DOI:10.20079/j.issn.1001-893x.220419009
引用格式:陈昊宇,胡宏林.雾网络中基于统计分布的内容缓存与交付方案[J].电讯技术,2023,63(12):1902-1910.[CHEN H Y,HU H L. Content caching and delivery policy with frequency distribution in fog radio access networks[J]. Telecommunication Engineering,2023,63(12):1902-1910.]

雾网络中基于统计分布的内容缓存与交付方案*

陈昊宇[1,2],胡宏林[1]

(1.中国科学院上海高等研究院,上海 201210;2.中国科学院大学 电子电气与通信工程学院,北京 100049)

摘 要:作为 5G 中的一种重要模型,雾无线接入网络(Fog Radio Access Network,F-RAN)通过设备到设备通信和无线中继等技术获得了显著的性能增益,而边缘设备中合适的缓存则可以让内容缓存用户(Caching Users,CUs)向内容请求用户(Requesting Users,RUs)直接发送缓存内容,有效减小前传链路的负担和下载延迟。考虑一个 F-RAN 模型下用户发出请求并获得交付的场景,将每个 CU 的内容请求队列建模为独立的 M/D/1 模型,分析导出 CUs 缓存命中率和平均下载延迟关于内容缓存与交付方案的表达式,证明 CUs 缓存命中率与内容统计分布之间的联系有助于实现前者的近似最优解。针对在一段时间内的期望视角下建立的优化问题,提出了基于统计分布的算法并注意了执行时的交付控制。仿真结果表明,相较于现有缓存策略,优化内容整体统计分布的方案能够最大化 CUs 缓存命中率,同时减小平均下载延迟。

关键词:5G;雾无线接入网络(F-RAN);雾计算;终端直通(D2D);内容缓存

开放科学(资源服务)标识码(OSID):

中图分类号:TN915.07 **文献标志码**:A **文章编号**:1001-893X(2023)12-1902-09

Content Caching and Delivery Policy with Frequency Distribution in Fog Radio Access Networks

CHEN Haoyu[1,2], HU Honglin[1]

(1. Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China;
2. School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: As an important model in 5G, the Fog Radio Access Network (F-RAN) achieves significant performance gains through technique among device-to-device (D2D) communications and wireless relays. Especially, caching appropriately in the edge devices allows content caching users (CUs) to send content to content requesting users (RUs) directly, which effectively reduces the burden of the fronthaul and download delay. The authors consider a scenario where users send content requests and get delivered under the F-RAN model. By modeling the content request queue at each CU as an independent M/D/1 queue model, the authors analyze and derive the cache hit probability of CUs and average download delay expressions under content caching and delivery policies. It is proved that the relationship between the cache hit probability of CUs and content frequency distribution can help realizing the approximate optimal solution of the former. On this basis, the authors establish an optimization problem under the expectation perspective over some time and propose a frequency distribution (FD)-based algorithm with delivery control when implementing to solve it. The simulation results show that compared with the existing caching policies, the policy optimizing the FD of all content can maximize the cache hit probability of CUs and reduce the average download delay.

Key words: 5G; fog radio access network (F-RAN); fog computing; device-to-device (D2D); content caching

* **收稿日期**:2022-04-19 **修回日期**:2022-06-10
基金项目:国家重点研发计划(2020YFB1806606);国家自然科学基金资助项目(61801460)
通信作者:胡宏林

0　引　言

随着智能移动设备(如智能手机、平板电脑、可穿戴设备等)的发展和普及,用户对低延迟移动应用程序和多媒体服务的需求大幅增加,这导致了数据流量的爆炸性增长,对未来网络设计形成了挑战[1]。文献[2]中预测:到 2023 年,连接到 IP 网络的设备数量将超过全球人口的 3 倍,其中物联网(Internet of Things,IoT)连接将占全球在线设备和连接数量的一半,5G 设备和连接则占据 10% 以上。通过整合云计算和无线接入网络,云无线接入网络(Cloud Radio Access Network,C-RAN)为满足无线通信系统发展带来的巨大算力与带宽需求提供了可能性[3]。然而,在 C-RAN 正面临着前传容量受限等一系列挑战的同时[4],集中式云计算能力的线性增速已逐渐与边缘数据的需求拉开差距[5]。为了解决上述问题,进一步整合云计算和雾计算则产生了新的雾无线接入网络(Fog Radio Access Network,F-RAN)架构[6]。它将大量存储、通信和控制功能转移,在网络边缘扩展了传统的云计算范式[7]。在 F-RAN 中,缓存用户(Caching Users,CUs)和雾无线接入节点(Fog Radio Access Point,F-AP)组成了容量有限的内容服务器,直接向请求用户(Requesting Users,RUs)交付已被缓存的内容。通过避免内容重复传输以及缩小与用户间的物理距离,F-RAN 可以节省大量核心网络和回程传输的资源消耗,有效减少用户的服务延迟,进而提高用户体验质量(Quality of Experience,QoE)[1,8]。

内容分发网络(Content Delivery Network,CDN)和信息中心网络(Information Center Network,ICN)等架构对缓存技术的成功实践证明了在移动网络中部署内容缓存的合理性与可行性[5]。然而,一方面互联网与移动网络内容缓存机制的巨大差异使得传统基于 CDN 的缓存技术在移动网络中无法直接应用;另一方面,与一般移动边缘计算(Mobile Edge Computing,MEC)网络相比,F-RAN 能够凭借更强的算力同时处理更多的数据[9]。因此,充分利用 F-AP 和 CUs 等边缘设备的缓存、计算和通信能力,方便 RUs 快速访问和检索内容,对于减轻前传、后传甚至云端和主网的流量负载,有效节约网络带宽和能耗具有重要意义[6,10]。作为提高 F-RAN 性能的关键组件,如何缓存有使用价值的数据几乎决定了不同终端的服务质量(Quality of Service,QoS)能否得到保证,这导致内容缓存策略的优化成为了 F-RAN 边缘缓存研究的热点[11-12]。

关于 F-AP 处的缓存策略,目前已经有丰富的研究[13-15],主要考虑了用户和缓存内容本身的各种特性。然而,这些 F-RAN 边缘缓存场景中并未考虑 D2D 技术的应用。事实上,CUs 可以与邻近的 RUs 共享缓存数据,利用 D2D 链路直接向 RUs 提供服务。该技术对于 F-RAN 在延迟和传输成本等内容交付方面的显著性能提升已经得到了研究和验证[11,16]。

上述关于 F-RAN 环境下 F-AP 与 D2D 设备合作缓存的研究主要从各个节点的角度考虑它们与内容集的具体映射关系,这相当程度上削弱了所有 D2D 设备形成的缓存空间的整体性。然而,若因系统随机性而采取一段时间内的期望视角,在 CUs 具备显著内容交付优势的前提下,直接优化 CUs 缓存命中率是必要的。因此,本文主要以一个包含单个 F-AP 和多个 D2D 设备的 F-RAN 系统为研究对象,通过数学证明探究有关建模与内容统计分布之间的联系,从而导出优化目标并设计关于一般优化问题的求解算法。针对边缘存储,本文提出了基于统计分布的内容缓存与交付方案,实现了 CUs 缓存命中率的近似全局最优解,同时验证了相应交付方案对于优化平均下载延迟的有效性。

1　F-RAN 系统模型

1.1　网络模型

本文考虑一个包含多个 D2D 设备的典型 F-RAN 系统,如图 1 所示,其中共包含了一个功能完善的云计算中心(Cloud Computing Center,CCC),一个 F-AP,多个 CUs 和多个 RUs。在该网络结构下,F-AP 位于其覆盖区域中心,可以通过前传链路传输并缓存来自 CCC 的内容,也可以通过无线链路直接服务这一区域内的所有 CUs 和 RUs。

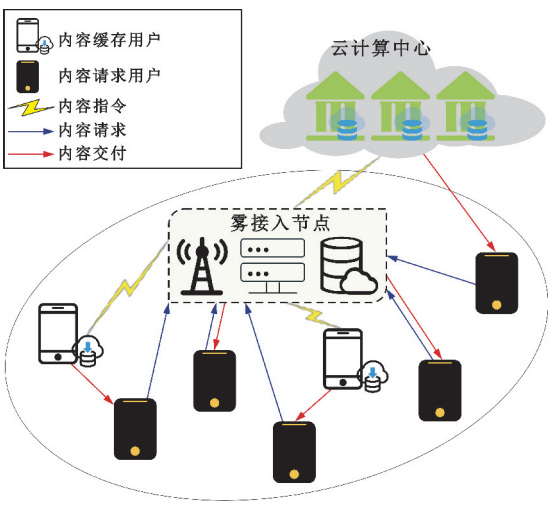

图 1　F-RAN 系统模型

假设 F-AP 覆盖区域的半径为 R_F,CUs 的最大传输距离为 R_C,且其分布服从密度为 γ_C 的齐次泊松点过程(Poisson Point Process,PPP)Φ_C;RUs 的分布服从密度为 γ_R 的独立 PPP Φ_R[16],因此 F-AP 和 CUs 的最大服务范围分别为 $A_F=\pi R_F^2$ 和 $A_C=\pi R_C^2$。对于不同范围下的 CUs 和 RUs 数量,本文采用 PPP 在指定区域中随机撒点个数的期望来进行表示,即 F-AP 可服务的 CUs 和 RUs 数量分别为 $K_{CF}=\gamma_C A_F$ 和 $K_{RF}=\gamma_R A_F$,CUs 最大通信范围内的 CUs 和 RUs 数量分别为 $K_{CC}=\gamma_C A_C$ 和 $K_{RC}=\gamma_R A_C$。

1.2 内容缓存与用户请求模型

本文将内容的流行程度视为用户请求相应内容的概率,并假定该信息可以由 CCC 端基于大量用户请求数据定期获取。当缓存空间有限时,设 CCC 服务器存储了一个大小为 N 的总体内容库 $C=\{c_1,c_2,c_3,\cdots,c_N\}$,其中 $c_i(i\in\{1,2,\cdots,N\})$ 表示所有内容中第 i 流行的内容文件。随着 RUs 实际请求的发生,该库及其包含的流行度信息将定期更新,下一步 F-AP 和 CUs 即可根据预先制定的缓存策略更新它们存储的内容集,以便 RUs 在后续的一段时间内通过不同链路直接取得已被缓存的内容。设 F-AP 本地存储空间大小为 N_F,执行最流行内容缓存(Most Popular Content Cache,MPC)策略[17],缓存了 CCC 端最流行的 N_F 个内容,则 F-AP 缓存的内容集为 $C_F=\{c_1,c_2,c_3,\cdots,c_{N_F}\}$。此外,假设单个 CU 本地存储空间的大小为 N_C,其缓存的内容集 C_C 为 C_F 的一个子集,由 F-AP 经无线链路直接传输获得。综上所述,显然有 $N>N_F>N_C$。CUs 内容集的形成方法即为本文主要讨论的内容缓存方案。

针对用户请求模型,本文规定 F-AP 覆盖区域内所有 RUs 都首先将它们的内容请求发送给 F-AP。不失一般性,令 CCC 服务器所存储的所有内容都有相同的大小 s,且用户请求内容 c_i 的概率服从 Zipf 分布,则 p_i 有表达式如式(1)所示:

$$p_i=\frac{i^{-\beta}}{\sum_{j=1}^{N}j^{-\beta}}。\tag{1}$$

式中:β 是控制受欢迎内容集中程度的 Zipf 指数[18],当 β 的值逐渐增大时,越来越多的用户请求会向越来越少的高流行度文件集中[19]。

1.3 内容交付模型

当 F-AP 接收到 RUs 的内容请求以后,系统将处理这些请求,然后交付相应内容,并于期间产生一定的下载延迟。此处有关下载延迟的具体定义为 F-AP 接收到内容请求至请求内容完成交付之间所经历的时间。假设每个 RU 请求到达都遵循独立泊松过程,且有相同的强度(到达率)λ(请求数/秒),即每个 RU 平均每秒发出 λ 个请求。对于不同 RU 的不同请求,F-AP 接收后有 3 种服务模式可供选择。

1.3.1 CCC 模式

当 RUs 请求的内容 $c_i\notin C_F$ 时,F-AP 就会通过前传链路向 CCC 端转发该信息。收到指令之后,CCC 端则可通过高功率节点(High Power Nodes,HPNs)直接向 RUs 提供此类流行度较低的突发性内容[6]。设 CCC 模式经历的下载延迟为固定常数 t。

1.3.2 F-AP 模式

当 RUs 请求的内容 $c_i\in C_F$ 时,若满足下列条件之一,则请求的内容由 F-AP 直接交付给 RUs:

1)在距离某个 RU 不超过 R_C 的范围内没有 CUs 存在,或存在若干个 CUs 但都未缓存内容 c_i 时;

2)在距离某个 RU 不超过 R_C 的范围内,存在一个缓存了内容 c_i 的 CU,但 F-AP 按照一定的概率判断其交付内容所需的时延过长,不适宜为该 RU 提供服务时。

本文假设 F-AP 有足够强的服务能力,可以同时传输大量内容,因此对于随机到达的 RUs 请求,F-AP 不会形成内容请求队列。令 F-AP 到 RUs 的数据传输速率为 r_F,则该模式经历的下载延迟为 c_i 从 F-AP 传输到相应 RU 所需要的时间,即 $t_F=s\cdot r_F^{-1}$。

1.3.3 CUs 模式

当 RUs 请求的内容 $c_i\in C_F$ 时,若在距某个 RU 不超过 R_C 的范围内存在一个缓存了 c_i 的 CU,且 F-AP 按照一定的概率判断其适合提供服务,则 c_i 由该 CU 交付给该 RU,具体分为两步:

1)F-AP 首先识别出适合的 CU,向其发送一条包含内容请求信息的指令,通知它准备交付内容;

2)收到指令后,CU 再通过 D2D 链路向发出请求的 RU 交付内容。

本文假设 CUs 服务能力有限,最多同时传输一个内容,因此对于 RUs 随机到达的请求,每个 CU 都将形成一个内容请求队列。故该模式经历的下载延迟包括内容请求在队列中的等待时间及其所需的服务时间(此处服务时间由指令从 F-AP 传输到相应 CU 和 c_i 从指定 CU 传输到相应 RU 两部分时延组成)。令 F-AP 发送指令部分的时延为 t_{FC},CUs 和 RUs 间的数据传输速率为 r_C,队列服务时间则可被表示为 $t_C=t_{FC}+s\cdot r_C^{-1}$。本文考虑把每个 CU 的内容请求队列都建模成独立的 M/D/1 排队模型,即请

求到达服从泊松过程且服务时间固定[20]。

综上所述,包括 F-AP 判断某个 CU 是否适合提供服务的概率的具体计算方法在内,完整的模式选择方法即为本文主要讨论的内容交付方案。

2　内容缓存与交付方案性能分析

2.1　缓存命中率

本文选取了缓存命中率作为分析方案性能的一项指标,用 P_{HC} 表示。它的定义为任意一个 RU 所请求的内容在其最大可通信范围内的某个 CU 处已被缓存的概率。因此,关于 CUs 缓存命中率的讨论与内容交付模型 CUs 模式下 F-AP 对相应 CU 是否适合提供服务的判断无关。同时,较大的 P_{HC} 意味着 CUs 可以通过 D2D 链路满足更多的 RUs 内容请求。下面,本文将从内容统计分布的角度导出 P_{HC} 的表达式。

假设缓存了内容 c_i 的 CUs 在所有 CUs 中的占比为 δ_i,$\delta_i\in[0,1]$,且任意一个 RU 位于一个半径为 R_{C} 的圆形区域的中心。故缓存了 c_i 的 CUs 在 F-AP 最大服务范围内的分布服从密度为 $\delta_i\gamma_{\mathrm{C}}$ 的齐次 PPP,根据二维泊松过程的性质,区域 A_{C} 内存在 m 个此类 CUs 的概率如式(2)所示:

$$P_i(m)=\frac{(\delta_i\gamma_{\mathrm{C}}A_{\mathrm{C}})^m\mathrm{e}^{-\delta_i\gamma_{\mathrm{C}}A_{\mathrm{C}}}}{m!}。\tag{2}$$

若令 $m=0$,则式(2)给出了区域 A_{C} 内不存在相应 CUs 的概率[21],即缓存未命中概率,那么至少有一个缓存了 c_i 的 CUs 位于 A_{C} 内的概率为 $1-P_i(m=0)$。此外,由于 RUs 内容请求的概率分布一致且只与内容本身有关,当任意一个 RU 请求内容 c_i 的概率都为 p_i 时,仅考虑 c_i 的缓存命中率如式(3)所示:

$$P_{\mathrm{HC},i}=p_i[1-P_i(m=0)]=p_i(1-\mathrm{e}^{-\delta_iK_{\mathrm{CC}}})。\tag{3}$$

在放置 CUs 缓存的过程中,各个内容之间不涉及相关性。结合式(3)及上述定义,若要计算 CUs 能够交付任意一个请求内容的概率,可以对所有 $c_i\in C_{\mathrm{F}}$ 的 $P_{\mathrm{HC},i}$ 求和。CUs 缓存命中率如式(4)所示:

$$P_{\mathrm{HC}}=\sum_{i=1}^{N_{\mathrm{F}}}P_{\mathrm{HC},i}=\sum_{i=1}^{N_{\mathrm{F}}}p_i(1-\mathrm{e}^{-\delta_iK_{\mathrm{CC}}})。\tag{4}$$

同时易得,δ_i 关于 CUs 本地存储空间大小的约束如式(5)所示:

$$\sum_{i=1}^{N_{\mathrm{F}}}\delta_i=N_{\mathrm{C}}。\tag{5}$$

2.2　平均下载延迟

本文还选取平均下载延迟作为分析内容缓存与交付方案性能的另一项指标,用 D_{AC} 表示。它的定义为所有 RUs 长期经历的平均下载延迟。基于 F-RAN 的 3 种服务模式,本文需要对不同的内容交付链路进行讨论。

首先是 CUs 的 M/D/1 排队模型,这里面包括两个参数:平均请求到达率和平均服务率(队列服务时间的倒数)。假设 G_{CU} 表示所有 CUs 组成的集合,G_{RU} 表示所有 RUs 组成的集合。对于任意一个 RU$x\in G_{\mathrm{RU}}$,到达任意一个 CU$y\in G_{\mathrm{CU}}$ 的内容请求遵循强度为 $\lambda_y(x)$ 的泊松过程。参数 $\lambda_y(x)$ 的表达式如式(6)所示:

$$\lambda_y(x)=\lambda\eta_y\sum_{c_i\in C_{\mathrm{F}}}g_{y,i}(x)p_i。\tag{6}$$

式中:$\eta_y\in(0,1]$ 为 CUy 的交付系数,它表示 F-AP 判断 CUy 适合为 RUs 提供服务的概率;$g_{y,i}(x)$ 为指示函数。若 $g_{y,i}(x)=1$ 则表示 CUy 是最接近 RUx 的缓存了内容 c_i 的一个 CU,且 CUy 和 RUx 之间的距离不大于 R_{C};否则 $g_{y,i}(x)=0$。

因此,对于 CUy 处的泊松过程,它的内容请求总平均到达率 λ_y 的表达式如式(7)所示:

$$\lambda_y=\sum_{x\in G_{\mathrm{RU}}}\lambda_y(x)。\tag{7}$$

根据 M/D/1 排队模型的性质[22]和 CUs 处的平均服务率,CUy 处的流量负载可以用系统中的平均内容请求总数 L_y 来表示。参数 L_y 的表达式如式(8)所示:

$$L_y=\lambda_yt_{\mathrm{C}}+\frac{(\lambda_yt_{\mathrm{C}})^2}{2(1-\lambda_yt_{\mathrm{C}})}。\tag{8}$$

其次是 F-AP 模式,它虽然并不涉及排队模型或队列等待时间,但是不妨同样利用平均请求到达率和平均服务率两个参数加以理解(服务时间部分仍然成立)。对于 F-AP 处的泊松过程,内容请求到达的总平均速率为其能直接交付的所有 RUs 请求除去已通过 CUs 模式交付的部分。结合平均服务率计算,F-AP 处的平均内容请求总数(即流量负载)始终为下载延迟期间内容请求到达的总量,故参数 L_{F} 的表达式如式(9)所示:

$$L_{\mathrm{F}}=t_{\mathrm{F}}(\lambda K_{\mathrm{RF}}\sum_{i=1}^{N_{\mathrm{F}}}p_i-\sum_{y\in G_{\mathrm{CU}}}\lambda_y)。\tag{9}$$

最后,考虑到 CCC 模式的下载延迟,CCC 处平均内容请求总数 L 的表达式同理可得如式(10)所示:

$$L=t(1-\sum_{i=1}^{N_{\mathrm{F}}}p_i)\lambda K_{\mathrm{RF}}。\tag{10}$$

综合上述 3 种模式下的流量负载和排队理论的利特尔法则(Little's Law),上述模型有传输延迟方程[23]:系统长期平均请求数=长期平均请求到达

率×系统长期平均请求等待时间。方程中有两项已知,即 3 种服务系统的总平均内容请求数,以及所有 RUs 请求的平均总到达率,而系统等待时间已经包括仅 CUs 模式涉及的队列等待时间和 3 种模式都涉及的服务时间两部分。因此,所有 RUs 内容请求经历的平均下载延迟如式(11)所示:

$$D_{\mathrm{AC}}=\frac{1}{\lambda K_{\mathrm{RF}}}(L+L_{\mathrm{F}}+\sum_{y\in G_{\mathrm{CU}}}L_y)。\tag{11}$$

3 基于统计分布的内容缓存与交付方案

本文对 F-RAN 中内容缓存与交付问题进行建模,导出了方案评价指标的量化表达式,目的是寻找关键影响因素以最大化 CUs 缓存命中率,并验证优化结果对降低平均下载延迟的有效性。为了求解该模型引出的一般优化问题,本文将提出基于统计分布的内容缓存与交付方案,尽可能地提升内容缓存与交付方案的性能和用户体验。

从 2.1 节中 P_{HC} 表达式导出的过程和结果可知,影响其大小的因素只有各个 c_i 对应的 δ_i。因此,为了最大化 CUs 缓存命中率,优化结果仅需考虑内容的整体统计分布,而与 C_{F} 和 G_{CU} 之间的映射关系无关。记缓存策略矢量 $\boldsymbol{\Delta}=(\delta_1,\delta_2,\delta_3,\cdots,\delta_{N_{\mathrm{F}}})$。本文所提方案的思路如下:通过计算求出最优 $\boldsymbol{\Delta}$ 使 P_{HC} 达到最大,再按定义将不同内容随机放置于一定数量的 CUs 本地存储空间以形成所有 C_{C}。基于统计分布的内容缓存与交付方案最优缓存命中率问题的最小化表述如式(12)所示:

$$\min_{\delta_i}\quad -P_{\mathrm{HC}}\tag{12a}$$

$$\delta_i\in[0,1],i=1,2,\cdots,N_{\mathrm{F}},\tag{12b}$$

$$\sum_{i=1}^{N_{\mathrm{F}}}\delta_i=N_{\mathrm{C}},\tag{12c}$$

$$t_{\mathrm{C}}\lambda_y=1-\xi,y\in\boldsymbol{G}_{\mathrm{CU}}。\tag{12d}$$

式(12b)表示内容分布的统计结果是 0~1 之间的比例值;式(12c)表示实际缓存要完全占用 CUs 的本地存储空间;式(12d)中 ξ 为一较小正数,用于使得 M/D/1 排队模型的服务强度(即服务能力利用率:平均到达率与平均服务率的比值)小于 1,以满足排队系统的稳定条件,保证不会形成无限队列。此处假定 F-AP 已经掌握了 $N_{\mathrm{C}},K_{\mathrm{CC}},\beta,\lambda,t_{\mathrm{C}}$ 等必要信息。记式(12a)中目标函数 $-P_{\mathrm{HC}}=W(\boldsymbol{\Delta})$,且一般而言 K_{CF} 足够大,足以将 δ_i 视作连续变量,故容易求得函数 W 的一阶梯度和 Hessian 矩阵如式(13)和式(14)所示:

$$\nabla_{\Delta}W=-\begin{pmatrix}p_1K_{\mathrm{CC}}\mathrm{e}^{-\delta_1K_{\mathrm{CC}}}\\p_2K_{\mathrm{CC}}\mathrm{e}^{-\delta_2K_{\mathrm{CC}}}\\\vdots\\p_{N_{\mathrm{F}}}K_{\mathrm{CC}}\mathrm{e}^{-\delta_{N_{\mathrm{F}}}K_{\mathrm{CC}}}\end{pmatrix},\tag{13}$$

$$\nabla_{\Delta}^2W=\begin{pmatrix}\dfrac{p_1K_{\mathrm{CC}}^2}{\mathrm{e}^{\delta_1K_{\mathrm{CC}}}}&0&\cdots&0\\0&\dfrac{p_2K_{\mathrm{CC}}^2}{\mathrm{e}^{\delta_2K_{\mathrm{CC}}}}&\cdots&0\\\vdots&\vdots&&\vdots\\0&0&\cdots&\dfrac{p_{N_{\mathrm{F}}}K_{\mathrm{CC}}^2}{\mathrm{e}^{\delta_{N_{\mathrm{F}}}K_{\mathrm{CC}}}}\end{pmatrix}。\tag{14}$$

由于 ∇_{Δ}^2W 为正定矩阵,且式(12b)和式(12c)均为关于 $\boldsymbol{\Delta}$ 的仿射函数,也是凸函数,故式(12)符合凸问题的标准形式[24]。对于此类具有等式和不等式约束的一般优化问题,本文采用带 KKT(Karush-Kuhn-Tucker)条件的拉格朗日乘数法(Lagrange Multipliers)求解,求得的 $\boldsymbol{\Delta}$ 即为全局最优解。定义拉格朗日函数 Q 的表达式如式(15)所示:

$$Q(\boldsymbol{\Delta},\boldsymbol{\mu},\tau,\lambda)=W(\boldsymbol{\Delta})-\boldsymbol{\mu}\boldsymbol{\Delta}^{\mathrm{T}}-\boldsymbol{\mu}(1-\boldsymbol{\Delta}^{\mathrm{T}})+\theta(\|\boldsymbol{\Delta}\|_1-N_{\mathrm{C}})。\tag{15}$$

式中:不等式的约束矢量 $\boldsymbol{\mu}=(\mu_1,\mu_2,\mu_3,\cdots,\mu_{N_{\mathrm{F}}})$ 和 $\tau=(\tau_1,\tau_2,\tau_3,\cdots,\tau_{N_{\mathrm{F}}})$ 各个分量均为非负,分别表示限制 $\boldsymbol{\Delta}$ 中各个分量不小于 0 和不大于 1 所对应的约束系数;非零参数 θ 则为式(12c)对应的等式约束系数。

进一步地,上述一般优化问题还需要的求解条件如式(16)所示:

$$\begin{cases}\|\nabla_{\Delta}Q\|_1=0\\\|\mu\circ\boldsymbol{\Delta}\|_1=0\\\|\tau\circ(\boldsymbol{I}^{\mathrm{T}}-\boldsymbol{\Delta})\|_1=0\end{cases}。\tag{16}$$

式中:运算符“∘”表示矩阵的哈达玛(Hadamard)积。

根据 Zipf 分布的性质,p_i 的大小将随着 i 的增大逐渐减小,因此,不难预计 δ_i 的变化也遵循相似态势,即 $\boldsymbol{\Delta}$ 中需要在流行度较高的内容中出现尽可能多的 1,而在流行度较低的内容中出现尽可能少的 0。本文由此对 δ_i 进行分段讨论。

假设当 $1\leqslant i\leqslant N_{\mathrm{L}}$ 时,$\delta_i=1$,显然 $\mu_i=0$,故此时关于 τ_i 的约束如式(17)所示:

$$\tau_i=p_iK_{\mathrm{CC}}\mathrm{e}^{-K_{\mathrm{CC}}}-\theta\geqslant 0。\tag{17}$$

假设当 $N_{\mathrm{L}}+1\leqslant i\leqslant N_{\mathrm{H}}$ 时,$0<\delta_i<1$,显然此时 $\mu_i=\tau_i=0$,故关于 δ_i 的约束如式(18)所示:

$$\theta=p_iK_{CC}e^{-\delta_iK_{CC}}。\tag{18}$$

假设当 $N_H+1\leqslant i\leqslant N_F$ 时,$\delta_i=0$,显然此时有 $\tau_i=0$,故关于 μ_i 的约束如式(19)所示:

$$\mu_i=\theta-p_iK_{CC}\geqslant 0。\tag{19}$$

综上,关于 δ_i 的求解方程组如式(20)所示:

$$\begin{cases}\sum\limits_{i=N_L+1}^{N_H}\delta_i=N_C-N_L\\ p_{N_H+1}\leqslant p_{N_L}e^{-K_{CC}}\\ \prod\limits_{i=N_L+1}^{N_H}p_ie^{-\delta_iK_{CC}}\leqslant(p_{N_L}e^{-K_{CC}})^{N_H-N_L}\end{cases}。\tag{20}$$

结合式(18),当 $N_L+1\leqslant i\leqslant\lfloor N_Le^{\frac{K_{CC}}{\beta}}\rfloor$ 时,可以求得 δ_i 的表达式如式(21)所示:

$$\delta_i=\frac{\frac{\beta}{K_{CC}}\sum\limits_{j=N_L+1}^{\lfloor N_Le^{\frac{K_{CC}}{\beta}}\rfloor}\ln\left(\frac{j}{i}\right)+N_C-N_L}{\lfloor N_Le^{\frac{K_{CC}}{\beta}}\rfloor-N_L}。\tag{21}$$

式中:变量 N_L 由不等式确定,具体计算方式为寻找满足式(22)的最大正整数。

$$\frac{\beta}{K_{CC}}\sum_{i=N_L+1}^{\lfloor N_Le^{\frac{K_{CC}}{\beta}}\rfloor}\ln\left(\frac{i}{N_L}\right)\geqslant\lfloor N_Le^{\frac{K_{CC}}{\beta}}\rfloor-N_C。\tag{22}$$

另外需要说明的是,在不同内容关于 CUs 本地存储空间的随机放置完成以后,上述算法有必要通过交付系数保证自身的成功执行。计算后得到各个 CU$y\in G_{CU}$ 的交付系数 η_y 如式(23)所示:

$$\eta_y=\min\left\{1,\frac{1-\xi}{\lambda t_C}\left(\sum_{x\in G_{RU}}\sum_{c_i\in C_F}g_{y,i}(x)p_i\right)^{-1}\right\}。\tag{23}$$

至此,本文提出的基于统计分布的内容缓存与交付方案的具体描述如下:

输入:单个 CU 的本地存储空间大小 N_C,CUs 最大通信范围内的 CUs 数量 K_{CC},Zipf 指数 β,每个 RU 请求遵循的相同到达率 λ,CUs 模式的队列服务时间 t_C

输出:缓存策略矢量 $\boldsymbol{\Delta}$,各个 CU$y\in G_{CU}$ 处的交付系数 η_y

1 初始化 $N_L=1$,$N_L'=\lceil N_Fe^{-\frac{K_{CC}}{\beta}}\rceil$//确立 N_L 的取值范围;
2 While $N_L'-N_L>1$
$\hat{N}_L=\lceil\frac{N_L+N_L'}{2}\rceil$
If $\hat{N}_L$ 满足式(22)
$N_L=\hat{N}_L$
Else
$N_L'=\hat{N}_L$
End if
End while//用二分法求解 N_L;
3 For $i=1,N_L$
$\delta_i=1$
End for//对 $\boldsymbol{\Delta}$ 赋值;
4 For $i=N_L+1,\lfloor N_Le^{\frac{K_{CC}}{\beta}}\rfloor$
按式(21)计算 δ_i
End for//对 $\boldsymbol{\Delta}$ 赋值;
5 For $i=\lceil N_Le^{\frac{K_{CC}}{\beta}}\rceil,N_F$
$\delta_i=0$
End for//对 $\boldsymbol{\Delta}$ 赋值;
6 For $y\in G_{CU}$
按式(23)计算 η_y
End for//计算所有的 η_y,共 K_{CF} 个。

算法执行完毕以后,则需要进行内容 c_i 的具体放置($i\in\{1,2,\cdots,N_F\}$):考虑内容 c_i,计算出 $K_{CF}\delta_i$ 的数值并取整,代表内容 c_i 应当被放置在$\lfloor K_{CF}\delta_i+\frac{1}{2}\rfloor$个 CUs 中,然后随机选择相应数量的 CUs,令它们缓存指定内容即可(这里实际上取了 $K_{CF}\delta_i$ 的近似值,故得到的是近似全局最优解)。倘若 CUs 本地存储空间已被完全利用且各个 CU$y\in G_{CU}$ 的 C_C 内部没有重复元素的条件得以保证,所有内容 $c_i\in C_F$ 的放置完成之后,形成的 CUs 内容集即为本文讨论的内容缓存方案的执行结果。当所有 RUs 都向 F-AP 发送内容请求时,F-RAN 系统最终执行的内容交付方案的完整模式选择流程如图 2 所示。

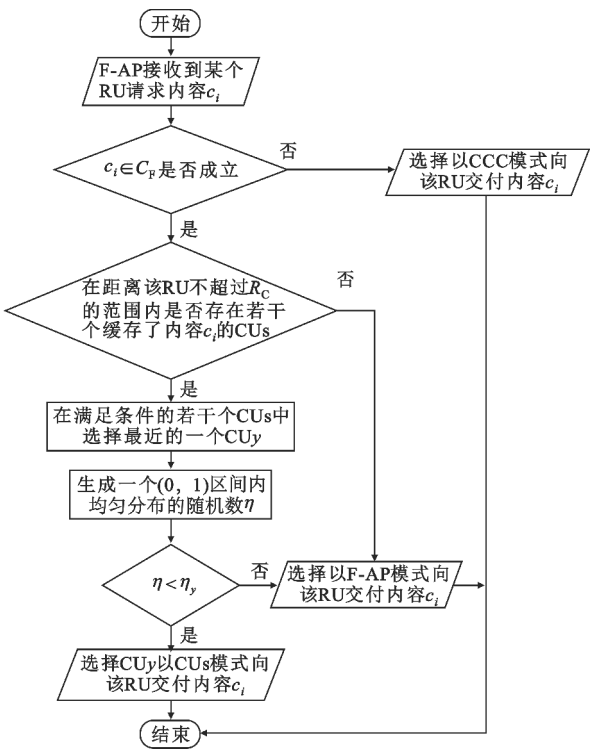

图 2　内容交付方案模式选择流程

4 仿真实验与结果分析

4.1 参数设置

针对基于统计分布的内容缓存与交付方案的仿真实验使用 Matlab 完成,通过执行具体 F-RAN 网络中的重复随机试验取得 1 000 组数据,依据其均值评估本文所提算法。仿真实验中使用的详细初始参数如表 1 所示。

表 1 仿真实验基本参数

参数	值
独立 PPP Φ_R 的密度 γ_R/(个/m^2)	10^{-3}
每个 RU 请求遵循的相同到达率 λ/(个/s^1)	3.5
F-AP 到 RUs 的数据传输速率 r_F/(Mb/s)	2[26]
CUs 和 RUs 间的数据传输速率 r_C/(Mb/s)	20
CCC 服务器存储所有内容的大小 s/Mb	5
F-AP 发送指令部分的时延 t_{FC}/ms	25
CCC 模式经历的固定下载延迟 t/ms	5 000
F-AP 覆盖区域的半径 R_F/m	150
CUs 的最大传输距离 R_C/m	30[16]
CCC 服务器存储的总体内容库的大小 N	4 000
F-AP 本地存储空间的大小 N_F	1 000
单个 CU 本地存储空间的大小 N_C	80
限制服务强度的较小正数 ξ	0.1

4.2 实验结果分析

本文选取了文献[12]中介绍的两种内容缓存与交付方案进行对比实验,通过仿真程序分别实现了基于概率(Probability-based,PB)的缓存策略、基于边缘缓存用户分类(Edge Caching Users Classification,EUC)的缓存策略以及基于统计分布(Frequency Distribution,FD)的内容缓存与交付方案,目的是验证本文所提方案分别在优化 CUs 缓存命中率和平均下载延迟方面的优越性和有效性。

图 3 展示了不同内容缓存与交付方案下,CUs 缓存命中率随 CUs 分布所服从的齐次 PPPΦ_C 密度变化的情况。此处设置变量 $\gamma_C \in [2\times10^{-4}, 10\times10^{-4}]$,Zipf 指数 $\beta=0.5$。由图 3 可以看出,CUs 缓存命中率会随着 CUs 分布密度的增加而不断增加,但鉴于取整运算等原因,导致图像出现了数值跳变。不难理解,当 γ_C 数值较小时,CUs 数量少,因而形成的整体缓存空间较小,能够缓存的内容也很少,这使得 P_{HC} 的数值偏低。然而当 γ_C 的数值增大时,除了 CUs 形成的整体缓存空间增大外,单个 RU 周边存在 CUs 的概率也在增加。此时,本文提出的 FD 方案尽可能地考虑了 CUs 可提供服务的多样性,避免了单个 RU 周边存在 CUs 的数量变多时缓存内容重复而可能出现的无效缓存现象(内容已被缓存但没有机会进行交付)。因此,相较于其他内容缓存与交付方案,FD 方案在 CUs 缓存命中率方面具有显著优势。

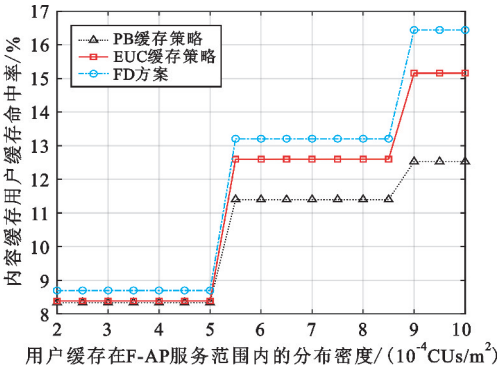

图 3 CUs 分布服从不同密度齐次 PPP 对 CUs 缓存命中率的影响

图 4 展示了不同内容缓存与交付方案下,CUs 缓存命中率随 Zipf 指数变化的情况。该部分的其他参数设置还包括变量 $\beta \in [0.15, 0.65]$ 以及 $\gamma_C = 5\times10^{-4}$。根据 Zipf 分布的性质,当 β 增大时,越来越少的高流行度文件将有越来越高的概率被 RUs 请求。此时,即使缓存内容保持不变,CUs 缓存命中率也会提高。仿真结果表明,这一过程中 FD 方案相较于其他缓存策略始终保持了显著的性能优势,这是因为它有效减少了 RUs 附近的 CUs 因缓存内容相同而造成的缓存空间冗余。从期望的视角观察,F-RAN 网络凭借 FD 方案的内容丰富性和它本身被请求概率增大两方面因素获得了更高的 CUs 缓存命中率。更重要的是,本文提出的算法可以达到 CUs 缓存命中率最大化问题的近似全局最优解。

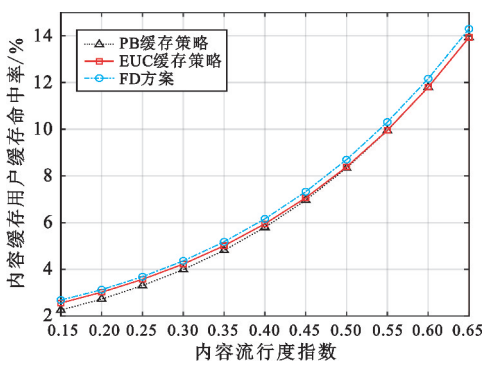

图 4 不同 Zipf 指数对 CUs 缓存命中率的影响

图 5 展示了不同内容缓存与交付方案下,平均下载延迟随 Zipf 指数变化的情况,其中补充参数的设置与图 4 部分相同。这项仿真在实验设计上具备的主要差异在于,由于 PB 和 EUC 缓存策略本身并不包含交付系数的概念,在假定各个 CU 的交付系数都为 1 的情况下,就需要舍弃会形成无限队列的部分数据,而仅保留使平均下载延迟计算有效的数据。图 5 的结果显示,当 CUs 缓存的内容变得更加流行以后,更多的 RUs 请求将会通过 CUs 模式交付,D2D 链路则会始终保持内容交付方面的性能优势。这意味着提高 CUs 缓存命中率可以通过让更多的流量负载向 CUs 转移从而减小平均下载延迟,证明了本文所提算法的有效性和执行边缘缓存的价值。除此之外,就 FD 方案而言,即使有队列等待时间,相比其他两种缓存策略,它用内容缓存方案中的随机性和内容交付方案中的交付系数对其进行控制与均衡,最终既没有浪费缓存空间,也实现了更优越的性能。仿真实验的设计还表明,限制单个 CU 处的内容请求到达率是必要的,FD 方案利用交付系数避免了无效数据,增强了算法的鲁棒性。

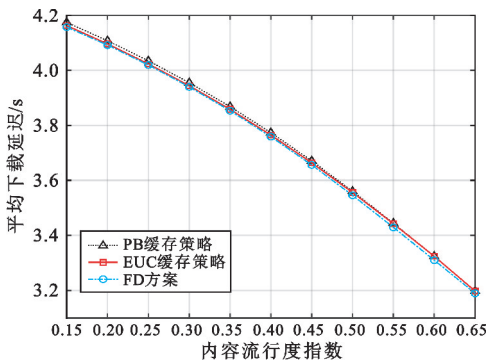

图 5　不同 Zipf 指数对平均下载延迟的影响

5　结束语

本文主要针对雾网络中的缓存进行性能优化,提出了基于统计分布的内容缓存与交付方案,利用带 KKT 条件的拉格朗日乘数法最大化了 CUs 缓存命中率,并且通过仿真实验分析了各个变量对有关性能指标造成的影响,验证了优化 CUs 缓存命中率与降低平均下载延迟的相关性。结果表明,本文设计的算法有效,实现了预期效果,达到了优化目标的近似全局最优解,可以始终保持性能优势。因此,在设计内容缓存与交付方案时,直接采取期望视角是必要的,本文从节点分布和用户请求等建模开始,到算法中考虑的内容在概率和统计分布方面的特征都体现了这一点。值得注意的是,为了保证内容缓存与交付方案的成功执行,需要将交付控制纳入考量,事实上它是方案性能的重要影响因素。

下一步将考虑流行度预测、用户移动性等对模型更精确的随机表示和更多的优化方法,探讨此类研究思路与本文的差异。

参考文献:

[1] JIANG W, FENG G, QIN S, et al. Multi-agent reinforcement learning based cooperative content caching for mobile edge networks[J]. IEEE Access, 2019, 7: 61856-61867.

[2] CISCO. CISCO annual Internet report (2018—2023) white paper[EB/OL]. (2020-03-09)[2022-03-19]. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf.

[3] PENG M, LI Y, ZHAO Z, et al. System architecture and key technologies for 5G heterogeneous cloud radio access networks[J]. IEEE Network, 2015, 29(2): 6-14.

[4] PENG M, WANG C, LAU V, et al. Fronthaul-constrained cloud radio access networks: insights and challenges[J]. IEEE Wireless Communications, 2015, 22(2): 152-160.

[5] 李碧瑶. 边缘网络下的计算卸载和边缘缓存方法研究[D]. 北京:北京邮电大学, 2020.

[6] PENG M, YAN S, ZHANG K, et al. Fog-computing-based radio access networks: issues and challenges[J]. IEEE Network, 2016, 30(4): 46-53.

[7] BONOMI F, MILITO R, ZHU J, et al. Fog computing and its role in the Internet of Things[C]//Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing. New York: Association for Computing Machinery, 2012: 13-16.

[8] WANG X, CHEN M, TALEB T, et al. Cache in the air: exploiting content caching and delivery techniques for 5G systems[J]. IEEE Communications Magazine, 2014, 52(2): 131-139.

[9] LI Z, CHEN J, ZHANG Z. Socially aware caching in D2D enabled fog radio access networks[J]. IEEE Access, 2019, 7: 84293-84303.

[10] LI Q, NIU H, PAPATHANASSIOU A, et al. Edge cloud and underlay networks: empowering 5G cell-less wireless architecture[C]//Proceedings of 2014 20th European Wireless Conference. Barcelona: IEEE, 2014: 1-6.

[11] WANG X, LENG S, YANG K. Social-aware edge caching in fog radio access networks[J]. IEEE Access, 2017, 5: 8492-8501.

[12] HUA H, CHU X. Content caching policy with edge caching user classification in fog radio access networks

[C]//Proceedings of 2021 IEEE Wireless Communications and Networking Conference. Nanjing: IEEE, 2021: 1-7.

[13] LIU T, LI J, KIM B, et al. Distributed file allocation using matching game in mobile fog-caching service network[C]//Proceedings of 2018 IEEE Conference on Computer Communications Workshops. Honolulu: IEEE, 2018: 499-504.

[14] ZHOU X, LIU Z, GUO M, et al. SACC: a size adaptive content caching algorithm in fog/edge computing using deep reinforcement learning[J]. IEEE Transactions on Emerging Topics in Computing, 2022, 10(4): 1810-1820.

[15] YAN S, JIAO M, ZHOU Y, et al. Machine-learning approach for user association and content placement in fog radio access networks[J]. IEEE Internet of Things Journal, 2020, 7(10): 9413-9425.

[16] JIANG F, ZHANG X, SUN C. A D2D-enabled cooperative caching strategy for fog radio access networks[C]//Proceedings of 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. London: IEEE, 2020: 1-6.

[17] CHEN Z, LEE J, QUEK T, et al. Cooperative caching and transmission design in cluster-centric small cell networks[J]. IEEE Transactions on Wireless Communications, 2017, 16(5): 3401-3415.

[18] YU K, MA Z, NI R, et al. A caching strategy based on many-to-many matching game in D2D networks[J]. Tsinghua Science and Technology, 2021, 26(6): 857-868.

[19] BRESLAU L, CAO P, FAN L, et al. Web caching and Zipf-like distributions: evidence and implications[C]//Proceedings of Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. New York: IEEE, 1999: 126-134.

[20] KINGMAN J. The first Erlang century—and the next[J]. Queueing Systems, 2009, 63(1/2/3/4): 3-12.

[21] ANDREWS J, BACCELLI F, GANTI R. A tractable approach to coverage and rate in cellular networks[J]. IEEE Transactions on Communications, 2011, 59(11): 3122-3134.

[22] CAHN R. Wide area network design: concepts and tools for optimization[M]. San Francisco: Morgan Kaufmann Publishers Inc., 1998.

[23] KLEINROCK L. Queueing systems[M]. Hoboken: Wiley-Interscience, 1975.

[24] BOYD S, VANDENBERGHE L. Convex optimization[M]. New York: Cambridge University Press, 2004.

作者简介：

陈昊宇 男，1998 年生于安徽合肥，2020 年获工学学士学位，现为硕士研究生，主要研究方向为无线通信、边缘计算。

胡宏林 男，1975 年生于安徽安庆，2004 年获工学博士学位，现为研究员、博士生导师，主要研究方向为移动通信技术、脑机通信技术。