

时延QoS要求约束的业务流量卸载算法

于宝珠, 焦子阳*

沈阳理工大学信息科学与工程学院, 辽宁 沈阳

收稿日期: 2025年2月10日; 录用日期: 2025年3月10日; 发布日期: 2025年3月17日

摘要

随着网络技术的不断发展, 通信网络中的业务流量和终端数量急剧增加, 导致网络传输的复杂性和负载压力显著上升。业务流量卸载通过将部分业务从主网络卸载至辅助网络或边缘计算节点, 以缓解主网络的压力, 提高整体网络的服务质量。本文研究了在时延QoS要求约束下, 业务流量的卸载算法。结合有效容量和有效带宽理论, 确保在高负载条件下满足业务的统计型时延QoS要求。通过排队论、有效容量和有效带宽理论构建时延分析模型, 设计了流量卸载算法。基于不同网络的服务速率, 求解相应的时延违反概率, 在此基础上, 计算不同网络的可支持的最大到达率, 对业务流量进行权重分配, 从而实现了对业务流量的动态卸载, 确保了各网络资源的优化利用和系统性能的提升。仿真结果表明, 在时延QoS约束条件下, 算法能够有效降低网络时延, 提高网络稳定性和服务质量。

关键词

排队论, 有效容量理论, 有效带宽理论, 业务流量卸载, 时延QoS

The Traffic Offloading Algorithm Design under the Delay QoS Constraints

Baozhu Yu, Ziyang Jiao*

College of Information Science and Engineering, Shenyang Ligong University, Shenyang Liaoning

Received: Feb. 10th, 2025; accepted: Mar. 10th, 2025; published: Mar. 17th, 2025

Abstract

With the continuous development of network technology, traffic and the number of terminals in the communication network has increased sharply, resulting in a significant increase in the complexity and load pressure of network transmission. Unload some services from the primary network to the secondary network or edge computing nodes to relieve the pressure on the primary network and

*通讯作者。

文章引用: 于宝珠, 焦子阳. 时延 QoS 要求约束的业务流量卸载算法[J]. 计算机科学与应用, 2025, 15(3): 72-82.
DOI: 10.12677/csa.2025.153059

improve the overall network service quality. This paper studies the traffic offloading algorithm under the constraint of delay QoS requirements. Combining the theory of effective capacity and effective bandwidth, the QoS requirements of statistical delay can be satisfied under high load conditions. Based on queuing theory, effective capacity theory and effective bandwidth theory, the delay analysis model is constructed, and the traffic offloading algorithm is designed. Based on the service rate of different networks, the corresponding delay violation probability is solved. On this basis, the maximum arrival rate supported by different networks is calculated, and the weight of service traffic is assigned. In this way, the dynamic unloading of service traffic is realized, and the optimal utilization of network resources and the improvement of system performance are ensured. Simulation results show that the algorithm can effectively reduce network delay and improve network stability and service quality under the delay QoS constraint.

Keywords

Queuing Theory, Effective Capacity Theory, Effective Bandwidth Theory, Traffic Offloading, Delay QoS

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

在当今的数字化时代, 移动互联网和物联网技术的蓬勃发展极大地推动了数据流量的爆炸式增长。这种增长不仅体现在数量上, 更体现在数据流的多样性和复杂性上。从简单的文本信息到高清视频通话, 不同类型的数据流对网络的要求各不相同, 尤其是在时延(延迟)和服务质量(QoS)方面。随着用户对网络服务质量要求的提高, 在保证时延 QoS 要求的前提下实现业务流量卸载, 是网络优化的重要课题。

业务流量卸载, 是将业务负载从相对拥塞的网络转移到相对空闲的网络(如 Wi-Fi), 是解决网络拥塞、提高数据传输效率的有效手段。然而, 不同的业务对时延的敏感度不同, 如实时视频通话对时延的要求远高于普通的数据下载。因此, 设计一个能够考虑时延 QoS 要求的业务流量卸载算法, 对于满足用户体验和提高网络效率具有重要意义。

本课题将聚焦于研究时延 QoS 要求约束的业务流量卸载算法。首先, 深入理解不同业务流量的特性及其对时延的敏感度, 以及这些因素如何影响业务流量卸载的决策过程。此外, 缺少精确的时延违反概率理论结果以指导流量卸载是算法设计中的一个关键挑战。本课题旨在提出一种新的卸载策略, 即保证时延 QoS 要求下, 合理分配不同网络间流量负载, 从而有效缓解网络拥塞, 提高数据传输效率。这对于提升用户的网络使用体验具有深远意义。

在 QoS 约束的业务流量卸载算法研究领域, 国内外学者们已经取得了一些重要成果。QoS 保障技术旨在通过合理的资源分配和调度, 确保各种应用和服务在网络中的性能需求得到满足。这些技术在多媒体通信、实时数据传输以及高可靠性应用中尤为重要。国内外学者对 QoS 保障技术进行了广泛研究。例如, 高建勋, 张永红等研究了基于服务质量的网络资源管理策略, 提出了一种动态资源分配方法来提高网络的服务质量[1]。Tolba B 等探讨了在异构网络环境下的 QoS 保障机制[2], 提出了一种多层次 QoS 模型, 能够在不同的网络环境中有效地保证服务质量。2003 年卡耐基梅隆大学的 D. Wu 等人提出了有效带宽理论的对偶形式即有效容量理论, 用来研究无线通信网络中的 QoS 保障问题[3]。

有效容量是指在给定的统计 QoS 要求下, 系统可以支持的最大到达速率。基于有效带宽/有效容量理

论, 美国德州农工大学 X. Zhang 和西安电子科技大学的程文驰等人, 针对 5G 异构网络提出一种新的统计时延 QoS 保障机制[4] [5]。文献[6]对时变衰落信道下的有效容量进行了研究。

文献[7]计算了实时媒体流在缓存区中的时延概率分布函数, 并给出了时延约束下实时媒体流的等效带宽的表达式。石文孝等学者在通信网理论与应用领域进行了深入研究, 为排队论的概念与规则提供了重要的阐述和应用[8]。此外, Chang C S 等学者对高速数字网络中的有效带宽进行了研究, 为理解网络性能提供了重要参考[9]。有效带宽和有效容量的定义和应用也受到了学者们的广泛关注。例如, 陈有汉等学者讨论了网络状态的不确定信息产生的原因, 分析并总结了目前考虑不确定信息的 QoS 单播和 QoS 多播路由算法[10]。这些研究为理解有效带宽在网络性能分析中的重要性提供了重要支持。

在卸载策略方面, 学者们也做出了一系列有益的研究。Xuefen Chi 等学者使用耦合带宽资源分配模型表明了 CBRA 在节省带宽和保障服务质量方面的有效性[11]。在路由策略方面, 张品等学者探讨了非确定环境下的 QoS 路由问题, 包括最大可能性路径、最优路径分解以及最优分解路径问题[12]。在国内, 余翔等学者研究了在不同网络环境下的业务流量卸载问题, 提出了一系列基于边缘计算的流量卸载方法, 有效地提高了网络的吞吐量和用户体验[13]-[15]。文献[16]以时延 QoS 为基础, 研究了信道的优化从而实现有效容量达到最大。文献[17]研究了功率分配优化的问题。

此外, 文献[18]构建了一种用于异构蜂窝网络中流量卸载的排队理论方法, Anand G, Faizan Q, Sang DD 则基于排序的网络选择在 5G 实现流量卸载[19]。文献[20]以最大化用户吞吐量为目标, 考虑用户的最小速率需求, 使用非合作博弈方法将异构蜂窝网络的功率控制问题分解为各基站各自的功率控制问题。文献[21]使用非合作博弈方法解决功率控制问题。文献[22]提出了一种基于微基站所处位置的功率控制方法。文献[23]利用微基站之间的相对位置信息, 动态的控制每个微基站的功率, 从而降低基站间的干扰, 提升网络的吞吐量。但文献[21]-[23]均未考虑用户的 QoS 需求。文献[24]提出了一种基于强化学习的流量卸载方案, 并基于这一方案提出了一种随机竞争的流量卸载算法。文献[25]以最大化异构蜂窝网络中所有基站群总容量为目标, 提出了一种基于强化学习的功率控制算法。

总而言之, 在利用传统方法解决流量卸载问题时, 诸多文献忽略了用户的 QoS 需求, 或者仅关注了用户的最小速率要求。

2. 时延 QoS 要求约束的流量卸载算法

本文研究了一种时延 QoS 约束的流量卸载算法, 通过 MATLAB 软件建模分析该算法的时延性能。本文基于泊松到达和常速率服务作为到达模型与服务机制, 通过有效容量理论描述无线信道在满足特定 QoS 要求下的最大可支持传输速率, 利用有效带宽理论衡量流量对网络资源的占用情况, 并通过大偏差理论分析网络在极端条件下的时延分布。

2.1. 到达过程的有效带宽与常速率服务建模

2.1.1. 到达过程有效带宽

设定随机过程和速率函数: 将业务流的到达过程建模为泊松过程, 平均到达率是 λ , 根据大偏差理论, 其速率函数 $\Lambda(\theta)$ 可以表示为:

$$\Lambda(\theta) = \lambda(e^{\theta} - 1) \quad (2.1)$$

根据到达过程有效带宽的定义, 泊松过程的有效带宽表示为:

$$E_B(\theta) = \frac{\lambda(e^{\theta} - 1)}{\theta} \quad (2.2)$$

其中 θ 为引入的变量。泊松过程的平均到达率是 λ ，为了防止系统过载、保持稳定性以及避免资源浪费，通常需要设定系统的处理能力(服务速率)要高于平均到达率。这是因为实际系统中流量可能存在突发性和波动。服务速率高于到达速率可以避免终端缓存数据的持续累积，而对于具有时延 QoS 要求的业务流，其所需的服务速率往往远超平均到达。

2.1.2. 常速率服务机制建模

常速率服务机制(Constant Rate Service Mechanism, CRSM)是一种在网络中提供恒定速率服务的机制，通常用于传输对时延敏感的业务流量。常速率服务机制的核心思想是在整个服务过程中，以固定的速率为业务流提供带宽保障，从而确保业务流量的稳定传输。

在数学建模中，常速率服务机制可以表示为一个恒定速率的服务过程。在这种机制下，业务流以固定速率 C 传输数据包，每个业务流的传输独立且互不干扰。根据有效容量理论，常速率服务的有效容量可以推导为： $E_C(\theta) = C$ 。

2.1.3. 队列溢出概率与时延违反概率

根据有效容量和有效带宽理论，我们可以得到队列溢出概率不等式：

$$\Pr(Q \geq \sigma) \leq e^{-\theta^* \sigma} \quad (2.3)$$

其中 Q 表示缓存中的数据包排队长度； σ 为队列阈值。队列阈值 σ 较小时，此式得出的队列溢出概率通常会高于实际值，因此此式可以视作实际队列溢出概率值的上界。 θ^* 与到达过程的有效带宽和服务过程的有效容量有关，定义为：

$$\theta^* = \{\theta > 0 : E_B(\theta) = E_C(\theta)\} \quad (2.4)$$

即为有效带宽等于有效容量情况下对应的 θ 。

时延违反概率是指系统中任务的等待时间(时延)超过某个预定阈值的概率。通常由 Little 公式和队列溢出概率推导而来，用于评估在高负载或突发情况下系统响应时间的可靠性。通过这种不等式，可以量化时延超出可接受范围的风险，从而帮助设计和优化系统以满足服务质量要求。

结合 little 公式，在到达过程为泊松过程的情况下，队列阈值与时延阈值的关系可以表示为： $\sigma = \lambda D_{\max}$ 。结合式(2.3)队列溢出概率不等式，可推导得到时延违反概率不等式为：

$$\Pr\{W \geq D_{\max}\} \leq \Pr\{Q \neq 0\} e^{-\theta^* \lambda D_{\max}} = \frac{\mu_A}{\mu_C} e^{-\theta^* \lambda D_{\max}} \quad (2.5)$$

其中 W 为系统时延， $e^{-\theta^* \lambda D_{\max}}$ 为违反概率的理论值， $D_{\max}$ 作为系统时延的阈值。

3. 流量卸载算法设计

3.1. 计算可支持的最大到达

二分法(Binary Search)是一种适用于有序数列的算法。原理是通过把搜索区间循环一分为二，不断减小区间范围来确定目标值。具体步骤如下：首先确定目标值的位置位于搜索区间的中间，再比较目标值与中间值。若目标值小于中间值，就在左半区间继续搜索；反正，在右半区间继续搜索。经过数次循环，能够找到目标值或区间缩小到无法再分。这种方法的时间复杂度为 $O(\log n)$ ，远优于线性搜索的 $O(n)$ 。

为了估计节点可支持的最大到达，采用常用的数值方法——二分搜索法。通过该方法逐步缩小搜索区间，逼近目标值。以下是详细的步骤说明：

- 1) 设定时延阈值 $D_{\max}$ ，时延违反概率要求 $W_{\max}$ ，容许误差概率 E ，两个接入节点的平均服务速率 C_1 和 C_2 。

2) 设定初始搜索区间: 首先, 需要根据实际情况对可支持最大到达的范围做出一个初步估计, 设定搜索的上下界。记下界为 low , 上界为 upper 。由于本文利用二分法搜索的是每个排队系统可支持的最大到达速率, 每个排队系统的平均服务速率均要大于平均到达, 因此下界 low 可设定为 0, 上界为该系统的平均服务速率 C_1 或 C_2 。

3) 计算搜索区间的中间值: 当 low 不大于 upper 时, 在每一步迭代中, 计算当前搜索区间的中间值 Mid , 公式如下:

$$\text{Mid} = \frac{\text{low} + \text{upper}}{2} \quad (3.1)$$

设定一个中间变量 R , 令 $R = \text{Mid}$, 作为本次迭代到达过程(泊松过程)的平均到达, 由式(2.4)求出 $E_B = E_C$ 的解 θ^* 。

4) 计算时延违反概率: 结合式(2.5)可得时延违反概率的表达式。将求得的 θ^* , 中间值 R , 时延阈值 $D_{\max}$ 代入表达式, 计算当前 R 下的时延违反概率 P 。

$$P = e^{-\theta^* D_{\max} R} \quad (3.2)$$

5) 比较容忍误差概率: 将计算得到时延违反概率 P 减去时延违反概率要求 $W_{\max}$ 并与误差阈值 E 进行比较, 即:

$$\left(e^{-\theta^* D_{\max} R} - W_{\max} \right) < E \quad (3.3)$$

如果等式成立就说明满足要求。 R 的值就等于当前的 Mid 值, 跳转到步骤(7)。

6) 更新上下界: 将求出的时延违反概率 P 与时延违反概率阈值 $D_{\max}$ 进行比较。若 $P < D_{\max}$, 则将 low 更新为 Mid , 反之 upper 更新为 Mid 。跳转步骤(3)重复循环。

7) 输出结果, R 值即为当前节点可支持的最大到达。

通过上述步骤, 我们能够有效地利用二分搜索法, 确定在给定时延违反概率阈值下的可支持最大到达。这种方法不仅计算效率高, 而且能够保证所求解的结果满足系统的服务质量要求, 从而在实际应用中提供可靠的性能保证。

3.2. 分配因子

负载的合理分配对于确保系统的服务质量(QoS)至关重要。本文通过对两种不负载服务速率的通信系统的流量卸载问题进行研究, 提出了一种利用分配因子的流量卸载方法。该方法通过二分搜索法求解不同节点的可支持最大到达, 并利用分配因子将总负载合理分配到各个系统中。

由 3.1 中介绍的系统二分法, 可以确定每个接入节点对应的排队系统可支持的最大到达速率。对于具有多个节点的排队系统, 可以将总平均到达在不同节点间进行分配, 以缓解时延 QoS 保障的压力。在本文中, 提出一种业务流量分配方案, 根据二分法搜索出的每个系统可支持的最大到达, 定义不同节点间的分配因子 W_1 和 W_2 , 若到达过程的总平均到达(负载为 R), 则总负载在两个服务系统中的分配因子定义如下:

$$W_1 = R \frac{R_1}{R_1 + R_2} \quad (3.4)$$

$$W_2 = R \frac{R_2}{R_1 + R_2} \quad (3.5)$$

其中, W_1 和 W_2 分别表示总负载 R 在乘上对应的分配因子在两个节点网络中的分配。这个分配方法

保证了根据各自的分配因子, 蜂窝网络和 Wi-Fi 网络能够公平且合理地分配总带宽, 从而实现流量卸载。

4. 仿真结果及分析

在本文中, 我们使用 Matlab 仿真平台验证所提出算法的有效性。首先分析不同节点服务速率下, 排队系统可支持的最大到达率。假设业务流量为泊松过程, 可支持的最大到达率指的是泊松过程的平均到达。需给定时延违反概率阈值等参数, 设置的时延违反概率要求为 10^{-3} , 二分法允许误差概率在 10^{-4} 范围内。时延阈值为 10 时隙。服务速率 C 的范围从 5 packets/slot 到 10 packets/slot。系统服务速率与可支持的最大到达率之间的关系如图 1 所示。

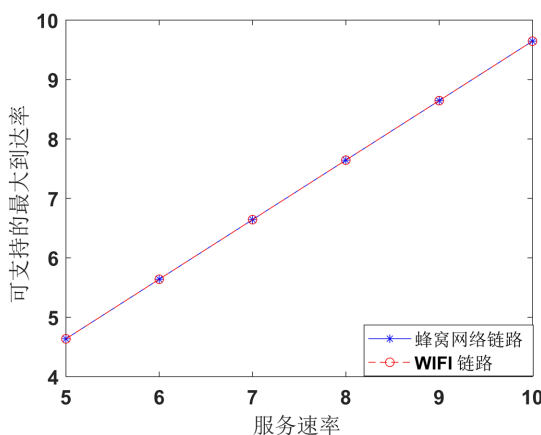


Figure 1. The service rates vs. supported maximum arrival rates

图 1. 服务速率与可支持的最大到达率

在图 1 中, 假设蜂窝链路和 Wi-Fi 链路对应的排队系统服务速率相同, 且都建模为常速率服务, 因此两个系统可支持的最大到达率也是相同的。由图看出, 可支持的最大到达率随着系统服务速率的增加而增加, 这表示随着蜂窝网络和 Wi-Fi 网络平均服务速率的提高, 系统可以承受更高的到达平均负载。这一结论也符合工程上对业务流进行时延 QoS 保障的直观规律, 对于具有复杂随机特性的业务流量, 只能通过提高服务能力的方式保障其可靠性要求。本文设计的算法能够为网络建设与优化提供一定的指导, 将系统服务能力与可支持的业务复杂上限进行映射。

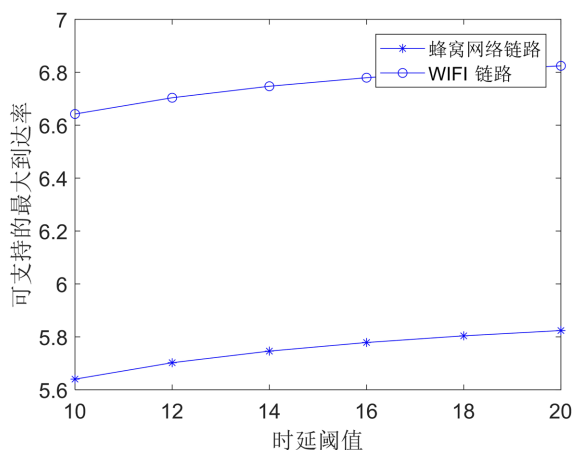


Figure 2. The delay thresholds vs. supported maximum arrival rates

图 2. 时延阈值与支持的最大到达的关系

图 2 给出了在两种节点的服务速率不同情况下, 分别对应的可支持的最大到达率。设置时延违反概率要求为 10^{-3} , 允许误差在 10^{-4} 范围内。蜂窝网络服务速率 C 设为 6 packets/slot, Wi-Fi 网络服务速率 C 设为 7 packets/slot。设置时延阈值范围从 10 到 20 时隙。

由图 2 可以看出, 随着时延阈值的增加, 系统可支持的最大到达率增大。时延阈值指的是系统允许时延上界, 阈值增大则意味着对时延的容忍程度增强, 从而使得在服务速率不变的情况下, 系统可以容纳更多的到达而不会超过设定的时延违反概率要求。由于 Wi-Fi 链路的服务速率设定高于蜂窝节点, 因此前者对应的排队系统可支持的到达负载更高。到达过程是一个具有复杂随机特性的随机过程, 服务速率变化后直观上我们无法得到相对于负载上限的变化情况, 而根据本文设计的算法, 可以在理论上给出负载与服务能力的对应关系。且本文设计的负载估计算法是在时延 QoS 约束下实现的, 可以保证只要系统的平均到达速率设置在可支持的最大到达率之下, 业务流量的 QoS 要求都可以得到保障。

在研究迭代次数与收敛速度的关系时, 设置的时延违反概率阈值为 10^{-3} , 允许误差在 10^{-4} 范围内。蜂窝网络服务速率 C 设为 6 packets/slot, Wi-Fi 网络服务速率 C 设为 7 packets/slot。设置时延阈值为 10 slots。二分法的迭代收敛情况如图 3 所示。

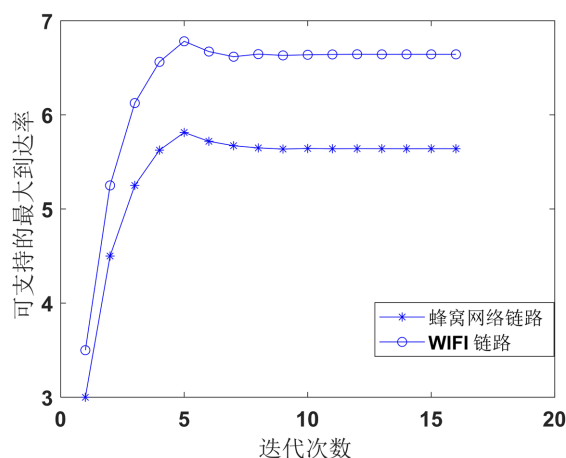


Figure 3. The iterations vs. supported maximum arrival rates

图 3. 迭代次数与可支持最大到达率的关系

图 3 展示了蜂窝网络和 Wi-Fi 网络两个系统对应的流量卸载算法的迭代情况。蓝色星号线表示蜂窝网络的可支持最大到达随迭代次数的变化。蓝色圆形线表示 Wi-Fi 网络能承受的负载随迭代次数收敛的趋势。由图 3 可知, 随着迭代次数增加, 两个系统对应的可支持最大到达率最终均收敛到一个稳定值。且迭代次数均较小, 本文设计的流量卸载算法可以快速收敛, 计算复杂度较低, 算法逻辑较简单轻量。从最后收敛结果来看, 两个系统可支持的最大到达率均低于系统各自的服务速率, 这是由于业务流量具有时延 QoS 要求。若业务的 QoS 较为严苛, 例如 5G 当中的高可靠低时延类业务, 网络系统只能以更高的服务能力来保障其 QoS 要求。在这种情况下系统的带宽利用率是较低的。因此在未来网络中, 系统只能以大带宽消耗为代价支持高可靠低时延场景的落地。

图 4 展示了不同时延阈值下, 蜂窝与 Wi-Fi 两个系统对应的时延违反概率理论值的变化情况。设定到达过程为泊松过程, 平均到达速率设为 5 packets/slot, 蜂窝网络服务速率 C 设为 6 packets/slot, Wi-Fi 网络服务速率 C 设为 7 packets/slot。设置时延阈值范围为 0 到 10 slots。图 4 中曲线为基于有效容量有效带宽理论推导的时延违反概率上界理论结果。

图 4 中两条曲线分别表示蜂窝网络和 Wi-Fi 网络的时延违反概率随时延阈值的变化情况。随着时延

阈值的增加, 时延违反概率逐渐减少。因为较大的时延阈值代表更高的容忍度, 从而降低了违反时延限制的概率。在时延阈值较小, 即时延 QoS 要求较为严格的情况下, 服务能力高的系统可以有效地降低时延违反事件发生的概率。而当时延阈值的要求放宽后, 服务速率对时延性能的优化可以达到更低的数量级。图 4 是执行负载分配算法后两个系统对应的时延违反概率情况, 可以发现负载重新分配后, Wi-Fi 网络的时延违反概率有显著下降, 而蜂窝网络的时延违反概率稍有下降。整体仍然控制在较低的水平, 这表明负载重新分配策略有效地改善了系统的性能。综上所述, 通过合理的负载分配, 可以有效降低蜂窝网络的时延违反概率, 同时保持 Wi-Fi 网络的较低时延违反概率, 从而优化整个系统的性能。

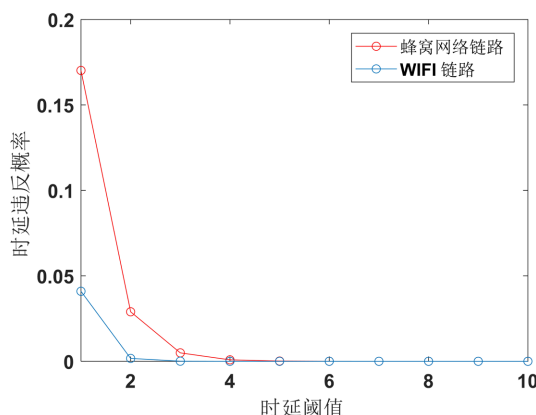


Figure 4. The delay thresholds vs. the delay violation probability

图 4. 时延违反概率与时延阈值的关系

5. 结论

本文以时延 QoS 约束为基础, 依靠排队论, 有效容量理论, 有效带宽理论等理论知识。研究了到达为泊松到达, 服务机制为常速率服务的情况下的一种新的卸载策略。基于有效容量有效带宽理论推导的时延违反概率理论上界, 设计了排队系统最大可支持到达率估计算法, 并以此设计了流量卸载算法。通过 MATLAB 建模分析, 对算法的性能表现进行了评估。分析结果, 卸载算法能在保证时延 QoS 约束的条件下, 有效降低时延, 提高信道资源利用率。

参考文献

- [1] 高建勋, 张永红. 通信网络资源管理思路探讨[J]. 中国新通信, 2021, 23(3): 27-29.
- [2] Tolba, B., Abo-Zahhad, M., Elsabrouty, M., Uchiyama, A. and El-Malek, A.H.A. (2024) Joint User Association, Service Caching, and Task Offloading in Multi-Tier Communication/Multi-Tier Edge Computing Heterogeneous Networks. *Ad Hoc Networks*, **160**, Article 103500. <https://doi.org/10.1016/j.adhoc.2024.103500>
- [3] Wu, D. and Negi, R. (2003) Effective Capacity: A Wireless Link Model for Support of Quality of Service. *IEEE Transactions on Wireless Communications*, **2**, 630-643.
- [4] Wang, J. and Zhang, X. (2016). Heterogeneous QoS-Driven Resource Adaptation over Full-Duplex Relay Networks. 2016 *IEEE Global Communications Conference (GLOBECOM)*, Washington, 4-8 December 2016, 1-6. <https://doi.org/10.1109/glocom.2016.7841869>
- [5] Zhang, X. and Wang, J. (2018) Joint Heterogeneous Statistical-QoS/QoE Provisionings for Edge-Computing Based Wifi Offloading over 5G Mobile Wireless Networks. 2018 *52nd Annual Conference on Information Sciences and Systems (CISS)*, Princeton, 21-23 March 2018, 1-6. <https://doi.org/10.1109/ciss.2018.8362265>
- [6] Chang, C. (2000) *Performance Guarantees in Communication Networks*. Springer-Verlag.
- [7] 张志飞, 王黎伟, 裴正定. 时延约束下的等效带宽估计[J]. 通信学报, 2000(12): 73-77.
- [8] 石文孝. 通信网理论与应用[M]. 第 2 版. 北京: 电子工业出版社, 2016.

-
- [9] Chang, C.S. and Thomas, J.A. (1999) Effective Bandwidth in High-Speed Digital Networks. *IEEE Journal on Selected Areas in Communications*, **13**, 1091-1100. <https://doi.org/10.1109/49.400664>
- [10] 陈有汉, 胡仲海, 桂志波. 考虑不确定信息的 QoS 路由算法综述[J]. 计算机工程与应用, 2006, 42(15): 125-128.
- [11] 迟学芬, 马丽君, 王雅辉, 等. 针对 M2M 业务的 IEEE 802.11 优化及其性能评价[J]. 吉林大学学报(工学版), 2014, 44(5): 1488-1497.
- [12] 张品. 网络中 QoS 路由问题的研究[D]: [博士学位论文]. 成都: 电子科技大学, 2004.
- [13] 余翔, 曲原宇, 杨路. 空天地边缘计算网络任务卸载策略[J/OL]. 电讯技术: 1-10. <https://doi.org/10.20079/j.issn.1001-893x.240110001>, 2024-05-31.
- [14] 杨守义, 李富康, 任瑞敏. 信任环境下考虑系统公平性的边缘计算卸载策略和资源分配[J]. 通信学报, 2024, 45(3): 142-154.
- [15] 张光华, 徐航, 万恩晗. 面向移动边缘计算的任务卸载方法研究[J]. 吉林大学学报(信息科学版), 2024, 42(2): 210-216.
- [16] Cheng, W., Zhang, X. and Zhang, H. (2016) Optimal Power Allocation with Statistical QoS Provisioning for D2D and Cellular Communications over Underlaying Wireless Networks. *IEEE Journal on Selected Areas in Communications*, **34**, 151-162. <https://doi.org/10.1109/jsac.2015.2476075>
- [17] Sun, W. and Li, J. (2023) Power Allocation Optimization for Hybrid Information and Energy Transfer with Massive MIMO Downlink. *IET Communications*, **17**, 1998-2008. <https://doi.org/10.1049/cmu2.12672>
- [18] 沈伟达. 时延 QoS 约束下的带宽估计与服务优化研究[D]: [硕士学位论文]. 吉林: 吉林化工学院, 2022.
- [19] Ya, M., Ase, A. and La, H. (2021) A Queueing Theory Approach to Traffic Offloading in Heterogeneous Cellular Networks. *AEU-International Journal of Electronics and Communications*, **139**, Article ID: 153910. <https://doi.org/10.1016/j.aeue.2021.153910>
- [20] Anand, G., Faizan, Q., Sang, D.D., *et al.* 基于排序的网络选择在 5G 异构网络中实现流量卸载[J]. 工程科技导报, 2019, 12(2): 9-16.
- [21] 孙洪亮, 赵创业, 付月. ALOHA 接入机制下时延 QoS 保障算法[J]. 软件, 2021, 42(7): 85-87.
- [22] 鲁宁宁. 5G 小蜂窝系统功率控制方法研究[D]: [硕士学位论文]. 南京: 东南大学, 2016.
- [23] Zhao, J., Ning, Y., Gong, Y. and Ran, R. (2017) Power Control with Power Budget for Uplink Transmission in Heterogeneous Networks. 2017 *IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Toronto, 24-27 September 2017, 1-5. <https://doi.org/10.1109/vtcfall.2017.8287973>
- [24] 王俊社, 蒋彤彤, 张琛, 等. 异构网络中基于非合作博弈的功率控制算法[J]. 科学技术与工程, 2019, 19(29): 160-165.
- [25] 占庆祥, 刘如通, 谭国平. 基于强化学习的异构网络流量卸载方案研究[J]. 电子测量技术, 2018(2): 66-71.

附 录

本文中设计的系统可支持最大到达率估计算法核心代码如下:

```

clc; clear all;
syms R;
threshold = 10-3; % The threshold of delay-violation probability.
acpErr = 10-4; % Acceptable error for bandwidth estimation.
C = 6:7;
delay Target = 10;
for k = 1:length (C)
    KS = C (k);
    for i = 1:length (delay Target)
        High = C (k);
        Low = 0; % Determine the lower bound.
        l = 1;
        while (Low < High) % Start to perform binary search algorithm.
            Mid (k, l) = (High + Low) / 2;
            f Theta = @(x) Determine Aggregate Theta (Mid (k, l), KS, x);
            initial Theta = Range Theta ();
            initial Theta Num = length (initial Theta);
            theta Tmp = zeros (initial Theta Num, 1);
            for j = 1:initial Theta Num
                theta Tmp (j) = f solve (f Theta, [initial Theta (j)], optimset ('Display', 'off'));
            end
            MTheta Value = max (theta Tmp);
            syms delay;
            f Delay = (-MTheta Value * Mid(k, l) * delay) - log (threshold );
            delay Value = solve (f Delay, delay);
            delay Value = double(delay Value);
            if (abs (delay Value - delay Target (i)) < acpErr)
                R_estimation (k, i) = Mid (k, l); % Determine the super martingale bandwidth.
                break;
            else
                if (delay Value - delay Target (i) < 0)
                    Low = Mid (k, l);
                else
                    High = Mid (k, l);
                end
                l = l + 1;
            end
        end
        if (High = Low)

```

```
        fprintf('this is an error');
        break;
    end
end
end
function [fsp] = Determine Aggregate Theta (lamda, C, theta)
    fsp = (lamda*(exp (theta) - 1)/theta) - C;
end
function [Initial Theta] = Range Theta ( )
    Initial Theta = [10e-10, 10e-9, 10e-8, 10e-7, 10e-6, 10e-5, 10e-4, 10e-3, 10e-2, 10e-1, 1, 10, 5e-10,
        5e-9, 5e-8, 5e-7, 5e-6, 5e-5, 5e-4, 5e-3, 5e-2, 5e-1, 5];
end
```