A Resource Allocation Perspective on Caching to Achieve Low Latency

Hsiang Hsu, and Kwang-Cheng Chen, *Fellow, IEEE*

***Abstract*—The rapid growth of Internet contents from wireless data networks, especially social media, results in unprecedented traffic volume, invoking a challenge to the load of cellular infrastructure. Moreover, instead of traditional quality of service (QoS), the demands of quality of experience (QoE) become a more practical norm, which could barely be improved under present resource allocation on radio spectrum. With this in mind, we more generally consider resource allocation by exploiting caching in radio access networks (RANs), and show that caching, as a sort of storage, could be viewed as a substitution of the communication spectrum. Thus, we propose a collaborative strategy to implement caching in infrastructure and in mobile devices simultaneously, which in general turns out to be device-to-device (D2D) communication. This new design paradigm enables great reduction of latency for requesting Internet contents and can be implemented via slight amendments to present cellular systems.**

***Index Terms*—Caching, RAN, D2D, latency, QoE, resource allocation, wireless networks.**

I. Introduction

INTERNET traffic is forecasted to increase nearly tenfold from 2014 to 2019, and hence dominates the wireless mobile data transmission volume; in particular, social media in Internet content transportation has already exceeded 50 percent of total mobile data traffic in 2012 [1]. To satisfy this dramatically growing mobile communication demands, telecommunication operators and developers have to fiercely increase cellular network capacity and backhaul bandwidth accordingly. Due to spectral usage for Long Term Evolution (LTE) standard approaching the Shannon limit, the potential solution widely studied is deploying small and femto cells, establishing a more complex structure of heterogeneous multi-layer cellular networks, with an aim of exploiting licensed and unlicensed radio resource more efficiently.

This radio access network (RAN) structure with respect to utilizing radio resource potentially introduces growing inter-cell interference (ICI) level; even more control signals like channel state information (CSI), interference alignment, power control, *etc.*, need to be imposed on present LTE system, where significant portion of wireless bandwidth is occupied by control signals [2]. In addition, as Internet traffic dominates in wireless mobile network, a more pragmatic metric than quality of service (QoS) appears to be quality of experience (QoE), particularly latency, which can barely be improved by merely increasing wireless capacity via radio resource. Furthermore, along with the shortage of radio spectrum, the insufficiency of backhaul network bandwidth and the layer-upon-layer protocols when requesting data from remote data centers remain a bottleneck for low latency [3]. Under present cellular network architecture, social media on demand have to travel from data centers behind the IP network, through packet gateway (P-GW), service gateway (S-GW), to base stations (BSs), and finally to the user equipments (UEs), as shown in Fig. 1. This detour leads to a structurally inevitable latency, which is hard to improve by merely air-interface technology.

As a result, we should not confine our view to radio resource (*i.e.* spectrum) only. Instead, we turn our focus on computation and storage (*i.e.* caching) in wireless network. Computation could be illustrated as an in-network resource for improving spectrum efficiency by preliminary computation at data aggregators, thereby deciding whether more radio resource should be used to transmit the signals or not [4]. Caching, an idea of using storage capacity, is another resource that can be easily imposed on present LTE system, and is the highlight in this letter. Due to the inhomogeneity of popularity of social media, caching objects can be considered as a complementary function to trade with radio resource; it not only alleviates the unwanted volume of backhaul traffic, but also provides a chance to satisfy QoE, *i.e.* primarily to improve latency.

Instead of treating caching separately in BS and in devices [5]–[8], in this letter, we provide a more comprehensive vision on utilizing caching as a general resource in the RAN in order to distribute social media with greatly reduced latency (Fig. 1). Thus, caching in BS (or other infrastructures) and caching in mobile devices simultaneously form a collaborative strategy, and surpass existing methods by only implementing one of them in QoE. This design actually follows top-down (application-driven) philosophy, which is different from existing philosophy. Furthermore, we show that in present LTE system, traffic monitoring function at Policy and Charging Rules Function (PCRF) and a proposed low-complexity algorithm facilitate the entire picture of resource utilization in mobile networks, regarding radio resource, backhaul resource, storage capacity, as a design paradigm shift for the next generation communication system.

II. Problem Formulation

A. Network Model

We consider a radio access network (RAN) consisting of N UEs $\{u_n\}_{n=1}^{N}$ and a base station (BS), as illustrated in Fig. 1. Within the network, K social media $\{c_k\}_{k=1}^{K}$, as the main traffic of Internet contents, are requested by the UEs. They can acquire the media either by reliable connection to the BS, or by establishing a linkage to other mobile devices who have stored the content and are able to share; that is, by device-to-device

Manuscript received June 22, 2015; accepted October 27, 2015. Date of publication November 9, 2015; date of current version January 7, 2016. This work was supported in part by the Ministry of Science and Technology under Grant MOST 104-2221-E-002-082 and in part by Mediatek Inc. under Grant MOST 103-2622-E-002-034. The associate editor coordinating the review of this paper and approving it for publication was D.W.K. Ng.

The authors are with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: r03942046@ntu.edu.tw; ckc@ntu.edu.tw).

Digital Object Identifier 10.1109/LCOMM.2015.2499193

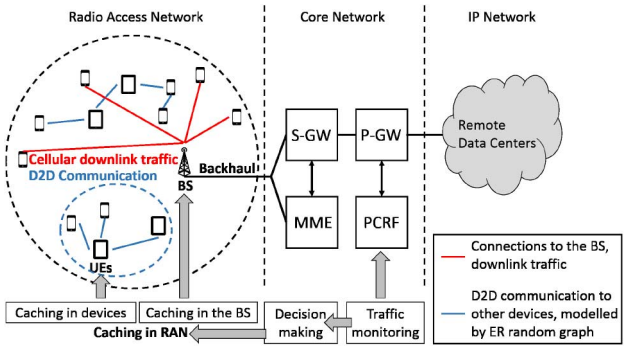

Fig. 1. System scenario. We consider caching within the RAN as an in-network resource, which could be utilized to reduce the cellular and backhaul traffic volume. Moreover, since caching reduce the need to retrieve data from remote data centers, the mean latency for data acquisition would be greatly decreased as well. This methodology could be carried out by traffic monitoring at PCRF and a new entity for decision making (see Sec. III-C).

(D2D) communication. For D2D communication, we adopt in-band overlay D2D sharing scheme, with total available licensed spectrum W in the RAN, divided into two non-overlapping parts: a fraction ηW is assigned to D2D communication while the other fraction $(1-\eta)W$ is used for cellular downlink traffic, where η is termed as the spectrum partition factor. Furthermore, pertaining to the network model among the mobile devices, we assume that for any two users in the cell, there is a probability ρ that they can communicate with each other. This assumption serves as an upper bound of the performance for D2D communication under cellular networks, since it overestimates the connection probability of two devices and neglects possible clustering structure among users. Therefore, the network topology of D2D communication can be modelled by Erdös-Rényi (ER) random graph [9] $G(N, \rho)$, as the least favorable case.

Generally, the popularities of the K social media during a period of time are non-homogeneous nowadays, and can be characterized by Mandelbrot-Zipf distributions [10]. We define p_k to represent the probability of requesting the k^{th} most popular (*i.e.* rank k) social media among the users, with skewness factor $\alpha \in [0, \infty)$, and then we have $p_k = \frac{1/(k+q)^\alpha}{H_{N,q,s}}$, $k = 1, 2, \cdots, K$, where $H_{N,q,s}$ is the normalization constant. $q \geq 0$ is the Plateau factor adjusting the Plateau shape of the left-most part of the distribution. This model degenerates to the Zipf distribution if $q = 0$. The popularities are reasonably assumed to be constants in the observed period. For the size $\{s_k\}_{k=1}^K$ of the social media, it is assumed to be Log-Normal distributed [11], *i.e.* $s_k \sim \ln \mathcal{N}(\mu, \sigma^2)$, $k = 1, 2, \cdots, K$. If an user device u_n requests c_k from other devices, the probability of successful acquisition depends on the connectivity among devices, and the popularities and the file size of the requested social media. The reason is that ρ and $\{p_k\}_{k=1}^K$ are the key factors for an user u_n requesting c_k to find other users who requested and hold c_k. Moreover, when the size of the media is large, it is less likely for u_n to obtain the social media entirely via D2D communication. Therefore, incorporate the considerations together, the existence of sharing links regarding u_n for c_k is characterized by a Bernoulli random variable $B_{nk}(\rho_k)$ with parameter $\rho_k = 1 - (1 - p_k \rho^{c s_k})^{n-1}$, $0 \leq \rho < 1$; c is a tunable variable.

For resource utilization, we consider whole caching resources in the RAN; to be more specific, a storage capacity sized M at the BS and distributed storage capacity $\{M_n\}_{n=1}^N$ at the user u_n. In the next paragraph, we show the formulation of the trade-off between caching resource and spectrum resources, especially for backhaul and downlink traffic.

B. The Multi-Objective Optimization Problem

The goal of this work is to determine how to utilize the BS cache and the bandwidth for D2D sharing. To be more specific, we have to decide the social media to be cached at the BS and to be shared among users directly by D2D communication. Since the latency for data acquisition mainly comes from latency in the core network and the BS, an efficacious method to reduce latency is to diminish the traffic volume to the BS and on the backhaul. That is, minimizing latency could be achieved by simultaneously minimizing the backhaul traffic and the downlink traffic. We subsequently introduce two binary indicators, which take values in $\{0, 1\}$ for each social media.

- θ_k: whether social media c_k is cached at the eNobeB.
- ϕ_k: whether c_k should be shared using D2D communication.

Furthermore, let $\delta_{nk} \in \{0, 1\}$ denote whether user u_n needs c_k, satisfying $\frac{\sum_n \delta_{nk}}{N} \approx p_k$, $\forall k$. The backhual traffic f_b is calculated by

$$f_b = \sum_n \sum_k \left[(1 - \theta_k - \phi_k)\, \delta_{nk} s_k + \phi_k\, (1 - B_{nk}\, (\rho_k))\, \delta_{nk} s_k\right] \tag{1}$$

where the first term in the summation is the traffic created by the media neither cached by the eNB nor shared by D2D, and the second term represents the traffic needed for retransmission if D2D sharing fails. Similarly, the downlink traffic is

$$f_d = \sum_n \sum_k \left[(1 - \phi_k)\delta_{nk} s_k + \phi_k (1 - B_{nk}(\rho_k))\delta_{nk} s_k\right] \tag{2}$$

Since the goal is to minimize f_b and f_d simultaneously, the two objective functions are combined using scalarization method [12]. Thus, we obtain

$$\underset{\{\theta_k, \phi_k\} \in \{0,1\}}{minimize}\ f = w f_b + (1 - w) f_d \tag{3}$$

where $w \in [0, 1]$, which stands for the importance of the objective function f_b. That is, larger w means larger cost of backhaul traffic f_b in the system.

The primary resources in cellular networks we take into consideration are the caching resources M and $\{M_n\}_{n=1}^N$, and the radio spectrum W with utilization constraints. As the total cached media should not exceed the finite cache memory size at the BS, we have $\sum_k \theta_k s_k \leq M$. Similarly, since successful D2D sharing means the media is stored at least at an user device, we also have the constraint $\sum_n \sum_k \delta_{nk} \phi_k B_{nk}(\rho_k)) s_k \leq \sum_n M_n$. Moreover, the bandwidth utilized for D2D social media sharing should be less than or equal to the fraction of cellular bandwidth for D2D communication, and the downlink traffic should not

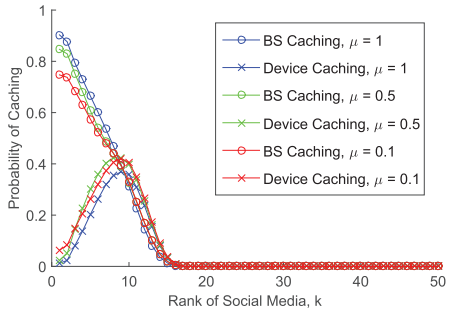

Fig. 2. Optimal strategy of caching utilization in the RAN. The strategy is to cache the most popular social media in the RAN; however, with respect to sizes, the utilization is different for caching in BS and caching in devices.

exceed the remaining fraction of the total cellular bandwidth, yielding

$$\sum_n \sum_k \phi_k B_{nk}(\rho_k)\delta_{nk} s_k (1+x) \le \eta W \gamma T, \text{ and}$$

$$\sum_n \sum_k \delta_{nk} \left[(1-\phi_k)s_k + \phi_k(1-B_{nk}(\rho_k))s_k\right] \le (1-\eta) W \gamma T$$

where $0 \le x \le 1$ is the percentage of previous controlling transmission volume comparing to the size; T is the observed period; γ is spectrum efficiency. Finally, as we do not consider any network coding scheme for social media, an user could only access the media could be either from the BS or from other users; thus we have $\theta_k + \phi_k \le 1, \ \forall k$.

This is exactly a Mixed-Integer Programming (MIP) problem, and can be precisely solved by Branch-and-Bound (B&B) type algorithm, which is commonly used for discrete combinatorial optimization problems.

III. Main Results

A. Optimal Strategy for Social Media

In the simulation, N is set to be 100, which is the number of active users supported by a BS. There are $K = 50$ different social media requested in this system. For the assumptions of social media, $q = 1$ and $\alpha = 2$ are chosen, as a common set of parameters [10]. The available bandwidth in a cell is $W = 20$ MHz with spectrum partition factor $\eta = 0.05$. We consider the observed period $T = 300$ seconds and the spectrum efficiency $\gamma = 16.32$ for 4×4 MIMO. The cache size at the BS is $M = 50$ Memory Units (MUs, *e.g.* MBytes. The MU provides a scalable model if we adjust the observed period T and K.) and $M_n = 0.5$ MU, $\forall n$. For the existence of D2D sharing links, we assume that $\rho = 0.1$, $c = 1$, and $x = 0.05$. In Fig. 2, we consider the backhaul and downlink traffic to be equally important ($w = 0.5$) and μ varies from 0.1 to 1 with the same dimension as MU and σ fixed at 1.5. We randomly generate 10,000 realizations of the network topology, accumulate each vector $\{\theta_k\}_{k=1}^K$, $\{\phi_k\}_{k=1}^K$, and show the probability that social media of rank k (k^{th} popular social media) is cached by the BS or shared by D2D communication. In the figure, it is clear that for the popular social media (high rank), the best strategy is to cache it in the BS. Intuitively, this strategy saves most of the traffic demands in the backhaul system and hence reduce the time needed for retrieving social media from remote data centers via core network [6]. Due to the instability of opportunistic links (*i.e.* $G(N, \rho)$) between devices, the result alternatively suggests that for the remaining social media with higher popularity and small size, they should take the chance by direct sharing via D2D communication. That is, sharing popular small-sized social media is the most efficient way to capitalize on caching in devices. By D2D sharing, the backhaul and cellular downlink traffic are saved and most importantly, the latency can be much more reduced than caching in BS, as discussed in the next subsection.

To sum up, the strategies for caching in BS and for caching in devices collaboratively facilitate distribution of social media in cellular networks, since we view caching as a general resource embedded in the RAN rather than consider caching in BS and in devices independently. For different pairs of (w_1, w_2), the results are similar. We summarize the following claims supplying heuristic guidelines of caching of social media in infrastructure and devices:

Claim III.1 (Caching in infrastructure): Cache the most popular social media in the BS, *e.g.* eNBs in LTE.

Claim III.2 (Caching in devices): Cache and share the remaining social media of high popularity rank; whereas, different from caching in BS, it is desirable to share social media of smaller content size, due to the probabilistic connections of D2D communication.

B. Optimal Flow and Latency Performance

We have shown the optimal strategy for social media caching; subsequently we focus on the minimizing traffic flow and the according latency performance. In Fig. 3(a), the minimized flow is illustrated for $\eta = 0$ to 0.02. As $M = 0$, the utilization of caching in devices accounts for the entire saving of traffic. Thus, the more bandwidth allocated to device sharing, the more traffic could be saved, provided the partition of bandwidth for cellular downlink traffic is enough. It is clear that caching, no matter in infrastructures or in devices, serves as the trade-off with spectrum resources.

In Fig. 3(a), the traffic saving apparently comes from caching in BS, and the contribution of device caching (The lines of *No caching in BS*) to overall traffic reduction appears to be limited. However, significant latency improvement is observed due to the collaboration of caching in the RAN, as shown in Fig. 3(b). To calculate the latencies, we set the latency of the data packet according to the latency requirements for LTE system [2]. The latency in RAN is 10 ms; the latency from eNB to P-GW is 20 ms, and the latency for direct device connection is 5 ms [13]. The significant elimination of latency comes from the fact that social media are distributed either by the BS or by devices in accordance with different characteristics (*i.e.* popularity and size) to increase the successful delivery probability. In other words, if all the caching resources in RAN are integrally utilized, the latency is greatly reduced near the optimum obtained by (3), because of such collaborative resource allocation.

C. Engineering Implementation

The collaborative resource allocation on the entire caching resources in the RAN can be implemented using existing but

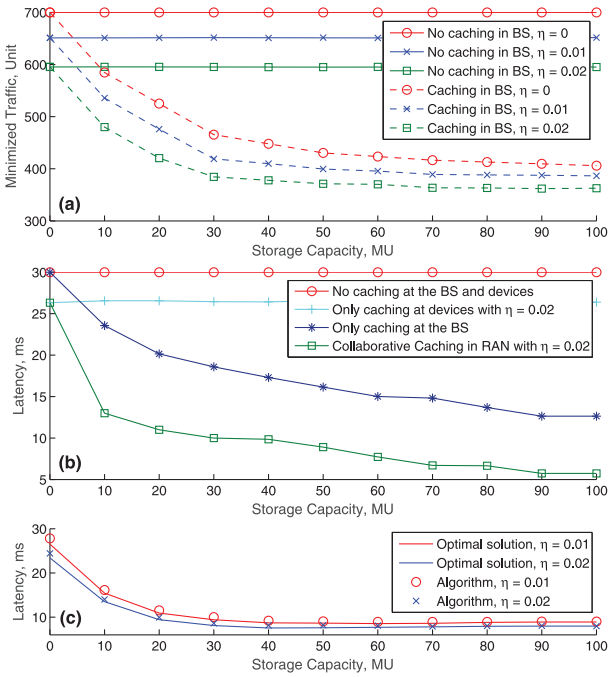

Fig. 3. (a) The minimized flow for different pairs of (M, η). It is obvious that large storage capacity in infrastructures and utilization of caching in devices (larger η) eliminate the traffic load for the RAN. (b) The latency performance. When we unify the caching resources in the RAN, the latency is reduced beyond the performance before the unification. (c) Shows that the latency performance of the algorithm and the optimal solution are closed, but with much lower complexity. It suggests that the algorithm provides suboptimal solutions.

not fully exploited function in present LTE systems. In the current structure, an entity called PCRF supplies the necessary information to make policies and charging decisions (Fig. 1). Two functions exist in the PCRF: the Policy Decision Function (PDF) and the Charging Rules Function (CRF), which perform accurate business activities based on the information of user data from the data center. It actually bridges the gap between content-aware services (*i.e.* charging) and the communication system. Together with PDF and CRF, the traffic flow of social media can be reasonably monitored, learned, and sorted [14]; that is, the sizes and popularities of the active social media can be obtained as two vectors $\mathbf{s}$ and $\mathbf{p} \in \mathbb{R}^K$, where K is the number of active social media in the RAN during a period of time. Therefore, we propose a new decision making entity (Fig. 1) in the core network for deciding the optimal strategy about caching. In addition, the optimization problem, suffering from high computational complexity, can be simply carried out by an algorithm of low complexity (**Algorithm 1**) based on the greedy nature inspected from Fig. 2.

In addition, device connectivity ρ is needed, and can be obtained by uncomplicated sensing of mobile users. A user device u_i senses in physical multicast channel (PMCH) and detects that it has $a_i \in \mathbb{N}$ connections, then an estimator for ρ appears to be $\hat{\rho} = \frac{\sum_{i=1}^{N} a_i/2}{N(N-1)/2}$, where $N(N-1)/2$ is the number of all possible links, and the estimator is the minimum variance unbiased estimator (MVUE) for the Bernoulli parameter ρ. With $\mathbf{s}$, $\mathbf{p}$, N, and ρ, the proposed new entity enables the algorithm and give the suboptimal strategy for resource allocation

Algorithm 1. The Greedy Algorithm

Input: ρ, N, and s_k, p_k, $\forall k$
Output: θ_k and ϕ_k, $\forall k$

```
cache, bandwidth ← 0
for k = 1 to k = K do                    ▷ Scan social media {c_k}
    if cache ≤ M then
        Cache c_k                        ▷ cache ← cache + s_k
for k = 1 to k = K do                    ▷ Scan social media {c_k}
    if c_k is not cached in eNB & bandwidth ≤ ηW then
        Share c_k via D2D communication
```

of caching. The performance of the proposed algorithm approximates that of the optimization in (3), while featuring low complexity, as shown in Fig. 3(c).

Subsequently, the BS, after receiving the suboptimal strategy from the on-line algorithm in the core network, could periodically broadcast a signal to the UEs by physical broadcast channel (PBCH) to tell them the which social media are suitable for direct device sharing. Upon receiving the signal, the UEs will attempt to access the social media by D2D communication first [13]. If D2D communication fails, the device will alternatively access the content from the BS. This completes the implementation of unifying caching resources in the RAN in present cellular systems.

REFERENCES

[1] Ericsson. (2015). *Ericsson Mobility Report* [Online]. Available: http://www.ericsson.com/mobility-report
[2] 3GPP. (2015). *3GPP Specification Series* [Online]. Available: http://www.3gpp.org/dynareport/36-series.htm
[3] I. V. Loumiotis *et al.*, "Dynamic backhaul resource allocation: An evolutionary game theoretic approach," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 691–698, Feb. 2014.
[4] S.-C. Lin and K.-C. Chen, "Improving spectrum efficiency via in-network computations in cognitive radio sensor networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1222–1234, Mar. 2014.
[5] Z. Ming, M. Xu, and D. Wang, "Incan: In-network cache assisted eNodeB caching mechanism in 4G LTE networks," *Comput. Netw.*, vol. 75, pp. 367–380, 2014.
[6] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.
[7] X. Zhuo, Q. Li, G. Cao, Y. Dai, B. Szymanski, and T. L. Porta, "Social-based cooperative caching in DTNS: A contact duration aware approach," in *Proc. Int. Conf. Mobile Adhoc Sens. Syst. (IEEE MASS)*, 2011, pp. 92–101.
[8] J. Cho, S. Oh, J. Kim, H. H. Lee, and J. Lee, "Neighbor caching in multi-hop wireless ad hoc networks," *IEEE Commun. Lett.*, vol. 7, no. 11, pp. 525–527, Nov. 2003.
[9] M. Newman, *Networks: An Introduction*. London, U.K.: Oxford Univ. Press, 2010.
[10] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, 1999, 126–134.
[11] P. Sobkowicz, M. Thelwall, K. Buckley, G. Paltoglou, and A. Sobkowicz, "Lognormal distributions of user post lengths in internet discussions—A consequence of the Weber-Fechner law?" *EPJ Data Sci.*, vol. 2, no. 1, pp. 1–20, 2013.
[12] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*. Reading, MA, USA:Addison-Wesley, 1973, vol. 28.
[13] W. Sun, E. G. Strom, F. Brannstrom, Y. Sui, and K. C. Sou, "D2D-based V2V communications with latency and reliability constraints," in *Proc. IEEE GLOBECOM Workshops*, 2014, pp. 1414–1419.
[14] E. Baştuğ, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," [Preprint]. arXiv:1503.05448, 2015.