Caching in Mobile HetNets: A Throughput–Delay Trade-off Perspective

Trung-Anh Do†, Sang-Woon Jeon‡, and Won-Yong Shin†
†Computer Science and Engineering, Dankook University, Yongin 448-701, Republic of Korea
‡Information and Communication Engineering, Andong National University, Andong 760-749, Republic of Korea
Email: dotrunganh@dankook.ac.kr; swjeon@anu.ac.kr; wyshin@dankook.ac.kr

Abstract—We analyze the optimal throughput–delay trade-off in content-centric mobile heterogeneous networks (HetNets), where each node moves according to the random walk mobility model and requests a content object from the library independently at random, according to a Zipf popularity distribution. Instead of allowing access to all content objects at base stations (BSs) via costly backhaul, we consider a more practical scenario where mobile nodes and BSs, each having a finite-size cache space, are able to cache a subset of content objects so that each request is served by other mobile nodes or BSs via multihop transmissions. Under the protocol model, we characterize a fundamental throughput–delay trade-off in terms of scaling laws by introducing our content delivery routing protocol and the corresponding optimal caching allocation strategy.

I. INTRODUCTION

Recently, data caching [1], which brings contents closer to users, has emerged as a promising technique that deals with the exponential growth of internet traffic. Thanks to the benefits of replicating popular contents across networks, caching techniques play an important role in maintaining the sustainability of future wireless networks.

As the number of users continues to grow dramatically, the capacity scaling law behavior has been widely studied in large wireless networks. In [2], it was shown that, for a static ad hoc network consisting of \( n \) randomly distributed source–destination pairs in the network area, the per-node throughput of the order of \( \sqrt{n \log n} \) is achievable using the nearest neighbor multihop transmission. Besides the multihop scheme, there have been various research directions to improve the per-node throughput up to a constant scaling by using hierarchical cooperation [3] and node mobility [4], [5].

Contrary to the studies on the conventional ad hoc network model in which source–destination pairs are given and fixed, investigating content-centric ad hoc networks would be quite challenging. As content objects are cached by numerous nodes over a network, finding the closest content holder of each request and scheduling between requests are of crucially importance for overall network throughput. The scaling behavior of content-centric ad hoc networks has received a lot of attention in the literature [6]–[9]. In static ad hoc networks, throughput scaling laws were analyzed using multihop communication [6], [7]. More specifically, a centralized caching allocation was presented in [6], while a decentralized random caching allocation was introduced in [7] yielding a significant performance gain over the single-hop caching scenario. In [8], throughput and delay performances were investigated for mobile ad hoc networks under various mobility models by showing that increasing the mobility degrees of nodes leads to a worse performance. Such analysis was extended to mobile heterogeneous networks (HetNets) [9] by introducing a caching allocation strategy, where each base station (BS) connected to the core network via infinite-speed backhaul is assumed to have an access to all content objects, or equivalently to be equipped with the infinite-size cache.

In this paper, we consider a content-centric mobile HetNet, where each node moves according to the random walk mobility model (RWMM) and requests a content object from the library independently at random, according to a Zipf popularity distribution. Instead of the infinite-speed backhaul-aided cache (or infinite-size cache), we assume that each mobile node and BS is equipped with a finite-size cache and is able to cache content objects in the library, where each BS has a relatively large-size cache. Under the protocol model, we characterize a fundamental throughput–delay trade-off in terms of scaling laws. By introducing our content delivery routing protocol and the corresponding optimal caching allocation strategy (i.e., the optimal content replication strategy), we find the optimal trade-off according to the identified operating regimes with respect to scaling parameters. When the total cache space at all BSs is greater than that at all nodes, our result indicates that popular contents are stored mainly in node caches while any request for less popular contents is fulfilled by BSs. The detailed description and all the proofs are omitted due to page limit.

II. NETWORK MODEL AND PROBLEM FORMULATION

A. Network Model

We consider a content-centric mobile HetNet consisting of \( n \) mobile nodes and \( f(n) = \Theta(n^\delta) \) static BSs, where \( 0 \leq \delta < 1 \). We assume that \( n \) nodes are distributed uniformly at random over a unit area and move according to the RWMM [5], and \( f(n) \) BSs are regularly placed over the same area. Each node and BS are assumed to be equipped with local caches, which are installed to store a subset of content objects.

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2014R1A1A2054577) and by the Ministry of Science, ICT & Future Planning (MSIP) (2015R1A2A1A15054248).

978-1-5090-1805-5/16/$31.00 ©2016 IEEE
in the library of size \( M = \Theta(n^\gamma) \), where \( 0 \leq \gamma < 1 \). Every content object is assumed to have the same size. In particular, each node and BS are able to cache at most \( K_n = \Theta(1) \) and \( K_{\text{BS}} = \Theta(n^\beta) \) content objects in their own finite-size cache space respectively, where \( 0 \leq \beta < \gamma \), which is well-suited to the deployment scenario of a massive number of small-cell BSs with very low operating costs. Besides, we assume that the total cache size in the mobile HetNet scales no slower than the pure ad hoc network with no BSs, i.e., \( \delta + \beta \geq 1 \), in order to analyze the impact and benefits of BSs equipped with a relatively large-size cache.

We assume that, at each time slot, every node requests its content object independently according to a Zipf popularity distribution, which typically characterizes a popularity of various kinds of real data such as web, file sharing, user-generated content, and video on demand. That is, the request probability of content object \( m \in \mathcal{M} \equiv \{1, \ldots, M\} \) is given by \( p_m = \frac{m^{-\alpha}}{H_\alpha(M)} \), where \( H_\alpha(M) = \sum_{i=1}^M i^{-\alpha} \) and \( \alpha > 0 \) is the Zipf exponent.\(^1\)

We consider the caching phase. Let \( A_m \) and \( B_m \) denote the number of replicas of content object \( m \in \mathcal{M} \) stored at nodes and BSs respectively, which will be optimized later. In order for a caching allocation to be feasible, \( \{A_m\}_{m=1}^M \) and \( \{B_m\}_{m=1}^M \) should satisfy

\[
\begin{align*}
\sum_{m=1}^M A_m &\leq nK_n, \\
\sum_{m=1}^M B_m &\leq f(n)K_{\text{BS}}.
\end{align*}
\]

Furthermore, we impose the following individual caching constraints:

\[
\begin{align*}
A_m &\leq n, \\
B_m &\leq f(n), \\
A_m + B_m &\geq 1
\end{align*}
\]

for all \( m \in \mathcal{M} \). Then, \( A_m \) and \( B_m \) replicas of content object \( m \) can be stored at the caches of \( A_m \) randomly chosen distinct nodes and \( B_m \) randomly chosen distinct BSs, respectively. Note that the last constraint in (2) is needed to avoid an outage event that a requested content object is not stored in the entire network.

Now, let us consider the delivery phase of content objects. At each time slot, each node downloads its requested content object (possibly via multihop) from one of the nodes or BSs storing the requested content object in their caches. We assume the protocol model in [2] for successful content delivery. In particular, let \( d(u,v) \) denote the Euclidean distance between nodes \( u \) and \( v \). Then content delivery from node \( u \) to node \( v \) is successful if and only if \( d(u,v) \leq r \) and there is no other active transmitter in a circle of radius \( (1 + \Delta)r \) from node \( v \), where \( r, \Delta > 0 \) are given protocol parameters. For analytical tractability, we also adopt the fluid model in [5]. In this model, the size of each content object is assumed to be arbitrarily small. Accordingly, the time required for content delivery between a node and its neighbor node or an assigned BS is much smaller than the duration of each time slot. Thus, all content objects waiting for transmission at a node will be transmitted by the node in one time slot. However, a content object received by a node at a given time slot cannot be transmitted by the node until the next time slot.

### B. Performance Metrics

For the content-centric mobile HetNet, an achievable scheme is composed of a sequence of policies \( \{\pi_n\} \), which determines the caching allocation and the scheduled communication that take place in each time slot. The throughput and delay are then defined as follows [9]:

**Definition 1** (Throughput). Let \( B_{\pi_n}(i,t) \) denote the total number of bits of all content objects received by node \( i \) during \( t \) time slots under the policy \( \pi_n \). Then, the average per-node throughput is defined as \( \lambda(n) = \mathbb{E}\left[ \frac{1}{n} \sum_{i=1}^n \liminf_{t \to \infty} \frac{1}{t} B_{\pi_n}(i,t) \right] \).

**Definition 2** (Delay). Let \( D_{\pi_n}(i,k) \) denote the delay of the \( k \)th requested content object of node \( i \) under the policy \( \pi_n \), which is measured from the moment that the requesting message leaves node \( i \) until the corresponding content object arrives at the node from the closest holder. The delay over all content requests for node \( i \) is \( \limsup_{r \to \infty} \frac{r}{\sum_{i=1}^n D_{\pi_n}(i,k)} \). Then, the average delay is defined as \( D(n) = \mathbb{E}\left[ \frac{1}{n} \sum_{i=1}^n \limsup_{r \to \infty} \frac{r}{\sum_{k=1}^r D_{\pi_n}(i,k)} \right] \).

### III. CONTENT DELIVERY ROUTING PROTOCOL

In this section, we describe our routing protocol to deliver content objects to requesting nodes. For multihop routing, the network of unit area is divided into \( a(n)^{-1} \) square routing cells of equal size, where \( a(n) = \Omega\left(\log\frac{n}{n}\right) \) and \( a(n) = O(1) \), so that each routing cell has at least one node with high probability (wp) [2]. The network is also divided into \( f(n) = b(n)^{-1} \) square cells of equal size so that each cell has one BS at its center.

Assume that multihop routing (topology discovery and data reachability) has already been accomplished in the network, so that each node knows to whom it can forward a requesting message in order to reach the desired content holder. Due to the node mobility, our routing protocol is built upon the routing scheme in [5], where packets chase their destinations, and reconstructed for our cache-enabled setting accordingly.

We can implement a multihop strategy in our content delivery by using other nodes as relays. Each routing cell of size \( a(n) \) or \( b(n) \) is activated regularly once every \( 1 + c \) time slot to avoid any collision, where \( c > 0 \) denotes a small constant independent of \( n \). Our content delivery routing protocol operates depending on the network topology, i.e., the initial distance between a requesting node and its closest holder. Let \( L_{H,R} \) denote the straight line connecting a requesting node and its closest holder, which can be a mobile node or a BS. Accordingly, a requesting message is delivered to the closest holder along the adjacent cells intersecting \( L_{H,R} \) via multihop
in forward direction, which corresponds to the first phase of the content delivery. Similarly, the desired content object hops in backward direction to the requesting node along the cells intersecting another $L_{H,R}$ due to the node mobility, which corresponds to the second phase. As a special case where the requesting node is inside the transmission range of any holder of the desired content object, the request will be responded by using a single-hop strategy in one time slot. Otherwise, the content delivery procedure consists of the following two steps:

**Step 1)** First delivery phase

a) If the closest holder is a node, then the requesting message is generated and chases the target node according to the following procedure. As depicted in Fig. 1(a), from cell $C^0$, the requesting message is transmitted via multihop along the adjacent cells toward cell $C^1$ containing the target node, where the per-hop distance is given by $\Theta\left(\sqrt{a(n)}\right)$. By the time the requesting message reaches cell $C^1$, the target node has moved to another position $C^2$ according to the RWMM. Thus, the message hops from cell $C^1$ to cell $C^2$. This continues until the message reaches the cell $C^3$ containing its target node.

b) If the closest holder is a BS, then the requesting message is delivered via multihop along the adjacent cells intersecting the straight line toward the coverage of the target BS, where the per-hop distance is given by $\Theta\left(\sqrt{b(n)}\right)$. When the requesting message arrives inside the cell coverage $b(n)$ of the target BS in cell $C^1$, the last relay node will send the message to the BS immediately using single-hop in one time slot, where the last hop distance to the target BS is given by $\Theta\left(\sqrt{b(n)}\right)$. Performing this long-distance hop for the last hop to the BS is the best we can hope for (which will be specified later). We refer to Fig. 1(b).

**Step 2)** Second delivery phase

a) By the time the target node receives the requesting message, the requesting node has moved to another position $C^4$ according to the RWMM. As illustrated in Fig. 2(a), the desired content object generated by the target node chases the requesting node by executing essentially the same procedure as the first delivery phase.

b) By the time the BS receives the requesting message, the requesting node has moved to another position $C^3$. As illustrated in Fig. 2(b), the desired content object generated by the BS is delivered to a relay node in cell $C^2$ along another $L_{H,R}$ toward the requesting node in one time slot. Thereafter, the relay chases the requesting node via multihop until reaching cell containing the requesting node.

Each time slot is divided into two sub-slots. The first and second phases of the content delivery procedure are activated during the first and the second sub-slots, respectively.

**IV. THROUGHPUT–DETERMINISTIC DELAY TRADE-OFF**

In this section, we characterize a fundamental throughput–delay trade-off in terms of scaling laws for the content-
centric mobile HetNet using the proposed content delivery routing. In cache-enabled networks, content objects tend to be generally pushed closer to users. Thus, the total number of replicas of content object \( m \in M \), \( A_m + B_m \), is a crucial factor which determines how far from a requesting node to its closest holder of the content \( m \). Similarly as in [8] and [9], from the fact that replicas of each content object are independently and uniformly distributed over the network, the average Euclidean distance from a requesting node to its closest holder of content \( m \) scales as the inverse square root of \( A_m + B_m \). By applying this argument to our framework, the requesting message is generated to be delivered to the closest holder with the initial distance \( \Theta \left( \frac{1}{\sqrt{A_m + B_m}} \right) \). As long as the transmission range \( \Theta \left( \sqrt{b(n)} \right) \) of each BS is less than the initial distance \( \Theta \left( \frac{1}{\sqrt{A_m + B_m}} \right) \), the value of \( b(n) \) does not influence the delay scaling law. For this reason, in our work, we set \( b(n) = \frac{1}{10 \sigma} \) as a typical cellular setting.

Let \( H_m \) denote the number of hops along the routing path between a requesting node and its closest holder of the content \( m \in M \). Then, for all \( m \in M \), it obviously follows that

\[
\mathbb{E}[H_m] = \Theta \left( \frac{1}{\sqrt{n \mathcal{a}(n)(A_m + B_m)}} \right).
\]

Let us consider the case where \( A_m \geq n \mathcal{a}(n)^{-1} \) and \( B_m = f(n) \). In this case, a node requesting the content \( m \) is able to find the desired content object in the same routing cell of size \( a(n) \) and \( b(n) \) whp, which thus corresponds to \( \mathbb{E}[H_m] = 1 \). Now, we establish our first main result.

**Theorem 1.** Suppose that the content delivery routing in Section III is used for the content-centric mobile HetNet. Then, the throughput–delay trade-off is given by

\[
\lambda(n) = \Theta \left( \frac{D(n)}{n \left( \sum_{m=1}^{M} \frac{p_m}{\sqrt{A_m + B_m}} \right)^2} \right),
\]

where

\[
\lambda(n) = O \left( \frac{1}{\sum_{m=1}^{M} p_m \sqrt{\frac{n \log n}{A_m + B_m}}} \right).
\]

Theorem 1 implies that the throughput–delay trade-off is influenced not by the per-hop distance but by the distributed caching space of each content object, \( A_m + B_m \). Due to the caching size constraints in (1) and (2), it is not straightforward how to optimally allocate the sets of replicas, \( \{A_m\}_{m=1}^{M} \) and \( \{B_m\}_{m=1}^{M} \), to show a net improvement in the overall throughput–delay trade-off. In the next section, we find the optimization caching allocation strategy to characterize the optimal throughput–delay trade-off.

**V. OPTIMAL CACHING ALLOCATION STRATEGY IN MOBILE HETNETS**

In this section, we characterize the optimal throughput–delay trade-off of the content-centric mobile HetNet by optimally selecting the replication sets \( \{A_m\}_{m=1}^{M} \) and \( \{B_m\}_{m=1}^{M} \). In our work, we mainly focus on the case where \( \alpha < 3/2 \). This is because that the optimal throughput–delay trade-off, \( \lambda(n) = \Theta(D(n)) \) where \( \lambda(n) = O(n^{-\epsilon}) \) for an arbitrarily small constant \( \epsilon > 0 \), is always achieved by only using node-to-node multihop communication for \( \alpha \geq 3/2 \). This corresponds to a special configuration of our HetNet with \( K_{BS} = 0 \) [8], [9].

From Theorem 1, it is seen that maximizing performance on the throughput and delay is equivalent to minimizing the term \( \sum_{m=1}^{M} \frac{p_m}{\sqrt{A_m + B_m}} \) in (3). Thus, we formulate the following constrained optimization problem:

\[
\min_{\{A_m\}_{m=1}^{M}, \{B_m\}_{m=1}^{M}} \sum_{m=1}^{M} \frac{p_m}{\sqrt{A_m + B_m}} \quad (4a)
\]

subject to:

\[
\begin{align*}
M & \sum_{m=1}^{M} A_m \leq n K_n, \quad (4b) \\
M & \sum_{m=1}^{M} B_m \leq f(n) K_{BS}, \quad (4c) \\
A_m & \leq n \text{ for } m \in \{1, \ldots, M\}, \quad (4d) \\
B_m & \leq f(n) \text{ for } m \in \{1, \ldots, M\}, \quad (4e) \\
A_m + B_m & \geq 1 \text{ for } m \in \{1, \ldots, M\}. \quad (4f)
\end{align*}
\]

For analytical convenience, given the optimal solution \( \{A_m^*\}_{m=1}^{M} \) and \( \{B_m^*\}_{m=1}^{M} \), we define two subsets \( M_1 \) and \( M_2 \), where \( M_1 \) and \( M_2 \) indicate the sets of content objects such that \( A_m^* + B_m^* = \Theta(A_m^*) \) and \( A_m^* + B_m^* = \Omega(f(n)) \), respectively, which will be specified in the following lemma.

**Lemma 1.** For \( \alpha < \frac{3}{2} \), the subsets \( M_1 \) and \( M_2 \) are given by

\[
M_1 = \{1, \ldots, m_1 - 1\},
\]

\[
M_2 = \{1, \ldots, m_2 - 1\},
\]

where \( m_1 = \Theta \left( n^{\frac{2(3 - \alpha)}{2(\alpha + 1)}} \right) \) and \( m_2 = \Theta \left( n^{\gamma - (\gamma - \frac{3}{2})} \right) \).

From the above lemma, one can solve the problem (4) by optimally finding the total number of replicas of content object \( m \in M \) at both nodes and BSs, i.e., \( \{A_m^* \}^{M}_{m=1} \) and \( \{B_m^* \}^{M}_{m=1} \). Now, we propose our replication strategy by individually choosing the replication sets \( \{A_m^*\}_{m=1}^{M} \) and \( \{B_m^*\}_{m=1}^{M} \). From the following proposition, our strategy is shown to guarantee the optimality as long as scaling laws are concerned.

**Proposition 1.** Suppose that the content delivery routing in Section III is used for our content-centric mobile HetNet. For \( \alpha < \frac{3}{2} \), the proposed optimal replication strategy is then expressed as

\[
A_m^* = \Theta \left( m^{\frac{3(\gamma - (\beta - \frac{3}{2}))}{2(\beta - \frac{3}{2})}} \right) \quad \text{if } m \in M_1
\]

\[
A_m^* = \Theta \left( \min \{m^{\frac{3(\gamma - (\beta - \frac{3}{2}))}{2(\beta - \frac{3}{2})}}, m^{\frac{3(\gamma - (\beta - \frac{3}{2}))}{2(\beta - \frac{3}{2})}} \} \right) \quad \text{if } m \in M \setminus M_1
\]
correspond to the cases where the terms $F_1$ and $F_3$ in (6) are dominant, respectively.

Note that in Regime I leading to the best throughput–delay trade-off, (4a) scales as $\frac{1}{\lambda}$ by using the optimal caching strategy in Proposition 1. Now, we are ready to characterize the optimal throughput–delay trade-off in the following theorem.

**Theorem 2.** Suppose that the content delivery routing in Section III and the optimal replication strategy in Proposition 1 are used for our content-centric mobile HetNet. Then, according to the Zipf exponent $\alpha$ and the scaling parameters $\gamma$, $\delta$, and $\beta$, the optimal throughput–delay trade-off is given by

$$\lambda(n) = \Theta \left( \frac{D(n)}{n^\beta} \right),$$

for an arbitrarily small constant $\epsilon > 0$. Here,

$$b = \begin{cases} 
0 & \text{in Regime I} \\
(1 - \delta)(\frac{3}{2} - \alpha) & \text{in Regime II} \\
1 - \delta - \beta + \min \{3 - 2\alpha, 1\} & \text{in Regime III.} 
\end{cases}$$

The impact and benefits of BSs equipped with a finite-size cache are explicitly addressed according to each operating regime on the throughput–delay trade-off.

**Remark 2.** In Regime I, the best performance $\lambda(n) = \Theta(D(n))$ is achieved by using the node-to-node multihop routing, and thus the use of BSs does not further improve the performance. On the other hand, in Regimes II and III, it turns out that the supplemental space $f(n) \frac{K_{BS}}{n^\beta}$ in our mobile HetNet significantly improves the network performance over the mobile ad hoc network scenario with no BSs [8], [9]. Interestingly, in Regime II, the optimal trade-off is shown to be the same as in the cache-enabled mobile HetNet case assuming BSs equipped with the infinite-size cache [9].

---

**REFERENCES**


