

doi: 10.11835/j.issn.1000-582X.2025.08.008

引用格式: 王诗, 曹大焱, 朱笑莹, 等. 车联网业务特性模型下卸载反馈策略的设计与评估[J]. 重庆大学学报, 2025, 48(8): 86-98.



车联网业务特性模型下卸载反馈策略的设计与评估

王诗^a, 曹大焱^{a,b}, 朱笑莹^{a,b}, 王铭宇^a, 王浩颖^a

(辽宁工程技术大学 a. 电子与信息工程学院; b. 研究生院, 辽宁葫芦岛 125100)

摘要: 随着移动边缘计算车联网中业务特征逐渐变多, 针对设计卸载反馈策略时需考虑业务特征的复杂建模问题, 实现服务器到端传输的性能评估是一大挑战。为实现性能的多维度综合评估, 笔者基于排队论和马尔可夫调制服务过程构建了一种考虑时变多类型业务的缓存调度策略评估框架。为适应各种通信环境, 提出的框架可灵活调控业务特性、双端处理速率和卸载反馈策略。基于该框架, 提出一种基于概率分布的卸载反馈策略。实验结果表明, 所提出策略的传输性能较传统优越 50%, 证明所提框架能在不同通信环境和硬件配置下为设计策略提供参考。

关键词: 车联网; 边缘计算; 业务模型; 马尔可夫调制服务; 资源分配

中图分类号: TN915.01

文献标志码: A

文章编号: 1000-582X(2025)08-086-13

Design and evaluation of an offload feedback strategy framework based on a service characteristic model in the internet of vehicles

WANG Shi^a, CAO Dayan^{a,b}, ZHU Xiaoying^{a,b}, WANG Mingyu^a, WANG Haoying^a

(a. School of Electronic and Information Engineering; b. Institute of Graduate, Liaoning Technical University, Huludao, Liaoning 125100, P. R. China)

Abstract: With the proliferation of diverse service characteristics in the internet of vehicles (IoV) under the mobile edge computing (MEC) paradigm, evaluating server-to-end transmission performance presents a significant challenge, particularly due to the complex modeling requirements that must account for service-specific traits in offload feedback strategies. To address this, a cache scheduling evaluation framework is proposed, incorporating time-varying and multi-type services based on queuing theory and a Markov-modulated service process. The proposed framework supports flexible adjustments to service characteristics, bidirectional processing rates, and offload feedback strategies, enabling it to adapt to various communication environments. Within this framework, an offload feedback strategy based on statistical prediction is proposed. Numerical simulations show that the proposed strategy improves transmission performance by approximately 50% compared with traditional approaches. These findings indicate that the proposed framework provides a valuable reference for designing adaptive strategies under diverse network conditions and hardware configurations.

收稿日期: 2024-03-11 网络出版日期: 2025-02-27

基金项目: 国家重点研发计划(2018YFB1403303); 辽宁省博士启动基金(2019-BS-114)。

Supported by Research Fund for National Key Research and Development Program of China (2018YFB1403303) and the Research Foundation for Doctoral Program of Liaoning Province (2019-BS-114).

作者简介: 王诗(1983—), 男, 副教授, 主要从事认知无线电、排队论、最优化理论方向研究, (E-mail)wangshi@lntu.edu.cn。

通信作者: 曹大焱(2000—), 男, (E-mail)houcdy@163.com。

Keywords: internet of vehicles; mobile edge computing; service modeling; Markov-modulated services; resource allocation

近年来,随着无线通信和物联网技术的发展,车联网(internet of vehicles, IoV)已成为5G的重要应用场景。在移动边缘计算技术背景(mobile edge computing, MEC)下,车联网中路边单元携带的MEC服务器和智能车辆配备的车载单元(on board unite, OBU)都具备计算和存储能力^[1]。基于MEC系统计算和缓存的功能,学者提出服务缓存^[2]和边缘缓存技术^[3]并衍生了任务卸载、资源分配等研究。在网络层上,任务卸载和资源分配问题被建模为最优化问题。针对最优化模型,张建军等^[4]提出一种多MEC联合卸载的方案,李方伟等^[5]提出了V2X(vehicle-to-everything)协同缓存与资源分配机制。由于引入多样化业务模型会使最优化模型出现计算成本高的问题,上述研究在完成资源分配时未考虑到业务相关性和优先级等多样化特征。然而,不同类型业务的网络需求和处理方式并不相同。例如业务为时延敏感和上下文敏感的应用程序^[6],则应卸载到MEC服务器,其他为安全性服务的重要业务应该在本地进行服务和保存。

目前车联网的典型业务包括:驾驶安全、交通效率、信息服务和管理综合4类业务^[7-8]。这些业务在网络端可定义为流量特性、可靠实时特性、忙时特性、移动性、触发特性和附着特性的量化组合^[9]。但相关研究^[10-11]尚缺少针对车联网中车辆业务通信的系统建模,都是在网络层通过多要素预测单一业务特性的变化。因此,对设计任务卸载策略而言,建立业务模型呈现车联网中多种业务特性对边缘计算效率的影响有重要意义。目前Zhu等^[12]将具备间歇传输特性的卫星业务传输过程建模为基于马尔可夫链蒙特卡罗的马尔可夫调制服务过程(markov chain monte carlo based markov modulated service process, MMSP)模型,证明了MMSP模型可用于链路层业务建模,并呈现到达业务状态之间的转移过程。

除业务类型外,边缘计算中业务卸载与MEC、OBU 2端的CPU计算周期速率也息息相关^[13]。相关学者对此进行了研究,Peng等^[14]在设计联合缓存和卸载策略时考虑了应用程序服务提供商的租赁成本和不同车辆计算能力的差异,戚艾林等^[15]在设计卸载策略时考虑了车辆快速移动导致的回传时延问题。由于CPU计算周期和分配策略分属物理层和网络层,难以在最优化模型中考虑CPU计算周期对数据量的变化影响,上述研究都假设业务卸载量不会导致OBU缓存溢出,不考虑双端计算速率的差异而进行业务卸载,卸载业务的计算数据容易导致OBU缓存溢出或欠载。OBU卸载任务时往往依据当前时刻的OBU缓存容量进行分配,但当MEC处理并回传任务数据时,OBU的缓存容量已发生变化。由于MEC计算任务和数据回传都存在时延且MEC的CPU计算周期数比OBU大。在MEC计算任务和进行数据回传时,OBU由于并行处理本地业务会导致缓存量较分配任务时增多或减少。在接收MEC数据时,OBU容易因业务卸载量分配有误出现数据溢出或欠载。为了设计跨层优化策略,链路层评估框架已证实是有效的。耿珂等^[16]将车辆高速移动的影响抽象为信道相关系数,采用蒙特卡罗方法进行链路层仿真。Zhang等^[17]为在考虑异构节点的同时分析网络性能,将各节点之间的数据传输建模为排队模型,基于马尔可夫分析方法建立了链路层策略评估框架。

考虑到业务常规的流量特性、可靠实时特性、触发特性和忙时特性,研究利用马尔可夫模型对车联网业务建模,通过离散排队分析完成业务时变性的量化业务模型。此外,借助MMSP模型,在对OBU和服务器的数据流排队分析时,考虑双端传输速率提出一种通用化评估框架。该框架完成了业务类型的模块化标准,实现传输环境的可配置性,提供诸如吞吐量、拒绝率和排队时延等系统指标。借助该框架,基于概率分布提出一种跨层的卸载反馈策略。

1 系统模型

1.1 基础模型

考虑移动的单个车辆和MEC服务器之间计算任务的数据包传输过程,如图1所示。车辆在移动过程中会产生娱乐业务、安全性业务等需求,这些业务需求被车载单元根据业务类型和流行度等因素卸载到本地或者MEC服务器进行计算,计算结果最终回传到车载单元进行处理显示。

假设在业务和数据传输过程中,所有数据都以离散数据包的形式传输,且数据传输过程都是如图2所示

的固定时隙结构。当到达业务量过多出现溢出时,溢出部分在服务器处理,其他业务在本地处理,否则将业务依据业务特性卸载到服务器或者 OBU。每个时隙由 3 个弱关联的流程组成:业务的产生到达流程、车辆对业务的处理流程及云端业务的处理反馈流程。假定车辆本地的业务处理和传输速率为时不变,且车辆本地接收到达业务和处理业务时不会发生业务和数据包遗漏。此外,假设 MEC 服务器的任务处理过程不存在业务请求和数据包丢失。

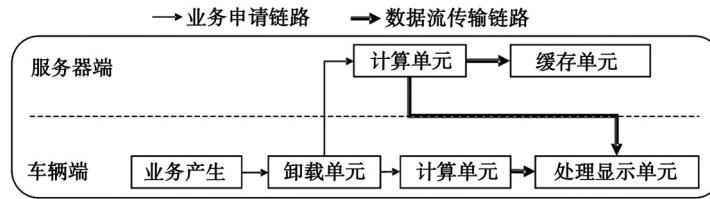


图 1 系统模型

Fig. 1 System model

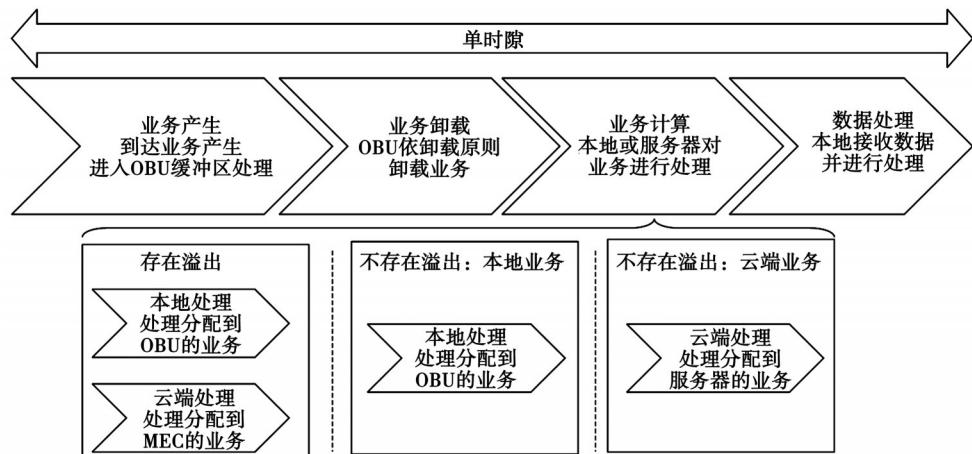


Fig. 2 Time slot diagram of business processing

考虑车辆和服务器处理传输速率差距的系统读写流程如图 3 所示。根据时隙的主要特征,或者占用时隙主要时长的阶段不同,可以把时隙分为写入业务、繁忙和溢出 3 种。写入时隙代表大量业务写入的时隙,繁忙时隙代表算力主要用来处理业务计算和处理的时隙,溢出时隙代表服务器传输的数据量过大导致 OBU 缓存区溢出。溢出时隙数据包会被丢弃,因此,下一时隙仍需要重新计算和回传数据包。

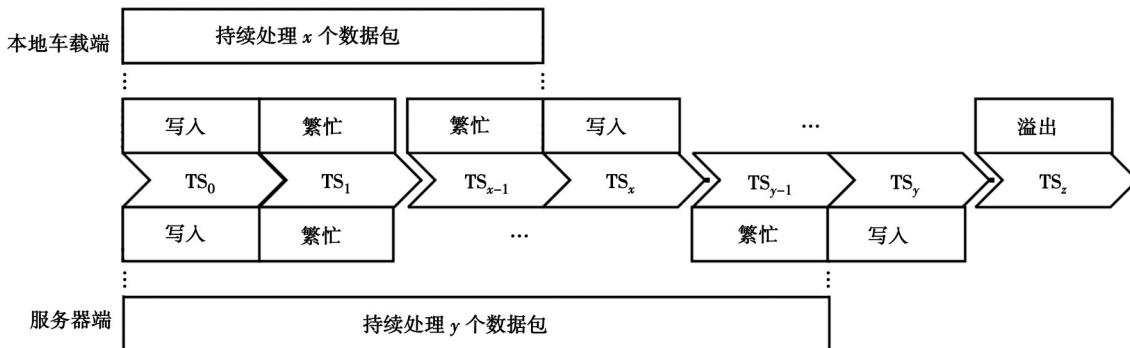


图 3 处理速率下系统的读写示意图

Fig. 3 Schematic diagram of system r/w at processing rates

1.2 马尔可夫调制业务模型

为了能在评估服务器到端传输性能时灵活调控到达业务的特性,从业务流量特性、触发特性、可靠实时

特性和忙时特性4个业务特征对所有业务进行重新定义。

假设OBU的容量为 L_b ,云端车载的容量为 $L_m=L_b+L_a$, L_a 为输入单元可以发送数据包的最大数量。业务流量特性的状态总数为 S_n ,第 i_n 种业务流量特性到达状态下业务包量 n_i 服从分布 $\xi_i^n=[p_n(0),\dots,p_n(j_n),\dots,p_n(L_a-1)]$, $\sum_{j_n=0}^{L_a-1}p_n(j_n)=1$,其中, $0\leq p_n(j_n)\leq 1$ 表示传输 j_n 个数据包的概率, $i_n\in D=\{1,2,\dots,S_n\}$, $j_n\in\{0,1,2,\dots,L_a-1\}$ 。业务流量特性到达状态之间的转移服从状态转移矩阵 I

$$I=\begin{bmatrix} p_{0\rightarrow 0} & \cdots & p_{0\rightarrow S_n-1} \\ \vdots & \ddots & \vdots \\ p_{S_n-1\rightarrow 0} & \cdots & p_{S_n-1\rightarrow S_n-1} \end{bmatrix}_{(S_n\times S_n)}, \quad (1)$$

式中: $\sum_{j=0}^{S_n-1}p_{i\rightarrow j}=1$, $0\leq p_{i\rightarrow j}\leq 1$ 表示当前时隙为第 i 种流量特性到达状态,下一时隙为第 j 种流量特性到达状态的概率。当前时隙业务可靠实时特性的到达状态是 $i_l\in K=\{1,2,\dots,S_l\}$ 时, S_l 是可靠实时特性到达状态的总数,可靠实时特性 i_l 服从分布 $\xi_l^i=[p_l(0),\dots,p_l(j_l),\dots,p_l(L_l-1)]$, L_l 表示该状态总数。业务可靠实时特性到达状态之间的转移服从状态转移矩阵 I

$$I=\begin{bmatrix} p_{0\rightarrow 0} & \cdots & p_{0\rightarrow S_l-1} \\ \vdots & \ddots & \vdots \\ p_{S_l-1\rightarrow 0} & \cdots & p_{S_l-1\rightarrow S_l-1} \end{bmatrix}_{(S_l\times S_l)}. \quad (2)$$

当前时隙业务触发特性的到达状态是 $i_z\in E=\{1,2,\dots,S_z\}$ 时,触发特性 z_i 服从分布 $\xi_z^i=[p_z(0),\dots,p_z(j_z),\dots,p_z(L_z-1)]$,其中, L_z 表示该状态总数, S_z 是触发特性到达状态总数。业务触发特性到达状态之间的转移服从状态转移矩阵 G

$$G=\begin{bmatrix} p_{0\rightarrow 0} & \cdots & p_{0\rightarrow S_z-1} \\ \vdots & \ddots & \vdots \\ p_{S_z-1\rightarrow 0} & \cdots & p_{S_z-1\rightarrow S_z-1} \end{bmatrix}_{(S_z\times S_z)}. \quad (3)$$

业务特性的忙时特性 $w\in O=\{1,2,\dots,8\}$ 是指某一特性业务在一天中一段时间内集中爆发的性质^[18]。依据业务特性,存在 $w_n\in\{0,1\}$ 、 $w_l\in\{0,1\}$ 和 $w_z\in\{0,1\}$ 分别表示业务流量特性、可靠实时特性和触发特性的忙时特性。为方便表示,令 $w=w_n+2w_l+4w_z$ 。各特性业务的爆发密度由忙时集中系数 $0\leq u\leq 1$ 呈现,爆发时长由忙时时长 $0\leq r\leq 24$ 呈现。假设所有特性业务的忙时集中系数阈值为 u_0 、忙时时长的阈值为 r_0 ,则依据实际忙时集中系数和忙时时长,可表示一段时间内流量特性业务是否存在集中爆发的特征。当忙时集中系数大于 u_0 ,爆发时长短于 r_0 时, $w_n=1$ 表示流量特性业务在短时间内存在集中爆发的特性,反之为不存在集中爆发的特性

$$w_n=\begin{cases} 0, & u < u_0, r > r_0, \\ 1, & u \geq u_0, r \leq r_0. \end{cases} \quad (4)$$

此外,当 $w_n=1$ 时,到达业务的流量特性根据马尔可夫调制得到。当 $w_n=0$ 时,到达业务的流量特性状态依据流量特性到达的稳态分布 π_n 采样获得。

1.3 缓存模型

为方便计算性能评估指标,将OBU和MEC服务器实体建模为数据包缓存队列。MEC服务器内存储的业务类型 l 服从分布

$$\psi=[p_\psi(0),\dots,p_\psi(i_\psi),\dots,p_\psi(S_h-1)], \quad (5)$$

式中: $i_\psi=w+8i_z+8S_zS_li_n$ 表示流量特性状态为 i_n ,可靠实时特性状态为 i_l ,触发特性状态为 i_z ,忙时特性为 w 的业务; $p_\psi(i_\psi)$ 表示业务类型为 i_ψ 产生的概率。 $S_h=8S_nS_lS_z$ 为到达业务特征状态的总数, S_n 为流量特性到达状态总数、 S_l 为可靠实时特性到达状态总数、 S_z 为触发特性到达状态总数。相似的,OBU端内存储的业务类型 \mathcal{L} 服从分布

$$\varphi = [p_\varphi(0), \dots, p_\varphi(i_\varphi), \dots, p_\varphi(S_\varphi - 1)]^T. \quad (6)$$

因此,根据业务类型的状态可得双端存储业务的4个特征的状态。当前时隙数据包在卸载前,OBU中数据包的数量被表示为 $b \in P = \{0, 1, \dots, L_b\}$,且 b 服从如下概率分布

$$\varphi_0^n = [\varphi_0^n(0), \dots, \varphi_0^n(b), \dots, \varphi_0^n(L_b)], \quad (7)$$

式中: $\varphi_0^n(b)$ 表示卸载前 OBU 中存在 b 个数据包的概率; $\sum_{b=0}^{L_b} \varphi_0^n(b) = 1$ 。 OBU 的 L_b+1 种缓存状态之间的转移概率可以用 \mathbf{B} 表示为

$$\mathbf{B} = \begin{bmatrix} p_{0 \rightarrow 0} & \cdots & p_{0 \rightarrow L_b} \\ \vdots & \ddots & \vdots \\ p_{L_b \rightarrow 0} & \cdots & p_{L_b \rightarrow L_b} \end{bmatrix}_{(L_b+1 \times L_b+1)}, \quad (8)$$

式中: $p_{i \rightarrow j}$ 表示上一时隙 OBU 缓存内存在 i 个数据包,且本时隙存在 j 个数据包的概率。分配后车载端存储的数据包数量满足的概率分布被 φ_1^n 表示,MEC 服务器向车载端传输后车载端内数据包数量服从的概率分布为 φ_2^n, φ_3^n ,是 OBU 处理数据包后其内部数据包数量满足的概率分布。当前时隙数据包分配前 MEC 服务器内包含的数据包数量描述为 $u \in U = \{0, 1, \dots, L_m\}$,且被规定服从概率分布

$$\psi_0^n = [\psi_0^n(0), \dots, \psi_0^n(i), \dots, \psi_0^n(L_m)]. \quad (9)$$

MEC 服务器的 L_m+1 种缓存状态之间的转移概率则可以用 \mathbf{J} 表示

$$\mathbf{J} = \begin{bmatrix} q_{0 \rightarrow 0} & \cdots & q_{0 \rightarrow L_m} \\ \vdots & \ddots & \vdots \\ q_{L_m \rightarrow 0} & \cdots & q_{L_m \rightarrow L_m} \end{bmatrix}_{(L_m+1 \times L_m+1)}, \quad (10)$$

式中: $q_{i \rightarrow j}$ 表示上一时隙 MEC 服务器缓存内存在 i 个数据包,且本时隙存在 j 个数据包的概率。

1.4 马尔可夫调制通信模型

车辆和服务器之间的通信采用取自 HIPERLAN/2 和 IEEE 802.11a 标准的 AMC 调制方式。单车辆模型考虑噪声信道,将传输信道的信噪比划分为 N_{SNR} 个 SNR 状态。考虑到传输信道的时变性,将 N_{SNR} 个 SNR 信道条件状态的演化建模为马尔可夫链^[19-20]。该模型由 $N_{\text{SNR}} \times N_{\text{SNR}}$ 的状态转移矩阵 \mathbf{F} 描述

$$\mathbf{F} = \begin{bmatrix} p_{1 \rightarrow 1} & \cdots & p_{1 \rightarrow N_{\text{SNR}}} \\ \vdots & \ddots & \vdots \\ p_{N_{\text{SNR}} \rightarrow 1} & \cdots & p_{N_{\text{SNR}} \rightarrow N_{\text{SNR}}} \end{bmatrix}_{(N_{\text{SNR}} \times N_{\text{SNR}})}, \quad (11)$$

式中: $p_{i \rightarrow j}$ 表示信道的信道条件状态由上一时隙状态 i 转移为本时隙状态 j 的概率。若当前时隙信道条件状态为 $c \in Q = \{1, 2, \dots, N_{\text{SNR}}\}$,根据调制方案可以获得信道条件状态为 c 的传输系统数据包传输概率分布 ξ_c^{channel}

$$\xi_c^{\text{channel}} = [p_c^{\text{channel}}(0), \dots, p_c^{\text{channel}}(j), \dots, p_c^{\text{channel}}(L_b)], \quad (12)$$

式中: $p_c^{\text{channel}}(j)$ 信道条件状态为 c 时,当前时隙内传输 j 个数据包的概率。

2 队列服务分析框架

2.1 队列分析

为推导出系统状态转移矩阵,笔者用排队理论分析数据包的传递过程。

1) 业务到达过程

在给定流量特性的忙时特性 w_n 基础上,依据马尔可夫调制过程获取当前时隙到达业务流量 n 服从的概率分布为

$$\mathbf{r} = [p_n(0), \dots, p_n(j_n), \dots, p_n(L_n - 1)], \quad (13)$$

式中: $\sum_{j_n=0}^{L_n-1} p_n(j_n) = 1$; L_n 表示数据到达过程可能到达的最大数量; $p_n(j_n)$ 表示在 i_n 状态下到达 j_n 个数据包的

概率。

2) 业务卸载过程

当前时隙到达业务的可靠实时特征为 l_i , 服从分布为 ξ_i^l 。 α_b 和 α_s 分别表示到达业务分配给 OBU 和 MEC 服务器的待处理数据包数量。分配过程可以用 2 者服从的概率分布向量 τ 和 y 描述。OBU 当前阶段的业务类型为 \mathcal{L}_0 , 其代表的业务流量类型 \mathcal{L}_0^n 服从分布

$$\varphi_0^n = [\varphi_0^n(0), \dots, \varphi_0^n(b), \dots, \varphi_0^n(L_b)]。 \quad (14)$$

当下业务的卸载主要依赖于业务的流量特征和可靠实时特性特征。当业务数据量过大时会溢出 OBU 缓存时, 溢出部分卸载到服务器处理。当业务数据量较小时, 可靠实时业务在本地处理, 其他业务在云端处理。当到达业务流量使 OBU 溢出时, 卸载到服务器的业务流量 α_s 服从分布 $y' = [p_y'(0), \dots, p_y'(i_y), \dots, p_y'(L_{\alpha}-1)]$, 该分布满足如下关系

$$p_y'(i_y) = \sum_{b < i_y + L_b} \varphi_0^n(b) p_n(i_y + L_b - b), \quad (15)$$

式中: $p_n(j_n)$ 表示在当前到达状态下业务到达 j_n 个数据包的概率。保留 OBU 端的业务流量 α_b 服从的分布 $\tau' = \text{Conv}(\tau, y'^{-1})$ 。因此考虑保存在 OBU 端的业务流量服从分布为

$$\tau = \begin{cases} 0, & E(l_i) < \lambda_i, E(n_i) < L_b - E(\mathcal{L}_0^n), \\ r, & E(l_i) \geq \lambda_i, E(n_i) < L_b - E(\mathcal{L}_0^n), \\ \tau', & E(n_i) \geq L_b - E(\mathcal{L}_0^n), \end{cases} \quad (16)$$

式中: λ_i 为可靠实时特性阈值, 对应的保存在服务器端的业务流量 α_s 服从的分布 $y = \text{Conv}(\tau, \tau'^{-1})$, 其中 τ'^{-1} 表示 τ 的转秩分布。考虑到车载端和服务器端处理速率的差距^[21], v_H 和 v_L 分别表示单次业务处理过程中 OBU 端和服务器端处理业务所需要的时隙数。OBU 的业务数据流量在业务卸载后服从分布 $\varphi_1^n = T(\varphi_0^n, \tau, v_H)$, 服务器端的业务数据流量服从分布 $\psi_1^n = T(\psi_0^n, \tau, v_L)$ 。其中, $T(x, y, z)$ 为

$$T(x, y, z) = \begin{cases} x, & t \neq kz, k = \mathbb{N}^+ \\ \text{Conv}(x, y), & t = kz, k = \mathbb{N}^+ \end{cases}。 \quad (17)$$

3) 结果反馈过程

服务器向 OBU 回传的数据量由任务卸载后的数据流量、信道条件和卸载反馈策略决定。依据卸载反馈策略, 服务器向 OBU 回传的数据量分布为

$$P_m = [p_m(0), \dots, p_m(i_m), \dots, p_m(L_m)]。 \quad (18)$$

考虑信道条件限制, 当信道条件已知时, 前时隙服务器发送数据包的调制方案为 $\xi_c = [p_c(0), \dots, p_c(j), \dots, p_c(L_b)]$ 。服务器可向 OBU 传输的数据包数量服从分布 $\omega = [p_\omega(0), \dots, p_\omega(i_\omega), \dots, p_\omega(L_m)]$, 该分布满足如下关系

$$p_\omega(i_\omega) = \sum_{j=i_\omega}^{L_b} p_c(j) p_m(i_m) + \sum_{i_m=i_\omega+1}^{L_m} p_c(j) p_m(i_m), \quad (19)$$

因此, 在服务器向 OBU 传输数据后, 服务器端的数据包数量服从概率分布为 $\psi_2^n = T(\psi_1^n, \omega^{-1}, v_L)$ 。考虑服务器和 OBU 处理速率的差异, 每时隙服务器完成结果反馈后, OBU 的数据包概率分布为 $\varphi_2^n = T(\varphi_1^n, \omega, v_H)$ 。

4) 数据处理过程

服务器端的触发业务会被保存一段时间。车载端部分可靠实时业务也需要保存, 如地图等较为重要的数据。假设可靠实时特性 l_i^2 大于 σ_i 的业务流需要被车载端保存, 触发特性 z_i^2 大于 ρ_z 的业务流会被服务器保存。则在完成数据处理后, 本地端和服务器端的数据包数量服从的分布为

$$\varphi_3^n = \begin{cases} \varphi_2^n, & E(l_i^2) > \sigma_i, \\ 0, & E(l_i^2) \leq \sigma_i, \end{cases} \quad (20)$$

$$\psi_3^n = \begin{cases} \psi_2^n, & E(z_i^2) > \rho_z, \\ 0, & E(z_i^2) \leq \rho_z. \end{cases} \quad (21)$$

2.2 状态空间

基于马尔可夫模型对系统进行分析,设置到达状态、OBU 缓存状态、服务器缓存状态和信道状态作为主状态,构成系统状态空间

$$\Phi = \{(i_\xi, b, u, c) \mid i_\xi \in \Omega, b \in P, u \in U, c \in Q\}。 \quad (22)$$

由于到达状态由业务流量特征状态、可靠实时特征状态、触发特征状态和忙时特性状态 4 个子状态构成,到达状态空间为

$$\Omega = \{(i_n, i_l, i_z, w) \mid i_n \in D, i_l \in K, i_z \in E, w \in O\}。 \quad (23)$$

到达状态空间大小为 $S_h = 8S_n S_l S_z$, 系统状态空间大小即为 $K = S_h (L_b + 1) N_{SNR} (L_m + 1)$ 。到达空间下各状态之间的转移服从到达状态转移矩阵

$$A = I \otimes \Gamma \otimes G = \begin{bmatrix} p_{1 \rightarrow 1} & \cdots & p_{1 \rightarrow S_h} \\ \vdots & \ddots & \vdots \\ p_{S_h \rightarrow 1} & \cdots & p_{S_h \rightarrow S_h} \end{bmatrix}_{(S_h \times S_h)}, \quad (24)$$

式中: I 是到达业务的流量特性状态转移矩阵; Γ 是到达业务的可靠实时特性状态转移矩阵; G 是到达业务的触发特效状态转移矩阵。该空间下各状态之间的转移服从系统状态转移矩阵 T

$$T = A \otimes B \otimes J \otimes F = \begin{bmatrix} T_{1 \rightarrow 1} & \cdots & T_{1 \rightarrow K} \\ \vdots & \ddots & \vdots \\ T_{K \rightarrow 1} & \cdots & T_{K \rightarrow K} \end{bmatrix}_{(K \times K)} \quad (25)$$

式中: A 是到达过程状态转移矩阵; B 是 OBU 缓存状态转移矩阵; J 是 MEC 服务器缓存状态转移矩阵; F 是信道状态转移矩阵; \otimes 是克罗内克积; $T_{i \rightarrow j}$ 表示系统状态编号上一时隙为 i , 本时隙变为 j 的概率。

2.3 基于蒙特卡罗的队列仿真

由于系统状态转移矩阵过大,为获取稳态分布,使用蒙特卡罗方法对服务器到端的数据包排队演化过程进行仿真,如图 4 所示。

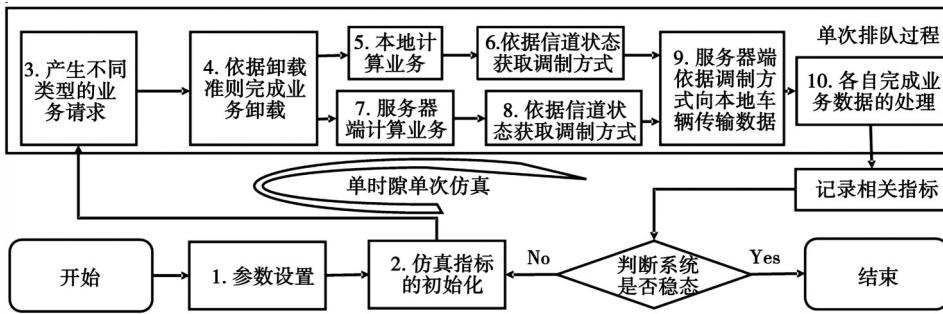


Fig. 4 Monte Carlo simulation block diagram

通过对队列模型的仿真,可得到系统的稳态分布为 $\pi = [\pi(1), \dots, \pi(S_\pi)]$, 其中 $\sum_{i=1}^{S_\pi} \pi(i) = 1$ 。稳态分布和状态转移矩阵的关系满足下式

$$\pi(T - E) = 0, \quad (26)$$

其中, E 是单位矩阵。

为了推导评价指标,规定编号为 i 的系统状态为 $\phi_i(a_i, b_i, u_i, c_i)$, 系统状态编号与到达状态、OBU 缓存状态、服务器缓存状态、信道状态满足如下关系

$$i - 1 = a_i + b_i S_h + u_i S_h (L_b + 1) + c_i S_h (L_b + 1) (L_m + 1)。 \quad (27)$$

2.4 性能指标推导

评估框架采用平均队长、平均吞吐、平均拒绝和平均时延作为性能评估指标。

1) 平均队长指标

平均队长指标定义为单位时隙内 OBU 缓存与服务器缓存的队长之和。结合稳态分布,平均队长指标为

$$\overline{Que} = \sum_{j=0}^{L_b+L_u} \sum_{b_i+u_i=j} \pi(i) j, \quad (28)$$

式中: $\pi(i)$ 表示稳定状态下系统状态为第*i*种状态的概率; b_i 表示第*i*种系统状态下的 OBU 队长; u_i 表示第*i*种系统状态下的服务器队长。

2) 平均吞吐指标

平均吞吐指标定义为单位时隙内 MEC 向 OBU 传输的数据包数量。在编号为*i*的状态下,吞吐*k*个数据包的概率 $P_o^b(i, k)$ 为

$$P_o^b(i, k) = \sum_{b=k}^{L_b} p_c(k) \varphi_2^n(b) + \sum_{b=k}^{L_b} p_c(b) \varphi_2^n(k), \quad (29)$$

式中:在给定系统状态*i*下,信道条件状态为*c*时 $p_c(k)$ 表示信道传输*k*个数据包的概率; $\varphi_2^n(k)$ 表示 MEC 服务器向 OBU 传输后 OBU 内数据包数量为*k*个数据包的概率。考虑稳态下部分系统状态会吞吐相同数量的数据包,吞吐*k*个数据包的概率 $P_{rate}^b(k)$ 为

$$P_{rate}^b(k) = \sum_{i=1}^{S_\pi} \pi(i) P_o^b(i, k), \quad (30)$$

平均吞吐为

$$\overline{Throughput} = \sum_{i=0}^{L_b} P_{rate}^b(i) i. \quad (31)$$

3) 平均拒绝指标

平均拒绝指标定义为单位时隙内 MEC 服务器向 OBU 传输却因 OBU 溢出而被丢弃的数据包数量。在编号为*i*的状态下拒绝*k*个数据包的概率 $P_{rejection}(i, k)$ 为

$$P_{rejection}(i, k) = \begin{cases} \sum_{j=0}^{L_b-b_i} p_m(j), & k=0, \\ p_m(L_b-b_i+k), & k>0, \end{cases} \quad (32)$$

式中: b_i 是在状态编号为*i*的情况下 OBU 的初始队长; $p_m(j)$ 是 MEC 向 OBU 回传时传输*j*个数据包的概率。

考虑状态空间的所有状态,有稳态下拒绝*k*个数据包的概率为 $P_{rejection}(k)$

$$P_{rejection}(k) = \sum_{i=1}^{S_\pi} \pi(i) P_{rejection}(i, k). \quad (33)$$

则传输系统的平均拒绝表示为

$$\overline{Rej} = \sum_{i=0}^{L_b} P_{rejection}(i) i. \quad (34)$$

4) 平均时延指标

高触发特性业务在数据处理时存在复用情况,因此,计算时延不考虑触发特性数据包情况,定义时延指标为单个数据包从到达传输系统到传输所用时隙数

$$\overline{Delay} = \frac{\overline{Que} - \sum_{j=0}^{L_b} \sum_{b_i=j, i_z>\rho_z} \pi(i) j + \frac{E(n)}{2}}{\overline{Throughput}}, \quad (35)$$

式中:*n*是当前时隙到达业务的流量特性;*i_z*表示第*i*种系统状态下到达业务的触发特性状态; ρ_z 为数据处理时会被服务器保存的触发特性状态阈值。

3 反馈策略的数学模型

3.1 策略的原理

研究将所有时隙分为如图 5 所示的直传时隙和调整时隙。目前 MEC 服务器在直传时隙完成对传输过程

的评估,并进行数据包传输。在调整时隙,MEC 服务器需要根据观测值统计结果调整预测的 OBU 缓存分布。

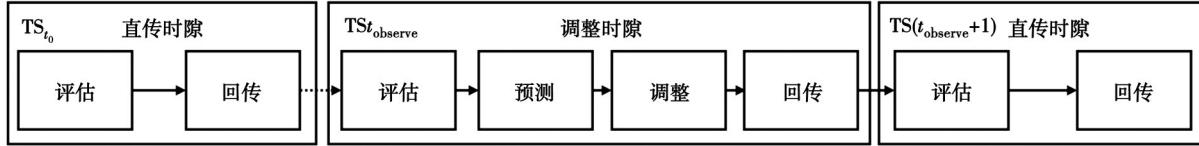


图 5 预测时隙图

Fig. 5 Prediction time slot plot

假设传输性能稳定所需的时间为 $t_{observe}$, 则 t_0 到 $t_{observe}-1$ 为直传时隙, 数据传输性能优于预设指标值, MEC 服务器直接将数据包传输至 OBU。 $t_{observe}$ 时隙为调整时隙, 数据传输性能劣于预设指标值, MEC 服务器根据观测时隙统计的排队情况对预测队长分布进行调整, 依据该分布发送数据包到 OBU。

3.2 两种反馈策略的数学模型

基于以上预测模型, 在每个观测时隙, MEC 服务器预测的车载端空余队长分布可表示为

$$\mathbf{P}_q^b = [P_q^b(q=0), \dots, P_q^b(q=i), \dots, P_q^b(q=L_b)], \quad (36)$$

式中: $P_q^b(q=i)$ 表示预测车载端可存入 i 个数据包的概率。结合反馈策略对预测空余队长分布进行修正, 可获取 MEC 服务器向 OBU 传输数据包数量遵循的概率分布向量

$$\mathbf{P}_m = [P_m(0), \dots, P_m(i), \dots, P_m(L_m)], \quad (37)$$

式中: $P_m(i)$ 表示 MEC 服务器向 OBU 传输 i 个数据包的概率。根据空余队长期望 $E_b = \sum_{i=0}^{L_b} P_q^b(i)i$, 可以建立传统的本地计算机制(all local computing mechanism, ALCM)、基于边缘缓存的全卸载计算机制(all offloading computing mechanism with caching, AOCM)、基于边缘缓存的 V2I 卸载机制(V2I collaborative caching and resource allocation, V2I-CCRA)^[5] 和提出的概率传输策略(probability transport strategy, PTS)数学模型。ALCM 策略需要将所有任务在车载端进行计算处理, AOCM 策略需要将所有任务都卸载到 OBU 进行计算。ALCM、AOCM、V2I-CCRA 和 PTS 策略可分别用 \mathbf{P}_m^1 、 \mathbf{P}_m^2 、 \mathbf{P}_m^3 和 \mathbf{P}_m^4 表示

$$P_m^1(i) = \begin{cases} 0, & i \neq 0, \\ 1, & i = 0; \end{cases} \quad (38)$$

$$P_m^2(i) = \begin{cases} 0, & i \neq L_m, \\ 1, & i = L_m; \end{cases} \quad (39)$$

$$P_m^3(i) = \begin{cases} 0, & i \neq E_b, \\ 1, & i = E_b; \end{cases} \quad (40)$$

$$P_m^4(i) = \begin{cases} \frac{0.2}{E_b}, & 0 \leq i < E_b, \\ 0.8, & i = E_b, \\ 0, & i > E_b. \end{cases} \quad (41)$$

4 数值分析

研究提出的评估框架最显著优点是可以灵活控制评估框架中的到达业务类型、信道环境、反馈策略、OBU 和 MEC 服务器的硬件设置。为了简化计算量, 对信道状态和所有业务特征都仅考虑 2 种到达状态, 对应状态矩阵也仅考虑 2 种: $\mathbf{H}_1 = \begin{bmatrix} 0.1 & 0.9 \\ 0.5 & 0.5 \end{bmatrix}$ 和 $\mathbf{H}_2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$ 。设定 OBU 处理时隙间隔为 $v_H = 1$ (时隙/业务), 服务器端处理时隙间隔为 $v_L = 2$ (时隙/业务), 到达业务流量特性、可靠实时特性和触发特性的忙时特性状态为 $w_n = w_l = w_z = 0$, 其他默认参数如表 1 所示。

表 1 默认参数

Table 1 Default parameters

符号	含义	参数
I	到达业务流量特性状态转移矩阵	H_2
ξ_1^a	数据量状态编号为 0 时的业务数据量概率分布	[0.2,0.6,0.2,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0]
ξ_2^a	数据量状态编号为 1 时的业务数据量概率分布	[0.0,0.0,0.0,0.0,0.0,0.0,0.2,0.25,0.3,0.2,0.05]
Γ	到达业务可靠实时特性状态转移矩阵	H_2
ξ_1^r	可靠实时特性状态编号为 0 时可靠实时特性概率分布	[0.1,0.9]
ξ_2^r	可靠实时特性状态编号为 1 时可靠实时特性概率分布	[0.9,0.1]
G	到达业务触发特性状态转移矩阵	H_2
ξ_1^t	到达触发特性状态编号为 0 时的触发特性概率分布	[0.3,0.7]
ξ_2^t	到达触发特性状态编号为 1 时的触发特性概率分布	[0.7,0.3]
F	信道状态转移矩阵	H_2
ξ_1^{channel}	信道状态编号为 0 时的服务器调制方案	[0.1,0.3,0.5,0.1,0,0,0]
ξ_2^{channel}	信道状态编号为 1 时的服务器调制方案	[0,0,0,0.3,0.35,0.25,0.1]

为了评估当前提出策略的性能,将其与 ALCM, AOCM 和 V2I-CCRA^[5]3 种策略进行性能比较。图 6 清晰地展示了反馈策略对传输性能的影响。设置参数如表 2 所示,考察大量、高可靠实时特性的不流行业务。设定 $F = \begin{bmatrix} p & 1-p \\ 0.5 & 0.5 \end{bmatrix}$, $p = 0, 0.1, \dots, 1$ 。从图中可以发现, V2I-CCRA 策略和提出的 PTS 策略的各项性能都比传统的 ALCM 和 AOCM 策略优越 50% 以上,而 PTS 策略和 V2I-CCRA 策略下的吞吐、拒绝指标则近似相当, PTS 策略下的时延指标比 V2I-CCRA 策略下优越 5%~30%。由于到达的重要不流行业务会在 OBU 端进行处理和传输,随着信道通信环境变差,传输系统的吞吐量变小,拒绝包数变多、系统时延变大。仿真结果证实了 PTS 策略的有效性,证明了本系统可以较明确地评估采用不同策略的传输系统在不同类型到达业务下的性能。

表 2 仿真 2 到达状态对应业务特性

Table 2 Characteristics of the arrival state in simulation 2

状态编号	I	w_n	Γ	w_f	G	w_z
到达状态 1	H_1	0	H_1	0	H_2	0
到达状态 2	H_1	0	H_2	0	H_2	0
到达状态 3	H_1	0	H_2	1	H_2	0
到达状态 4	H_1	1	H_2	1	H_2	0
到达状态 5	H_2	1	H_2	1	H_2	0
到达状态 6	H_2	1	H_2	1	H_2	1
到达状态 7	H_2	1	H_2	1	H_1	1

为了验证到达模型的合理性和系统对到达模型的兼容性,设置参数如表 3 所示,图 7 展示了当不同类型业务到达时,随着 OBU 容量变大,传输系统在给定反馈策略和卸载策略下的性能。当 OBU 容量逐渐变大时,由于不重要业务会优先存储在 MEC 服务器, OBU 吞吐量不变。由于反馈策略和 OBU 容量的限制, MEC 服务器反馈数据包的数量存在上限。当不重要业务到达量逐渐增多时, OBU 终端拒绝的数据包也会变多。时延指标则体现了系统对不重要业务的存储量存在最大限度。此外,到达状态 2~5 在所有性能上波动都不超过自身的 1%~2.5%,说明传输系统中可靠实时特性和流量特性变化对系统性能影响较小,在设计卸载策略和反馈策略时可以较少考虑。但相较到达状态 6 和 7 下的传输性能,差距达到自身的 25%~33%,可以认定流行特性的忙时特性对策略和传输系统的设计较为重要。

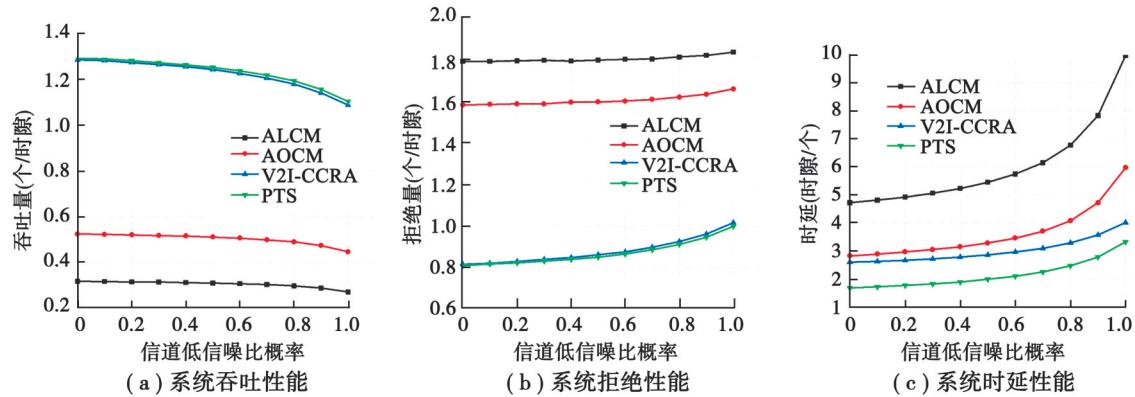


图 6 在信道低信噪比不同概率(p)下,采用 ALCM、AOCM、V2I-CCRA 和 PTS 的传输系统的传输性能的对比

Fig. 6 Comparison of the transmission performance of ALCM, AOCM, V2I-CCRA and PTS transport systems with different low SNR states of channels (p)

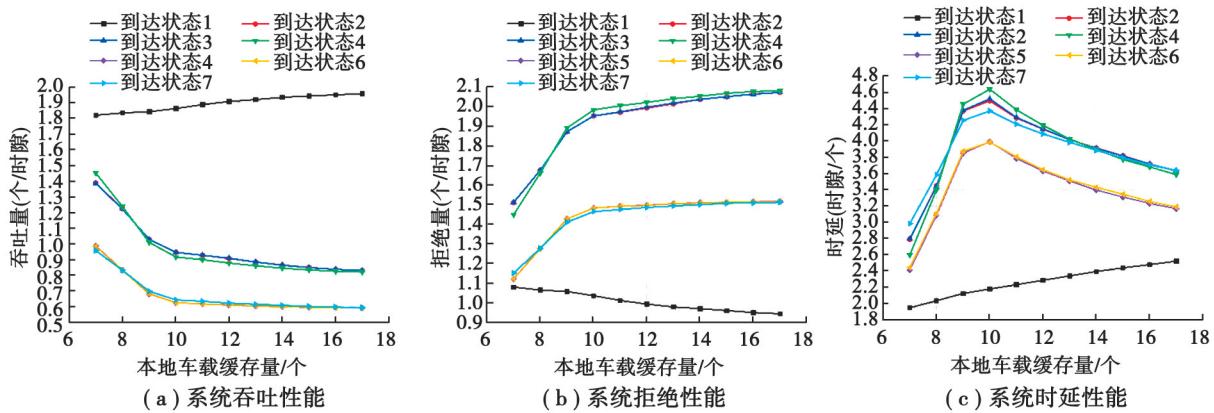


图 7 在不同 OBU 缓存量(L_b)下,采用 PTS 的传输系统的传输性能的对比

Fig. 7 Comparison of the transmission performance of a transmission system using PTS under different OBU buffer amounts (L_b)

为了探究服务器和 OBU 处理器速率对系统性能的影响,在图 8 中逐渐降低 OBU 传输速率,并控制服务器的处理速率比 OBU 处理速率低一个时隙,考虑到达业务到达状态如表 3 所示。如图 8 所示,随着 OBU 和服务器的处理速率降低,系统的传输性能普遍降低为初始性能的 80% 左右。但在到达状态 7 下系统的性能一直保持较好的水平且波动不超过最好状态的 15%,说明当到达业务具备量少、触发特性和忙时特性时,可以采用处理性能较差的传输系统。

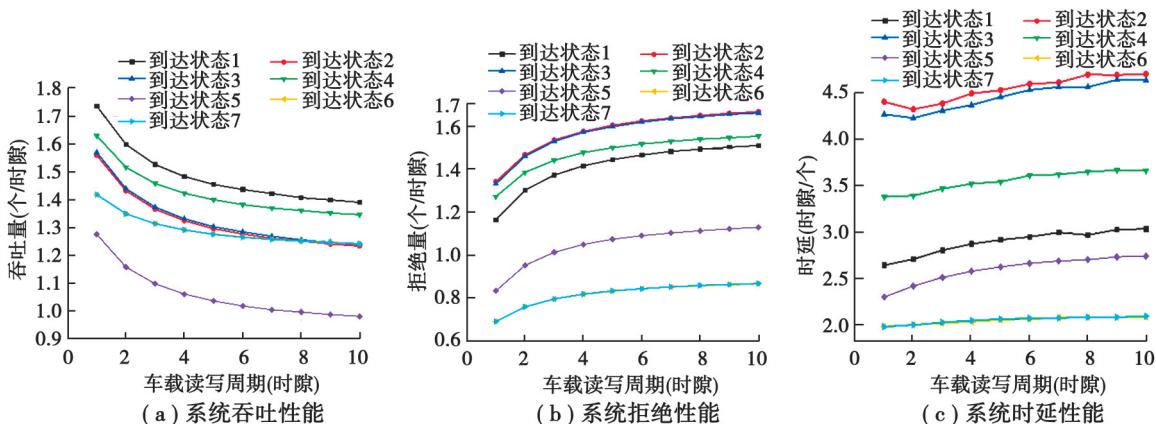


图 8 在不同 OBC 传输速率下,采用 PTS 的传输系统的传输性能的对比

Fig. 8 Comparison of the transmission performance of a transmission system using PTS under different types of arrival services at different OBC transmission rates

表3 仿真3业务到达状态对应业务特性

Table 3 Characteristics of the arrival state in simulation 3

状态编号	I	w_n	Γ	w_l	G	w_z
到达状态1	H_1	0	H_1	0	H_2	0
到达状态2	H_1	0	H_1	0	H_1	0
到达状态3	H_1	0	H_1	0	H_1	1
到达状态4	H_1	1	H_1	0	H_1	1
到达状态5	H_2	1	H_1	0	H_1	1
到达状态6	H_2	1	H_2	0	H_1	1
到达状态7	H_2	1	H_2	1	H_1	1

综上所述,所提的框架可以在评估服务器到端的传输性能时灵活调控卸载反馈策略、传输环境等多种参数,对到达业务的各项特性量化效果良好,为设计卸载反馈策略、评估传输性能提供参考。

5 结论

针对车联网中业务类型的爆发式增长和时变特性,在一定成本限制下使卸载反馈策略的设计更能满足传输需求,笔者提出了时变多类型业务下的通用缓存调度分析框架。该框架是基于马尔可夫链蒙特卡罗的马尔可夫调制服务过程模型,实现了业务类型在量和时变程度上的极大可配置性,针对不同的卸载反馈策略提供传输性能指标,为更好设计卸载反馈策略提供参考。基于该框架,研究提出一种卸载反馈策略,并探究了不同卸载反馈策略、不同处理器硬件配置和到达不同类型业务在传输性能上的相关性,指出卸载反馈策略以及不同的处理器配置适合处理业务的类型。实验结果表明,该框架可以在考虑多类型时变业务情况下,适应不同的卸载反馈策略并量化处理器速率以呈现卸载反馈策略和处理器速率对传输系统的影响,为选取处理器和设计卸载反馈策略提供参考。研究可以为建立车联行为的综合分析模型提供参考,为6 G多传感器通感一体化前景提供仿真模型。但目前由于计算复杂度,框架仅能考虑一对边缘服务器和车辆,后续研究以降低框架复杂度、建立车联网网络模型作为研究目标。

参考文献

- [1] 黄永明,郑冲,张征明,等.大规模无线通信网络移动边缘计算和缓存研究[J].通信学报,2021,42(4): 44-61.
Huang Y M, Zheng C, Zhang Z M, et al. Research on mobile edge computing and caching in massive wireless communication network[J]. Journal on Communications, 2021, 42(4): 44-61.
- [2] Xu J, Chen L, Zhou P. Joint service caching and task offloading for mobile edge computing in dense networks[C]//IEEE INFOCOM 2018 Conference on Computer Communications. Honolulu: IEEE, 2018: 207-215.
- [3] Zhang S, Li J, Luo H, et al. Low-latency and fresh content provision in information-centric vehicular networks[J]. IEEE Transactions on Mobile Computing, 2022, 21(5): 1723-1738.
- [4] 张建军,代帅康,张本宏.车联网中基于任务紧急性的联合卸载方案[J].电子测量与仪器学报,2020,34(11): 66-71.
Zhang J J, Dai S K, Zhang B H. Joint offloading method based on task urgency in the VANETs[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(11): 66-71. (in Chinese)
- [5] 李方伟,张海波,王子心.车联网中基于MEC的V2X协同缓存和资源分配[J].通信学报,2021,42(2): 26-36.
Li F W, Zhang H B, Wang Z X, V2X collaborative caching and resource allocation in MEC-based IoV[J]. Journal on Communications, 2021, 42(2): 26-36. (in Chinese)
- [6] Liu Y, Yu H, Xie S, et al. Deep reinforcement learning for offloading and resource allocation in aehicle edge computing and networks[J]. IEEE Transactions on Vehicular Technology, 2019, 68(11): 11158-11168. (in Chinese)
- [7] 董振江,古永承,梁健,等.C-V2X车联网关键技术与方案概述[J].电信科学,2020,36(4): 3-14.
Dong Z J, Gu Y C, Liang J, et al. Overview on key technology and solution of C-V2X for internet of vehicles[J]. Telecommunications Science, 2020, 36(4): 3-14. (in Chinese)

- [8] 王海陶, 宋小明, 卢纪宇. 物联网业务特征及业务模型研究[J]. 广西通信技术, 2012(3): 43-49.
Wang H T, Song X M, Lu J Y. The research on characteristics and service model of internet of things[J]. Guangxi Communication Technology, 2012(3): 43-49. (in Chinese)
- [9] 程一凡, 曲至诚, 张更新. 低轨卫星星座物联网业务量建模[J]. 电子与信息学报, 2021, 43(4): 1050-1056.
Cheng Y F, QU Z C, Zhang G X. Traffic modeling for low earth orbit satellite constellation internet of things[J]. Journal of Electronics and Information Technology, 2021, 43(4): 1050-1056. (in Chinese)
- [10] 张海霞, 李腆腆, 李东阳, 等. 基于车辆行为分析的智能车联网关键技术研究[J]. 电子与信息学报, 2020, 42(1): 36-49.
Zhang H X, Li D D, Li D Y, et al. Research on vehicle behavior analysis based technologies for intelligent vehicular networks [J]. Journal of Electronics and Information Technology, 2020, 42(1): 36-49. (in Chinese)
- [11] 侯世武, 谭献海. 典型物联网业务流量特性研究分析[J]. 物联网技术, 2017, 7(6): 40-42, 46.
Hou S W, Tan X H, Research and analysis of traffic characteristics of typical IoT services[J]. Internet of Things Technologies, 2017, 7(6): 40-42, 46. (in Chinese)
- [12] Zhu Y, Sheng M, Li J. Modeling and performance analysis for satellite data relay networks using two-dimensional markov-modulated process[J]. IEEE Transactions on Wireless Communications, 2020, 19(6): 3894-3907.
- [13] Silva L, Magaia N, Sousa B, et al. Computing paradigms in emerging vehicular environments: a review[J]. IEEE/CAA Journal of Automatica Sinica, 2021, 8(3): 491-511.
- [14] Peng K, Nie J, Kumar N, et al. Joint optimization of service chain caching and task offloading in mobile edge computing[J]. Applied Soft Computing, 2021, 103: 107142.
- [15] 戚艾林, 李旭杰, 陆睦, 等. 基于遗传算法的5G车联网的数据协作分发策略研究[J]. 国外电子测量技术, 2019, 38(01): 33-37.
Qi A L, Li X J, Lu M, et al. Research on data cooperative distribution strategies of 5G networked vehicles based on genetic algorithms[J]. Foreign Electronic Measurement Technology, 2019, 38(01): 33-37. (in Chinese)
- [16] 耿珂, 胡坤, 高强, 等. 复杂环境下双向协作通信研究[J]. 电子测量技术, 2019, 42(01): 116-120.
Ke J, Hu K, Gao Q, et al. Research on bidirectional cooperative communication in complex environment[J]. Electronic Measurement Technology, 2019, 42(01): 116-120. (in Chinese)
- [17] Zhang M, Zhu X, Zhang B, et al. A cross-layer performance evaluation system for spectrum sensing and allocation strategies in CR-WSN[J]. IEEE Sensors Journal, 2024, 24(9): 15355-15366.
- [18] Metzger F, Hoßfeld T, Bauer A, et al. Modeling of aggregated IoT traffic and its application to an IoT cloud[J]. Proceedings of the IEEE, 2019, 107(4): 679-694.
- [19] Zhang M, Zhu X Y, Wang S, et al. A channel allocation framework under responsive pricing in heterogeneous cognitive radio network[J]. IEEE Transactions on Cognitive Communications and Networking, 2023, 9(4): 872-883.
- [20] Wang S, Lan H, Zhu X, et al. A performance evaluation system of channel allocation protocol based on probability vectors for cognitive radio network[C]//2021 IEEE 4th International Conference on Electronics Technology (ICET). Chengdu, China: IEEE, 2021: 1062-1067.
- [21] 罗峰, 马逸飞, 郭怡, 等. 车载时间敏感网络链路冗余调度性能分析[J]. 仪器仪表学报, 2023, 44(02): 278-287.
Luo F, Ma Y F, Guo Y, et al. Analysis of time-sensitive network link redundancy scheduling performance[J]. Chinese Journal of Scientific Instrument, 2023, 44(02): 278-287. (in Chinese)

(编辑 侯湘)