. 集群计算, 2025, 28 (04): 813-827.
- [10] 周浩. 面向智慧园区管控系统的边缘计算任务实时调度研究 [D]. 广州: 广东工业大学, 2024.

Research on Real-time Optimization Strategy of Intelligent Management Tasks under Edge-Computing Environment

Qian Ziyang

(School of Management, Guangdong Ocean University, Zhanjiang, Guangdong 524088)

Abstract: With the popularization of internet-of-things sensors and intelligent management terminals, massive sensing data and intelligent control tasks are gradually delivered to edge-computing nodes. Random bandwidth fluctuation, link latency, heterogeneous computing capacity of edge nodes and data-transmission congestion break the sensing-computation-feedback data closed-loop, leading to response timeout of multiple intelligent-management tasks and hindering the deployment of smart security, equipment supervision, environmental monitoring and park-operation services. This paper analyses the structural influence of bandwidth fluctuation and communication delay on the timeliness of data closed-loop. Four major optimization approaches are explored, including dynamic priority-based task management, layered deployment and computation offloading, lightweight compression and self-adaptive precision adjustment of inference models, multi-node cooperative caching and data prefetching. The combined multi-dimensional scheduling scheme can cut down queuing delay, transmission overhead and model-inference latency while improving resource-utilization ratio of edge servers. These strategies provide technical support for edge-intelligent management systems such as smart-city facilities and marine-environment monitoring, and guarantee low-latency delivery of control instructions.

Keywords: edge computing; intelligent-management task; latency optimization; task offloading; cooperative cache; model lightweight