

# Joint Online Edge Caching and Load Balancing for Mobile Data

## Offloading in 5G Networks

ICDCS 2019, CCF-B

Yiming Zeng, Yaodong Huang, Zhenhua Liu, Yuanyuan Yang

纽约州立大学石溪分校

### 题目解释

Online, 相比 Offline, 其未知未来的信息或者只能预测未来近期的数据。

Joint..., 同时考虑 edge caching 和 load balancing, 分别需要存储设备和计算设备。

Data Offloading, 指的是将流量从 BS (基站) 或者回程网络中卸载下来。

### 背景

近年来, 智能手机和平板电脑等智能**边缘设备的数量**呈爆炸式增长。

全球**移动数据流量**预计将在 2016 年至 2021 年期间增长 7 倍[2], 并可能耗尽蜂窝网络的通信能力。

**网络流量具有时变性**, 这给低流量期间执行缓存更新提供了机会, 从而降低了峰值流量需求和传输延迟[7]。

**5G 无线网络**设计具有带宽大、天线规模大、频率复用高、网络密度大等特点[4]。越来越多的小型基站(SBSs), 例如 microcell、picocell 和 femtocell 基站[5], 正在部署中, 以便以更低的成本获得更好的服务质量。

5G 网络中的 SBSs 配备了移动计算服务器, 既有存储设备也有计算设备[6]。计算设备使 SBS 能够在相互竞争的用户请求之间谨慎地平衡负载。

### 相关工作

1. Poularakis 等人定义了一个由一个 BS 和几个 SBSs 组成的网络, BS 的运营商租用可用的缓存和带宽资源来满足 MUs 的请求, 该问题被表示为一个两阶段 Stackelberg 博弈, 以最小化总服务成本[8]。
2. Du 等人设计了激励流量卸载和资源分配契约, 以激励 SBS 最大化其效用[9]。
3. 在[10]中, 作者考虑了一个异构蜂窝网络, 并设计了一个分布式缓存问题, 通过信念传播来最小化下载延迟。

[8] K. Poularakis, G. Iosifidis, and L. Tassiulas, "A framework for mobile data offloading to leased cache-endowed small cell networks," in Mobile Ad Hoc and Sensor Systems (MASS), 2014 IEEE 11th International Conference on. IEEE, 2014, pp. 327–335.

[9] J. Du, E. Gelenbe, C. Jiang, H. Zhang, and Y. Ren, "Contract design for traffic offloading and resource allocation in heterogeneous ultra-dense networks," IEEE Journal on Selected Areas in Communications, vol. 35, no. 11, pp. 2457–2467, 2017.

[10] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink

## 动机

上述大多数研究解决的是一个缓存问题，没有考虑传输带宽限制，这是不现实的。同时，他们也没有考虑系统存储成本和替换成本。

## 挑战

1. **缓存策略和负载均衡是相互依赖的，需要共同优化。**缓存策略依赖于 MUs 的负载均衡决策来决定要(重新)放置哪些内容。MUs 需要知道每个 SBS 是否缓存了请求的内容，以便在 SBSs 之间分发它们的请求
2. 边缘缓存涉及整数规划，在许多特定的情况下，**整数规划是 NP 难的**。需要提出一个**计算效率高的算法**（非暴力法）。
3. 内容的受欢迎程度、用户需求和和其他因素可能随着时间会发生变化。所以这是一个**在线优化问题**，解决整数约束下的在线优化是一个具有挑战性（NP 难）的问题。

## 贡献

1. 提出了 5G 网络中移动数据卸载的**联合缓存策略和负载均衡的在线优化问题**。
2. 提出了一种基于**对偶分解的离线问题分解算法**。第一个子问题(缓存问题)是一个可以直接求解的标准凸优化问题。我们将另一个子问题（负载均衡）中的整数变量松弛为连续变量，并证明了这种松弛是精确的，即，保证了最优整数解。
3. 为了在有限预测的情况下在线解决这一问题，引入了三种很有前途的在线算法，即滚动时域控制 (RHC)[19]、Averaging Fixed Horizon Control (AFHC)和 Committed Horizon Control (CHC)。**证明理论界**。
4. 通过大量的仿真验证了在线算法的性能。结果表明，与现有算法相比，**该算法降低了系统成本(最多 27%)**

## 系统建模

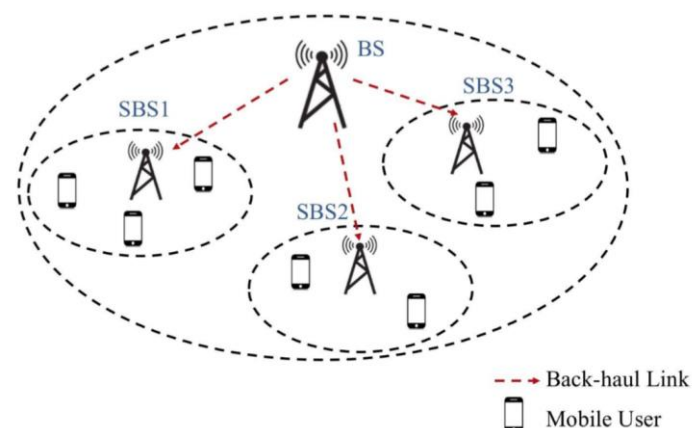


Fig. 1. An example of our proposed system model.

假设 SBSs 不会干扰 BS，SBSs 信号覆盖范围互不相交。

Content 大小均为 1。

当 Mus 的请求的 content 在 SBS 的缓存中且有足够的带宽进行传输时，首选 SBS。否则，请求将由 BS 提供服务。

存储容量约束：

$$\sum_{k \in \mathcal{K}} x_{n,k}^t \leq C_n, \forall n \in \mathcal{N}, t \in \mathcal{T}. \quad (1)$$

带宽约束：

$$\sum_{k \in \mathcal{K}} \sum_{m_n \in \mathcal{M}_n} \lambda_{m_n,k}^t y_{m_n,k}^t \leq B_n, \quad \forall n \in \mathcal{N}, t \in \mathcal{T}. \quad (2)$$

缓存约束：

$$y_{m_n,k}^t \leq x_{n,k}^t, \quad \forall n \in \mathcal{N}, m_n \in \mathcal{M}_n, k \in \mathcal{K}, t \in \mathcal{T}, \quad (3)$$

which implies that if  $x_{n,k}^t = 0$ , then  $y_{m_n,k}^t = 0$ .

缓存占比总和为 1：

$$z_{m_n,k}^t + y_{m_n,k}^t = 1, \quad \forall n \in \mathcal{N}, \forall m_n \in \mathcal{M}_n, \forall k \in \mathcal{K}, \forall t \in \mathcal{T}. \quad (4)$$

$$z_{m_n,k} = 1 - y_{m_n,k}^t$$

## 求解方法

目标函数：分为三部分

BS 处的开销函数：

$$f_t(Y^t) = \sum_{n \in \mathcal{N}} \left( \sum_{m_n \in \mathcal{M}_n} \omega_{m_n} \sum_{k \in \mathcal{K}} (1 - y_{m_n,k}^t) \lambda_{m_n,k}^t \right)^2. \quad (5)$$

SBS 处的开销函数：

$$g_t(Y^t) = \sum_{n \in \mathcal{N}} \left( \sum_{m_n \in \mathcal{M}_n} \hat{\omega}_{m_n} \sum_{k \in \mathcal{K}} y_{m_n,k}^t \lambda_{m_n,k}^t \right)^2. \quad (6)$$

SBS 处的从 t-1 时刻到 t 时刻的缓存替换开销函数：

$$h(X^t, X^{t-1}) = \sum_{n \in \mathcal{N}} \beta_n \sum_k \left( x_{n,k}^t - x_{n,k}^{t-1} \right)^+. \quad (8)$$

$$d(x_n^t, x_n^{t-1}) = \beta_n \sum_k \left( x_{n,k}^t - x_{n,k}^{t-1} \right)^+, \quad (7)$$

其中  $\beta_n$  包括来自不同来源的成本，例如，缓存替换的能源成本、更新期间发生的延迟成本、从 BS 下载内容的流量成本。

最优化求解方法：

采用原始对偶算法。将原始的整数规划问题（NP-Hard）加上松弛的变量，转变成线性规划问题的标准型，根据“对偶问题的最优解就是原始问题的最优解”，对对偶问题进行求出一个可行解。再将该可行解代入对偶问题中，迭代进行从而解出最优解。

## 实验设置

BS 个数为 1，SBS 个数为 1

Content 个数  $K=30$

时间槽个数为 100

SBS 的存储容量均为 5

SBS 的带宽均为 30，即一个 SBS 在一个时间槽内最多能传输 30 个 content

MUs（移动用户）个数为 30，在 SBS 的信号覆盖范围内呈正态分布。

BS 与 MU 之间的传输效率为  $\omega_{m_n}$ ，[0,1]区间内随机选取。

BS 与 MU 之间的传输效率为  $\hat{\omega}_{m_n}$  设置为 0，意为传输能量消耗或时延可以忽略不计。

用齐夫-曼德尔布罗分布去模拟 MUs 的请求模式。

## 实验分析

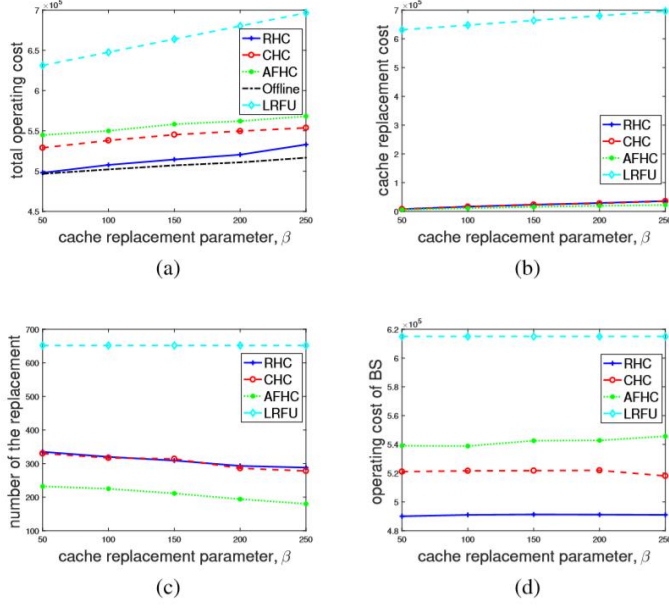


Fig. 2. The impact of the cache replacement cost  $\beta$ . (a) The total operating cost. (b) The cache replacement cost. (c) The number of cache replacement times. (d) The operating cost of BS.

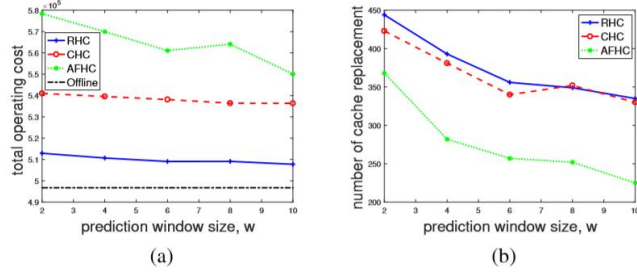


Fig. 3. The impact of the predict window  $w$ . (a) The total operating cost. (b) The number of cache replacement times.

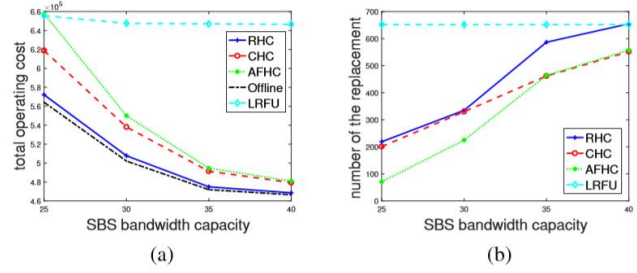


Fig. 4. The impact of the SBS bandwidth. (a) The total operating cost. (b) The number of cache replacement times.

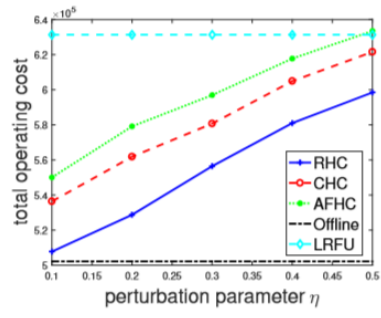


Fig. 5. The impact of the perturbation parameter  $\eta$