Hierarchical Edge Caching in Device-to-Device Aided Mobile Networks: Modeling, Optimization, and Design
Xiuhua Li，Xiaofei Wang. 英属哥伦比亚大学，天津大学
IEEE Journal on Selected Areas in Communications (IEEE JSAC) 2018, CCF-A.

Backhaul，在分层电信网络中，网络的回程部分包括核心网络或骨干网络与整个分层网络的“边缘”处的小型子网络之间的中间链路。

背景：

随着在线社交流行和普及，来自移动用户的内容请求（如视频、音频、相片等）呈爆炸式增长；
然而，有效地支持这些大量的请求对 MNOs 是一个巨大的挑战：
网络资源缺乏，特别是在 RANs 和 backhaul 网络。

相关工作

1. Uncoded/coded FemotCaching in BSs；
2. 在 D2D 层面 cache 策略；
3. …

这些 cache 策略都只关注了 BSs 层面和 D2D 层面中的的一种；
也有使用分层的 caching 策略的，但却在 web caching 或者是有线通信的情景下；

本文工作考虑的方面：
考虑了在 D2D 层面和 BSs 层面的分层的 edge caching；
考虑了社交行为和用户偏好，设计出 D2D 层和 BSs 层的协作机制；

Motivation

分层的 caching 已被广泛并有效地应用到 web caching 系统中；
D2D 通信优点：提升网络频谱效率、开销降低网络能耗；
Edge cache 优点：卸载网络流量、减少系统开销、提高 QoS 和 QoE，

挑战：

NP-hard

系统建模 System Modeling

a. 分层边缘缓存 Architecture 和 Topology

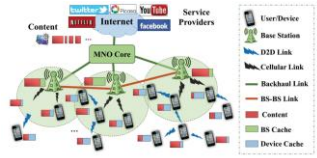


Fig. 1. Illustration of hierarchical edge caching architecture in D2D aided mobile networks.

MNO 核心网-BS 之间、MNO-SP 之间采用 backhaul 网络进行通信;
BSs 之间采用全连接的方式使用高速电缆或者光纤进行通信;
BS-User 之间采用蜂窝网络进行通信;
User-User 之间采用 wifi 直连或者蓝牙进行通信。

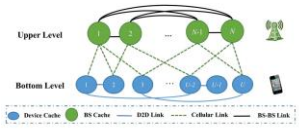


Fig. 2. Topology of hierarchical edge caching in D2D aided mobile networks.

b. 内容流行度和用户偏好
内容流行度指的是所有用户内容请求的概率分布，其符合 MZipf(曼德尔布罗-齐夫)分布;
用户偏好指的是一个用户对于内容请求的概率分布;

c. D2D Sharing Model

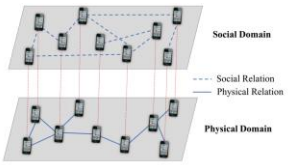


Fig. 3. An illustration of user relation in both physical domain and social domain.

Physical domain，是从地理位置角度考虑用户之间的连接情况;
根据连接时长，计算用户 u 和 v 的相遇概率:

$$E_{uv}^{D}=\frac{\sum_i T_{uv}^{D}(i)}{T_{\text{tot}}},\quad \forall u\in\mathcal{U},\ \forall v\in\mathcal{U}, \tag{3}$$

Social domain，是从用户社交关系（偏好、关系类型）考虑用户之间的连接情况。
根据内容请求，使用余弦相似度公式计算用户偏好相似度:

$$C_{uv}=\frac{\sum_{f\in\mathcal{F}} q_u^f q_v^f}{\sqrt{\sum_{f\in\mathcal{F}}(q_u^f)^2}\sqrt{\sum_{f\in\mathcal{F}}(q_v^f)^2}},\ \forall u\in\mathcal{U},\ \forall v\in\mathcal{U}. \tag{4}$$

使用 k-means 聚类，划分用户关系类型:

$$R_{uv}=\begin{cases}1, & \text{self } (u=v),\\ \alpha_1, & \text{close friends},\\ \alpha_2, & \text{normal friends},\\ \alpha_3, & \text{strangers},\end{cases} \tag{5}$$

d. Association of Uers and BSs
这里假设了当 local BSs cache 能满足 user 时，user 只从 local BSs 中取内容;
否则 local BSs 请求其他的 BSs;
若还不行，则从 Internet 上下载内容。
根据连接时长统计，用户 u 请求 BS n 的概率:

$$p^{B}\{u|\text{BS } n\}=\frac{\sum_i T_{un}^{B}(i)}{\sum_{n\in\mathcal{N}}\sum_i T_{un}^{B}(i)},\quad \forall u\in\mathcal{U},\ \forall n\in\mathcal{N}, \tag{8}$$

e. Cache Scheme

Cache scheme 分为两个阶段：Content Placement 和 Content Delivery.

（1）Content Placement:

在 BSs 中，采用具有 MDS 性质的纠删码，将 content 进行编码后放置在 BSs 中；

在 user 设备中，整份 content 进行缓存。

为什么在 user 设备中不进行纠删编码呢？

从 分享起来便捷、用户 QoS/QoE、复杂度 角度考虑。

（2）Content Delivery

D2D sharing —> Uers-BS 交付 —> BS-BS 协作 —> 直接下载

f. Content-Centric Control and Management

MNO Core 管理着计算和通信资源，并对 content placement 和 content delivery 做出决策。

分层边缘缓存框架设计

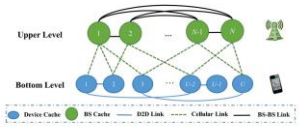


Fig. 2. Topology of hierarchical edge caching in D2D aided mobile networks.

分层缓存问题，是一个 NP-hard 的问题。本文采用分解 NP-hard 问题，并对子问题进行求最优解或次优解的方式，得到分层缓存的近似解。

（1）Cooperation in Bottom Level

$$L_u^f = x_u^f \lambda_u^f s_f + \sum_{v \in \mathcal{U}} p_{uv}^D [x_u^f (1 - x_v^f)] \lambda_v^f s_f, \ \forall u \in \mathcal{U}, \ \forall f \in \mathcal{F}, \tag{10}$$

平均流量负载：

将这个 二元整数约束线性二次规划问题 转为 二元整数线性规划问题，再转化为两面凸的最优化问题，再用额外惩罚方法(EPM)求解。

（2）Cooperation in Upper Level

工作场景

本文的应用场景是缓冲相当长时间内流行度不发生变化的 content，如 video、music 等。

工作流程

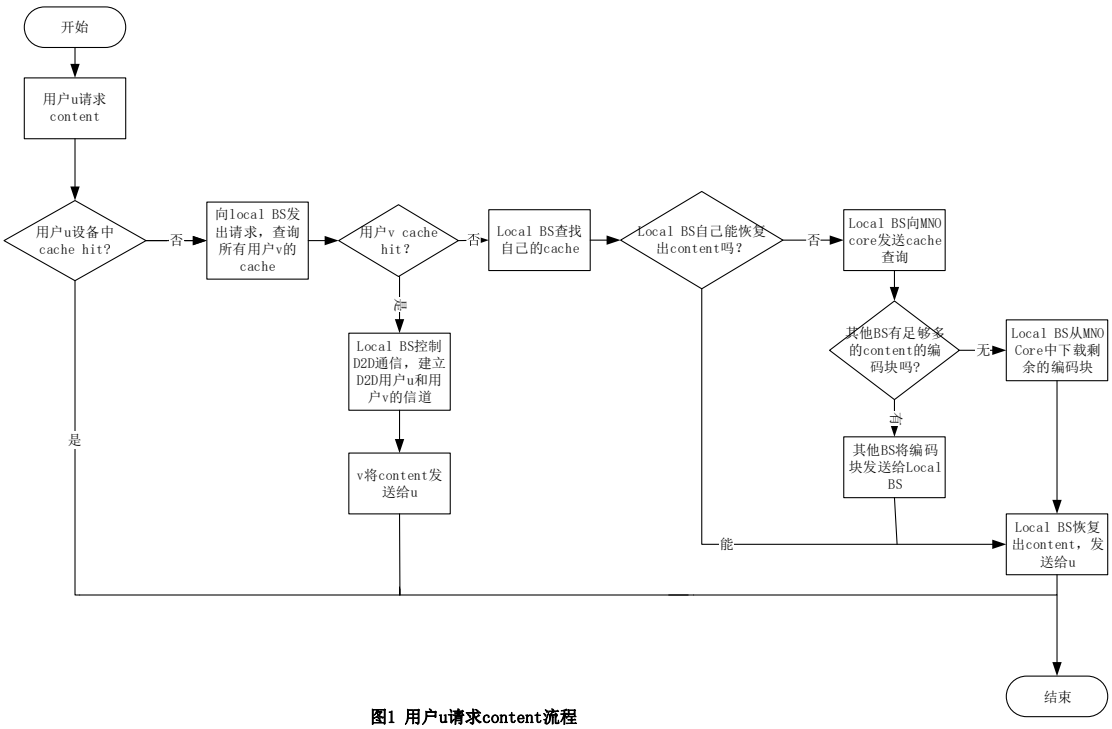


图1 用户u请求content流程

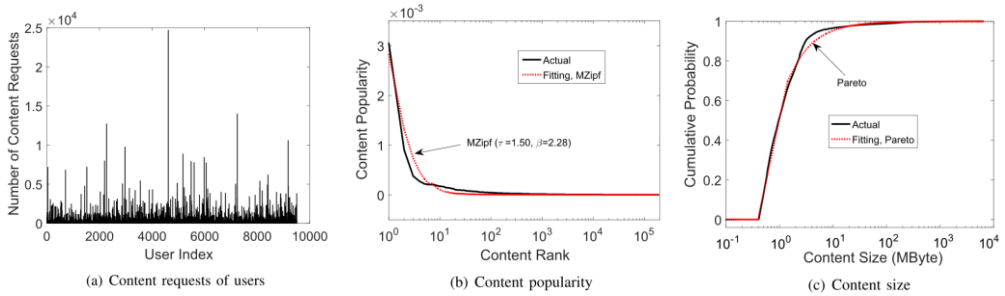
实验仿真设置

实验数据：Xender，2016 年 2 月
仿真器：用 python 实现自建的仿真器；
所有的用户设备中的 cache 大小相同，BS 同理；
缓存在 BS 中的被编码的数据块大小为 64KB；
BS-BS 的链路带宽为 1Gbps；
…

实验数据的展示

（1） 实验的数据集是：2016 年 2 月份的 Xender 的跟踪，包含了 9514 台用户设备，188447 份 content 文件，2107100 次用户请求。

（2） 对数据集进行了分析，结果如下：



(a) Content requests of users (b) Content popularity (c) Content size

以上三个图分别表明了：用户的请求次数分布均匀、content 流行度符合 MZipf 分布、content 的大小呈 Pareto 分布（即二八定律，此处指小的 content 更多）。这也验证了实验的对于内容流行度呈 MZipf 的假设是合理的。

（3） 性能提升

本文工作目标：

降低在 backhaul 网络上的流量负载，以降低所带来的经济成本.

实验设置：

所有的用户设备缓存大小相同，固定为 1GB ； BS 数量为 10 个，编码块大小为 64KB，BS-BS 链路带宽为 1Gbps，所有的 BS 缓存大小相同。以下实验研究随着 BS cache size 增大，四个性能表现（缓存系统在 backhaul 流量卸载百分比、缓存系统支持请求次数百分比、缓存系统成本降低的百分比、缓存系统中 BS-BS 链路的利用率百分比）的提升比例（与无缓存系统相比）。

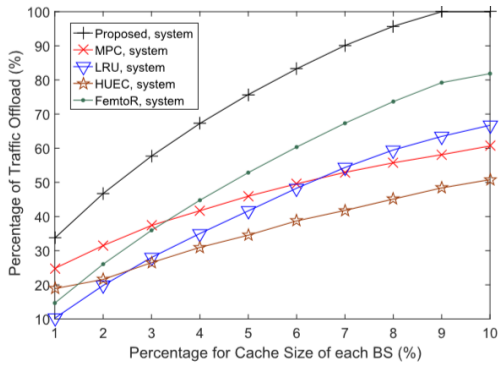
Baseline：

MPC：Most Popular Caching,该缓存系统使用本文架构，在 Upper 层和 Bottom 层均采用缓存流行度最高的 content 的 cache 替换策略；为了突出本文 cache 替换策略的有效性。

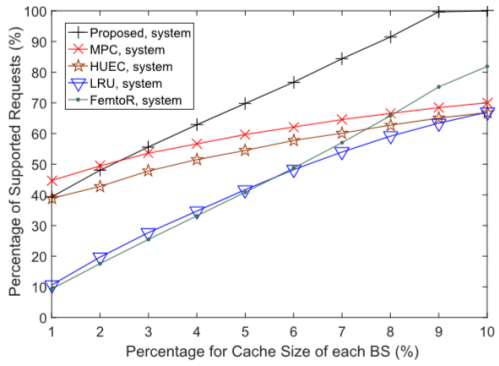
LRU：Least Recently Used,该缓存系统使用本文架构，在 Upper 层和 Bottom 层均采用 LRU cache 替换策略；为了突出本文 cache 替换策略的有效性。

HUEC：Hierarchical Uncoded Edge Caching，将其缓存系统架构修改成本文中的两级缓存架构，但在 Upper level 不对 content 进行编码，使用其论文中的 cache 替换策略；为了突出在 Upper level 使用纠删码的有效性。

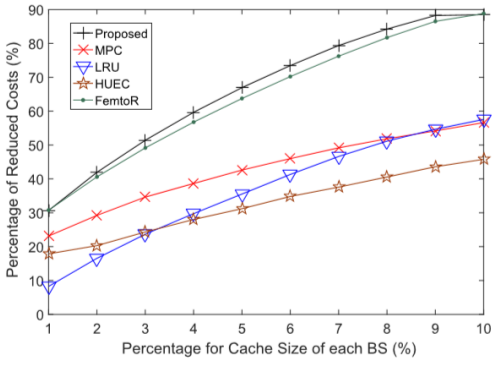
FemtoR：Revised FemtoCaching，该缓存系统使用本文架构，但不使用本文中的 D2D sharing 机制，且只在 upper level 对 content 进行编码，使用其论文中的 cache 替换策略。为了突出 D2D sharing 的有效性。



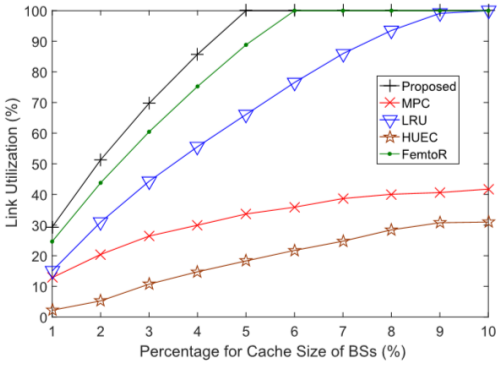
(a) Percentage of traffic offload



(b) Percentage of supported requests



(c) Percentage of reduced costs



(d) Link Utilization

以上四个图，横轴表示单个 BS 占数据集中 content 总大小的比例；纵轴分别表示：与无 cache 的系统相比，缓存系统在 backhaul 流量卸载百分比、缓存系统支持请求次数百分比、缓存系统成本（依据 backhaul 和 BS-BS 链路上的流量负载计算）降低的百分比、缓存系统中 BS-BS 链路的利用率百分比。

实验分析:

图 a 表明本文方法优于所有的 baseline，因为本文方法是最小化了（由于是 NP-hard 问题，故精确地表述是求得了最小 backhaul 流量的次优解），即尽可能地查找整个缓存系统中的 content 或者 content 编码块。

图 b 中，MPC 方法在横坐标为 1 时，是优于本文方法的。因为 Content Popularity 是根据采样时间内 User Request 统计得来的，故当 BS 不能比较好发挥出其缓存作用时（即 BS 缓存空间较小时），MPC 能达到最优的 Supported Requests 指标；但当 BS 逐渐增大时，本文方法便超过了 MPC。

图 d 中，本文方法与 FemtoR 的 Link Utilization 相近，这是因为 FemtoR 利用了由于纠删码所带来的 BS-BS 之间的协作机制。

图 c 中，本文方法与 FemtoR 的 Reduced Costs 相近，这是因为 FemtoR 的 BS-BS 链路利用率高，需要从 Backhaul 下载编码块的数据量小（Costs 主要由 Backhaul 上的流量决定的）。

总结

本文提出了一种基于 D2D 辅助移动网络的协作分层边缘缓存框架。

基于

移动用户的社会行为和偏好,

分析异构缓存大小,

派生系统拓扑结构。

最大化网络容量（network capacity），实现了

卸载网络流量,

降低系统成本,

支持移动用户在网络中请求内容,

并从工程实现的角度提出了低复杂度的解决方案。

基于跟踪的仿真结果表明，所提出的分层边缘缓存框架具有良好的性能，在卸载网络流量、满足内容请求和降低系统成本方面优于考虑的四种基线方案。