Capacity of Cache Enabled Content Distribution Wireless Ad Hoc Networks∗†‡

Benyuan Liu∗ Victor Firoiu† Jim Kurose‡ May Leung† Soumendra Nanda†

∗Department of Computer Science, University of Massachusetts, Lowell, MA, USA 01854
†Advanced Information Technologies, BAE Systems, Burlington MA, USA 01803
‡Department of Computer Science, University of Massachusetts, Amherst, MA, USA 01002

Abstract—While wireless ad hoc networks have a wide range of applications in environment monitoring, military operations, and disaster recovery, etc, the full potential of such networks is hindered by their diminishing capacity behavior as the network size scales up. Content caching has been previously proposed to improve the availability of contents in a network and thus helps to alleviate the load on content custodians, reduce access latency, and improve the network capacity. This paper studies the scaling laws of the capacity of cache-enabled content distribution wireless ad hoc networks. We consider two basic content access schemes, namely, the Nearest Caching Node scheme where a request is satisfied by the nearest node to the requestor that has the content in its cache, and the Transparent Enroute Caching scheme where a content request is routed towards the content custodian and is satisfied by an intermediate node (or custodian) along the path that has the content in its cache. We first establish the capacity of content distribution wireless ad hoc networks without content caching as a baseline for investigating the benefit of caching. We then obtain the scaling laws of the capacity for the above two content access schemes with content caching. Based on the results we further explore their design and performance implications. Our results show that the capacity exhibits distinct scaling behaviors under different scaling regimes of the network parameters. Under certain conditions increasing the cache size of the nodes can effectively improve the capacity while under other conditions the improvement can be negligible. The characterizations of the capacity allows us to identify the bottleneck of the content access capacity for given network scenarios and choose effective approaches to improve the capacity.

I. INTRODUCTION

A wireless ad hoc network is formed by nodes in a self-organized manner to forward data packets on behalf of each other when there is no fixed infrastructure such as base stations. While wireless ad hoc networks have a wide range of applications in environment monitoring, military operations, and disaster recovery, etc, the full potential of these network applications is inherently hindered by their diminishing capacity behavior as the network size scales up [1]. Specifically, in a wireless ad hoc network of \( n \) nodes, under random node placement and traffic pattern assumptions, the per-node capacity decreases at a rate of \( \frac{1}{\sqrt{n \log n}} \) as the number of nodes goes to infinity.

Content distribution networking (CDN) has achieved tremendous commercial success on the Internet where popular contents are cached across the network to improve the performance of data delivery services. Similar techniques have been proposed and studied for wireless ad hoc networks. A content may be opportunistically cached at multiple nodes in the network and access requests to the content can be satisfied by a nearby node instead of the content custodian. As a result, the improved content availability helps to alleviate the load on the content custodian, reduce access latency, and improve the network capacity. While there have been many studies on the design of various cache placement schemes, the fundamental performance limit of the capacity of wireless ad hoc networks with content caching remains to be thoroughly investigated.

In this paper, we study the scaling behaviors of the capacity and delay of a cache-enabled content distribution wireless ad hoc network. We consider the following two fundamental content access schemes: (i) Nearest Caching Node scheme where a content request is satisfied by the nearest node that has the content in its cache, (ii) Transparent Enroute Caching scheme where a content request is routed towards the content custodian and satisfied by the first node along the path that has the content in its cache. If no intermediate node has the requested content, the request will finally reach and be served by the content custodian.

We are interested in the following research questions:

- How does the capacity scale with the various network parameters including the number of nodes, number of unique contents, and cache size of each node?
- How does the performance of these two schemes compare with each other and with the case without content caching?
- What are the design and performance implications of the results?

Answers to these questions will improve our understanding of the fundamental performance limit of cache-enabled content distribution wireless ad hoc networks, and will lead to better network design and planning tools for a range of
wireless ad hoc network applications. The contributions of this paper are summarized as follows.

We first establish the capacity of content distribution wireless ad hoc networks without content caching as a baseline for investigating the benefit of caching. We then obtain the scaling laws of the capacity for the above two content access schemes with content caching. Based on the results we further explore their design and performance implications.

For the Nearest Caching Node scheme, under the uniform content access pattern where each content is equally likely to be requested and cached in a node, the per-node capacity is upper-bounded by $\sqrt{C/B \log n}$ where $n$, $B$, and $C$ represent the number of nodes, number of unique contents, and cache-size of each node, respectively. The capacity increases with the cache-size of each node at a rate of $\sqrt{C}$, while at the same time the packet delay decreases at the same rate. An $m$-fold increase in cache size will result in a $\sqrt{m}$-fold increase or decrease in capacity and packet delay, respectively.

For the Transparent Enroute Caching scheme, the capacity is limited by two factors: the cache-size to content ratio ($C/B$) and the inherent network transport capacity constrained by the number of simultaneous transmissions that can be supported by the network $(1/\sqrt{n \log n})$. Depending on the scaling behavior of the number of unique contents ($B$), the capacity falls into one of the two scaling regimes. When the number of contents is asymptotically smaller than $C \sqrt{n \log n}$, the request is satisfied enroute by an intermediate node before it reaches the content custodian, and the capacity is limited by and increases linearly with the cache-size to content ratio, i.e., $C/B$. An $m$-fold increase in the cache size of each node will result in an $m$-fold gain in capacity.

On the other hand, when the number of unique contents is asymptotically larger than $C \sqrt{n \log n}$, a request is satisfied by the content custodian instead of an intermediate node almost surely. In this case the benefit of content caching is negligible and the capacity is inherently constrained by the number of simultaneous transmissions that can be supported by the network.

The characterizations of the capacity allows us to identify the bottleneck of the data access capacity for given network scenarios and choose proper approaches to improve the capacity. For example, under the Transparent Enroute Caching scheme, when the capacity is constrained by the cache-size to content ratio, it is beneficial to configure more memory for content caching. On the other hand, if the capacity is constrained by the network transport capacity, allocating more memory for caching will not result in substantial capacity improvement.

The rest of the paper is structured as follows. Section II provides an overview of the previous studies on the capacity and content caching of wireless ad hoc networks. Section III describes the network model used in our study. Section IV presents our analysis of the capacity and delay for cache-enabled wireless ad hoc networks. Finally, we conclude the paper in Section VI.

II. RELATED WORK

The capacity of a wireless network is a fundamental performance metric that characterizes the long-term throughput of each node that can be supported by the network. The scaling laws of the capacity are dictated by many different aspects of the network, for example, network architecture, radio interference model, node communication paradigm, traffic locality, node mobility, power and bandwidth constraints, etc.

The scaling law of the capacity of wireless ad hoc networks was first investigated by Gupta and Kumar in [1] and has since triggered a myriad of studies along this line of research. The gap between the upper and lower bounds of the capacity was later closed in [2] using a percolation theory approach. A constant capacity can be achieved by exploiting node mobility as a type of multiuser diversity [3]. The capacity of a wireless ad hoc network can be improved by adding infrastructure support [4], [5], [6], [7], [8]. The scaling behaviors of the capacity under a variety of different traffic patterns have also been investigated, including the locality of traffic, broadcast, multicast, relay traffic, many-to-one traffic, and a generalized many-to-many traffic model [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. The capacity is affected by hardware technologies and constraints such as power and topology constraints [20], [21], multiple-input multiple-output (MIMO) techniques [22], directional antennas [23], [24], number of channels [25], network coding [26], etc. For mobile ad hoc networks, there are intrinsic tradeoffs between the capacity and delay [27], [28], [29], [30], [31], [32], between the security and capacity [33], [34].

Content caching can significantly improve the data availability in a wireless ad hoc network, resulting in reduced delay and increased capacity. Many different caching schemes have been proposed and studied for wireless ad hoc networks. In [35], an aggregated caching mechanism and a broadcast-based simple search algorithm are proposed for improving the information accessibility and reducing average communication latency. In [36], the authors propose both centralized and distributed algorithms using a benefit-based approach for cache placement problem of minimizing total data access cost with multiple data items and nodes with limited memory capacity. Cooperative caching schemes [37], [38], [39], [40], [41] that allow the sharing and coordination of cached data among multiple nodes have been proposed to further explore the potential of caching techniques.

Although there have been many studies on the capacity analysis and caching algorithm design for wireless networks, the capacity of cache enabled content distribution ad hoc networks remains to be fully analyzed. To this end, [42] studied the performance of on-path caching for the content distribution in wireless network environment with GPSR
as the underlying routing protocol [43]. The asymptotic laws for joint content replication and delivery in wireless networks was investigated in [44]. In [45], the authors studied the throughput capacity of an information-centric network when the data cached in each node has a limited lifetime.

III. NETWORK MODEL

We consider a wireless ad hoc network where \( n \) nodes are distributed in a unit area square uniformly at random. There is a total number of \( B(n) \) unique pieces of content that may be requested by nodes in the network. For simplicity we assume all contents have the same unit size and each node has a cache size that can hold \( C(n) \) units of content. In practice, the cache size of a node is limited by the available memory that can be used for content caching. Thus \( C(n) \) is usually upper bounded by the memory size of a node and may be considered as a constant for practical purpose. In general the contents usually have different sizes. As mentioned in [44], the variable-sized contents can be split into multiple unit segments and each segment can be treated as separate contents and cached independently of each other. As the number of nodes in the network increases, the number of unique contents that nodes may be interested in usually increases as well. So we assume \( B(n) \) is an non-decreasing function of \( n \). For conciseness we denote \( B(n) \) and \( C(n) \) as \( B \) and \( C \) in the rest of the paper, respectively.

Each content has a custodian node that maintains a permanent copy of that content. The custodian of each content is randomly located among the \( n \) nodes in the network. Provided that a content requestor has a network path to the content custodian, a content request can always be satisfied from its custodian. As the requested content is downloaded, intermediate nodes as well as the requesting node may opportunistically cache a copy of the downloaded content, and subsequent requests may be satisfied by one of the caching nodes.

We consider the following two basic content access schemes.

- **Nearest Caching Node**: A request to a content is flooded to find and retrieve the content from the nearest node that has the content in its cache. The flooding overhead can be reduced using the expanding ring search mechanism where the hop-count-limit of the request messages is successively incremented at each round until a node with the content in its cache is found. Note that the size of a request message is usually much smaller than that of the content. Also, the content caching information (which node has what contents) can be maintained by nodes and subsequent requests can be directly routed to the nearest caching node without flooding. In this paper we assume the flooding overhead is negligible and do not take it into account in our analysis.

- **Transparent Enroute Caching**: A request to content is always routed towards the content custodian along the shortest path. When receiving a request, intermediate forwarding nodes check their local caches for a copy of the requested content. If the content is found, the request will be satisfied from that node, rather than from the custodian, and the node may begin transmitting the content to the requestor. In the worst case, no intermediate node has the content in its cache, and the request will finally reach and be satisfied by the content custodian.

These two content access schemes are illustrated in Figures 1 and 2, respectively.

![Figure 1. Nearest caching node: content request is routed to and satisfied by the nearest node that has the content in its cache.](image1)

![Figure 2. Transparent Enroute Caching: content request is routed toward custodian and satisfied by the first node that has the content in its cache.](image2)

In this paper we consider a uniform content access pattern where each node chooses a content uniformly at random to...
download. Each content is equally likely to be requested and downloaded by a node. While studies have shown that content popularity on the Internet is not uniform but typically follows a Zipf’s law [46], the access pattern of a wireless ad hoc network depends on specific application scenarios. There are ample examples in which the uniformly random content access model makes sense. For instance, the geographic hash table (GHT) mechanism has been proposed for data-centric storage in wireless sensor networks [47]. In this approach, sensor data is routed and stored at nodes rather uniformly across the network using a distributed hashing table system, resulting in a uniform data access pattern.

As a result of the uniform access pattern, it is reasonable to assume that each content is equally likely to be cached by a node in the steady state. Since there are $B$ unique contents and each node can hold $C$ contents in its cache, we assume that the probability that any given content is in the cache of a node is $C/B$. Since there are a total number of $n$ nodes in the network, each content is cached in the network uniformly at random with density $nC/B$.

All of the nodes share a common wireless channel. Each node has a common transmission range $r$. For simplicity, we assume each node can transmit data at a rate of 1 bit/sec over the channel when scheduled. A time-division multiplexing (TDMA) scheme is used for scheduling data transmission over the wireless channel. Time is divided into fixed duration slots. In each time slot, certain nodes are scheduled to send data. We adopt the Protocol Interference Model introduced in [1]. A transmission from node $X_i$ is successfully received by node $X_j$ if the following two conditions are satisfied: (i) Node $X_j$ is within the transmission range of node $X_i$, i.e., $|X_i - X_j| \leq r$, where $|X_i - X_j|$ represents the distance between node $X_i$ and node $X_j$ in the plane; (ii) For every other node $X_k$ that is simultaneously transmitting over the same channel, $|X_k - X_j| \geq (1 + \Delta)|X_i - X_j|$ for some $\Delta > 0$.

Throughout the paper we use $O(g(x))$, $\Omega(g(x))$ and $\Theta(g(x))$ to represent the asymptotic upper, lower and tight bounds of function $g(x)$.

IV. CAPACITY OF CONTENT DISTRIBUTION NETWORK WITHOUT CACHING

In this section we establish the capacity of content distribution ad hoc networks without caching, which will be used as a baseline for our later investigation of the benefit of caching for content distribution in the next section.

Theorem 1: Consider a wireless ad hoc network of $n$ randomly located nodes and $B$ unique contents. Assume each content is equally likely to be requested by a node. The scaling laws of the capacity ($\lambda$) and packet delay ($d$) for the number of transmission hops for content requests are characterized as follows:

- The per-node capacity is upper-bounded by

$$\lambda(n,B) = O\left(\frac{B}{\sqrt{n \log n}}\right)$$  

(1)

- The average packet delay (number of transmission hops) between a requestor and content custodian is

$$\bar{d}(n,B) = O\left(\frac{n}{\sqrt{\log n}}\right)$$  

(2)

Discussion of results. Similar to the discussions in [19], depending on the scaling behavior of the number of unique contents, $B$, the upper bound of capacity is constrained by different factors and exhibits different scaling laws. Specifically, we have the following two cases:

- $B = \omega(\sqrt{n \log n})$. When the number of unique contents is asymptotically larger than $\sqrt{n \log n}$, the upper bound becomes:

$$\lambda(n,B) = O\left(\frac{1}{\sqrt{n \log n}}\right).$$

In this case the traffic induced by the contents is enough to saturate the network transport capacity, and the content download capacity is limited by the network transport capacity, which recovers the results in Gupta and Kumar [1].

- $B = O(\sqrt{n \log n})$. When the number of unique contents is asymptotically smaller than $\sqrt{n \log n}$, the upper bound becomes:

$$\lambda(n,B) = O\left(\frac{B}{n}\right).$$

In this case the traffic induced by the contents does not saturate the network transport capacity, and the content download capacity is limited by and increases linearly with the number of contents. The network has enough transport capacity to support traffic to all of the custodian nodes.

We now present the proof of Theorem 1.

Proof. The capacity of wireless ad hoc networks with a varying number of source and destination nodes has been established in [19]. In particular, consider the scenario where there are $s(n)$ randomly selected source nodes and $d(n)$ randomly selected destination nodes, the per-node capacity is upper bounded by

$$\lambda(n,s(n),d(n)) = O\left(\min\left(\frac{d(n)}{s(n)}, \frac{\sqrt{n \log n}}{s(n)}\right)\right)$$  

(3)

When $B \leq n$, for upper bound we can consider the case where each of the $B$ unique contents resides on different node, and the above results can be directly applied. When $B > n$, the upper bound can be achieved by assuming the contents are evenly distributed among the $n$ nodes. Therefore, the theorem is proved by plugging $s(n) = n$, and $d(n) = \min(B, n)$ in equation (3).

As the source and destination nodes are independently and uniformly distributed in the unit area network, and each source node chooses the destination node uniformly at
random, the mean path length between a source-destination pair is \( \Theta(1) \). It has been shown in [49] that the transmission radius of a node, \( r(n) \), should be at least in the order of \( \sqrt{\log n/n} \) to guarantee that the network is connected as \( n \) goes to infinity. Thus, the average delay (number of transmission hops) between a source-destination pair is given by

\[
\bar{d}(n, B) = O\left(\frac{1}{r(n)}\right) = O\left(\sqrt{\frac{n}{\log n}}\right)
\]

\( \square \)

V. Capacity of Content Distribution Networks with Caching

In this section we present our results on the scaling laws of the capacity, discuss the design and performance implications of the results, and then provide the proof of the results for each of the two content access schemes.

A. Nearest Caching Node

We first present the main results for the Nearest Caching Node scheme.

**Theorem 2:** Consider a wireless ad hoc network of \( n \) randomly located nodes and \( B \) unique contents of unit size, with each node having a cache size of \( C \) units. Assume each content is equally likely to be requested and cached by a node. Under the Nearest Caching Node scheme, the scaling laws of the capacity and packet delay (number of transmission hops) for content requests are characterized as follows:

- The per-node capacity is upper-bounded by

\[
\lambda(n, B, C) = O\left(\sqrt{\frac{C}{B \log n}}\right) \tag{4}
\]

- The average packet delay between a requestor and the nearest caching node is

\[
\bar{d}(n, B, C) = O\left(\sqrt{\frac{B}{C \log n}}\right) \tag{5}
\]

**Discussion of results.** Based on the capacity result as shown in equation (4), for a fixed network size \( n \) and number of unique contents \( B \), if we can increase the cache size of each node \( C \), the capacity increases at a rate of \( \sqrt{C} \), while at the same time the packet delay decreases at the same rate. An \( m \)-fold increase in cache size will result in a \( \sqrt{m} \)-fold increase or decrease in capacity and packet delay, respectively.

When compared with the baseline scenario where there is no caching, the capacity gain due to content caching is given by the ratio of equations (4) and (1)

\[
g = \frac{\sqrt{C/B \log n}}{1/\sqrt{n \log n}} = \sqrt{\frac{nC}{B}}
\]

Based on the uniform content access assumption, the probability that a content is in the cache of any given node is \( C/B \). Thus, \( \rho = nC/B \) corresponds to the average number of cache copies for each content in the network, and the capacity gain of content caching increases as \( \sqrt{\rho} \).

As the network size grows, if the number of unique contents increases asymptotically slower than \( nC, B = o(nc) \), the gain goes to infinity as a function of \( n \).

Note that in the special case when \( B = \Theta(Cn) \), i.e., on average each content is only cached by a constant number of nodes. The average distance between a requestor and the nearest node that has the content in its cache is \( \bar{d}(n, B, C) = \Theta(\sqrt{n/\log n}) \), which is on the same order as the dimension of the unit square network. There is no performance gain by downloading the content from the nearest caching node. The capacity becomes

\[
\lambda(n, B, C) = O\left(\sqrt{\frac{C}{B \log n}}\right) = O\left(\frac{1}{\sqrt{n \log n}}\right)
\]

which recovers the Gupta and Kumar result [1] for the random wireless ad hoc networks.

We now present the proof of Theorem 2.

**Proof of Theorem 2.** In a wireless ad hoc network of \( n \) nodes each with a cache size of \( C \) units and \( B \) unique unit size contents, based on the uniform content access assumption, the probability that any given content is in the cache of any given node is \( C/B \), and hence each content is cached in the network uniformly at random with density \( \rho = nC/B \).

In the Nearest Caching Node scheme, a content request is satisfied by the nearest node that has the content in its cache, as shown in Figure 1. The distance between the requestor and the nearest node that has the content in its cache, denoted as \( X \), follows a circular contact distribution [48]. The corresponding probability density function (PDF) is given by

\[
f(x) = 2\rho \pi x e^{-\rho x^2}
\]

Therefore, the expected shortest distance between a requestor and nearest content copy is

\[
\bar{X} = \int_{0}^{\infty} x f(x) dx = \int_{0}^{\infty} 2\rho \pi x^2 e^{-\rho x^2} dx
\]

Using the fact \( \int_{0}^{\infty} x^2 e^{-\alpha x^2} dx = \frac{\sqrt{\pi}}{4\alpha} \) for \( \alpha > 0 \), and integral by part, we have

\[
\bar{X} = \frac{1}{2\sqrt{\rho}} = \frac{1}{2} \sqrt{\frac{B}{nC}}
\]

It has been shown in [49] that the transmission radius of a node, \( r(n) \), should be at least in the order of \( \sqrt{\log n/n} \) to guarantee that the network is connected as \( n \) goes to infinity. Thus, the average delay (number of transmission hops)
hops) between a content requestor and nearest node with the content in its cache is given by
\[
\bar{d}(n,B,C) = \frac{\bar{X}}{r(n)} = \Theta\left(\sqrt{\frac{B}{C \log n}}\right)
\]

In this paper we focus on the scenario where \( B = \omega(C \log n) \), i.e., the number of unique contents is at least \( \log n \) times the cache size of a node asymptotically. Otherwise if \( B = O(C \log n) \), the number of unique contents is within a logarithm factor of the cache size of a node, and in the steady state each node has a copy of all the contents in the cache of itself or one its one-hop neighbors almost surely due to the following reason. The transmission range of each hop to guarantee connectivity is \( r(n) = \sqrt{\log n/n} \). Thus the number of nodes with the transmission range of each hop is \( n\pi r^2(n) = \Theta(n \log n) \). The request can be satisfied if any one of the nodes within the transmission range has a copy of the content in its cache. The probability that any given content is in the cache of a node is \( C/B \), the probability that a request is satisfied within a transmission round is

\[
\text{Prob(a request is satisfied in a transmission round)} = 1 - (1 - C/B)^{\log n} = 1, \quad \text{if } B = O(C \log n), \ n \to \infty
\]

In the network we assume each node randomly chooses a content for downloading. Denote the per node capacity as \( \lambda \). Since each node receives the requested content data at a rate of \( \lambda \) bits per second, there are \( n \) nodes in the network, and each bit needs to be relayed on the average by at least \( \bar{d}(n,B,C) \) hops, the total amount of bandwidth requirement is at least \( n\lambda \bar{d}(n,B,C) \). The per-node capacity is upper-bounded by \( O(1/\sqrt{\log n}) \). Therefore, we have
\[
\begin{align*}
    n\lambda \bar{d}(n,B,C) & = O\left(\frac{1}{\sqrt{\log n}}\right) \\
    \lambda & = O\left(\sqrt{\frac{C}{B \log n}}\right)
\end{align*}
\]

\( \square \)

**B. Transparent Enroute Caching**

The main results on the capacity of the Transparent Enroute Caching scheme are presented as follows.

**Theorem 3:** Consider a wireless ad hoc network of \( n \) randomly located nodes and \( B \) unique contents of unit size, with each node having a cache size of \( C \) units. Assume each content is equally likely to be requested and cached by a node. Under the Transparent Enroute Caching scheme, the scaling laws of the capacity and packet delay (number of transmission hops) for content requests are characterized as follows:

- The per-node capacity is upper-bounded by
\[
\lambda(n,B,C) = O\left(\max\left(\frac{C}{B}, \frac{1}{\sqrt{n \log n}}\right)\right)
\]

- The average packet delay between a requestor and the first node on the path with the content in its cache is
\[
\bar{d}(n,B,C) = O\left(\min\left(\frac{B}{C \log n}, \sqrt{\frac{n}{\log n}}\right)\right)
\]

**Discussion of results.** The per-node capacity under the Transparent Enroute Caching scheme is limited by two factors, the cache-size to content ratio (\( C/B \)), and the inherent transport capacity constrained by the maximum number of simultaneous transmissions that can be supported by the network \( (1/\sqrt{n \log n}) \). When compared with the baseline scenario where there is no caching, the capacity gain due to Same content caching is given by the ratio of equations (6) and (1), i.e.,

\[
g = \max\left(\frac{C \sqrt{n \log n}}{B}, 1\right)
\]

Depending on the scaling behavior of the number of unique contents, the capacity and its gain over the no caching scenario falls into one of the two scaling regimes:

- \( B = O(C \sqrt{n \log n}) \): When the number of unique contents is asymptotically smaller than \( C \sqrt{n \log n} \), the distance from a requestor to a node with the content in its cache is smaller than that from the requestor to the content custodian, i.e., the request is satisfied enroute by an intermediate node before it reaches the content custodian with an average delay \( \bar{d}(n,B,C) = O\left(\frac{B}{C \log n}\right) \). In this case, the capacity is limited by and increases linearly with the cache-size to content ratio, i.e., \( \lambda(n,B,C) = O(C/B) \). For a fixed number of unique contents (\( B \)), an \( m \)-fold increase in cache size will result in an effective \( m \)-fold gain in capacity. It is beneficial to allocate more memory for content caching if available.

- \( B = \omega(C \sqrt{n \log n}) \): When the number of unique contents is asymptotically larger than \( C \sqrt{n \log n} \), a request is satisfied by the content custodian instead of an intermediate node almost surely due to the following reason. Since the minimum transmission range to ensure network connectivity is \( \sqrt{n \log n} \), the number of hops between a requestor and the content custodian is \( O(\sqrt{n \log n}) \). In the proof of Theorem 3, it is shown that the probability that a request is not satisfied in a transmission round is given by \( (1 - C/B)^{\log n} \). Therefore, the probability that a request is satisfied before reaching
the content custodian approaches zero as $n$ goes to infinity, i.e.,

$$\text{Prob(a request is satisfied before custodian)} = 1 - ((1 - \frac{C}{B})^{\log n}) \sqrt{n / \log n}$$

$$= 1 - (1 - \frac{C}{B}) \sqrt{n / \log n}$$

Since $B = \omega(C \sqrt{n \log n})$, we have $\frac{C \sqrt{n \log n}}{B} \to 0$, as $n \to \infty$. As a result, we have

$$(1 - \frac{C}{B}) \sqrt{n / \log n} = (1 - \frac{1}{B/C}) \frac{C \sqrt{n \log n}}{n}$$

$$= e^{-\frac{C \sqrt{n \log n}}{n}} \to 1$$, as $n \to \infty$.

Therefore, we have

$$\text{Prob(a request is satisfied before custodian)} = 1 - (1 - \frac{C}{B}) \sqrt{n / \log n} \to 0$$, as $n \to \infty$.

In this case the benefit of content caching is negligible and the capacity is inherently constrained by the number of simultaneous transmissions that can be supported by the network, i.e., $\lambda(n,B,C) = O\left(\frac{1}{\sqrt{n \log n}}\right)$.

We now compare the capacity and delay of the two content access schemes, namely, the Nearest Caching Node (NCN) scheme and the Transparent Enroute Caching (TEC) scheme. Note that in the proof of Theorem 2 it is mentioned that we will focus on the scenario where $B = \omega(C \sqrt{n \log n})$, otherwise the number of unique contents is almost on the same order as the cache size of a node and a request is almost surely satisfied within a transmission hop, which is of less interest. Also, for the Transparent Enroute Caching, we will consider the case where $B = O(C \sqrt{n \log n})$ and the capacity is constrained by the cache-size to content ratio ($C/B$), otherwise a request is almost surely satisfied by the content custodian and the capacity is same as the Gupta-Kumar result.

Since in the Nearest Caching Node scheme a requestor always locates and downloads a content from the nearest node with the content in its cache, the average packet delay (denoted as $\bar{d}_{\text{NCN}}(n,B,C)$) is necessarily smaller than that of the Transparent Enroute Caching scheme (denoted as $\bar{d}_{\text{TEC}}(n,B,C)$), i.e.,

$$\frac{\bar{d}_{\text{NCN}}(n,B,C)}{\bar{d}_{\text{TEC}}(n,B,C)} = \sqrt{\frac{C \log n}{B}} \to 0$$, as $n \to \infty$.

As a result, the per-node capacity under the Nearest Caching Node scheme (denoted as $\lambda_{\text{NCN}}(n,B,C)$) is asymptotically larger than that of the Transparent Enroute Caching scheme (denoted as $\lambda_{\text{TEC}}(n,B,C)$), i.e.,

$$\frac{\lambda_{\text{NCN}}(n,B,C)}{\lambda_{\text{TEC}}(n,B,C)} = \sqrt{\frac{B}{C \log n}} \to \infty$$, as $n \to \infty$.

It is worth noting that the Nearest Caching Node scheme incurs certain content search overhead as the requestor needs to first locate the nearest node with the content in its cache. In our analysis the content search overhead is ignored, since the size of a content request message is usually much smaller than that of the content itself, and the content caching information (which node has what contents) can be maintained by nodes and thus subsequent requests can be directly routed to the nearest caching node without flooding. In this paper we assume that nodes are stationary after deployment. However, in mobile ad hoc network scenarios, the maintenance costs for content caching information can be high and in that case the overhead cannot be ignored in the analysis.

**Proof of Theorem 3.** In the Transparent Enroute Caching scheme as shown in Figure 2, a request is routed towards the content custodian along the shortest path. The request is satisfied by the first intermediate node on the path that has the content in its cache. If no intermediate node has the content in its cache, the request will eventually reach be satisfied by the content custodian.

With the same argument as in the proof of Theorem 2, the probability that a request is satisfied in a transmission round is

$$\text{Prob(a request is satisfied in a transmission round)} = 1 - (1 - \frac{C}{B})^{\log n}$$

The expected number of hops a request traverses before it is satisfied by a node is thus

$$d(n,B,C) = \frac{1}{1 - (1 - \frac{C}{B})^{\log n}}$$ (8)

Since we have $B = \Omega(C \log n)$, we have

$$(1 - \frac{C}{B})^{\log n} = 1 - \frac{C \log n}{B}$$, as $n \to \infty$ (9)

Plugging equation (9) into (8), we obtain

$$d(n,B,C) = \frac{B}{C \log n}$$

Since each content has a custodian in the network, in the worst case, the request will be routed to and satisfied by the custodian. The delay is upper bounded by $1/r(n) = \sqrt{n / \log n}$. Therefore, we have

$$\bar{d}(n,B,C) = \min\left(\frac{B}{C \log n}, \sqrt{n / \log n}\right)$$

Denote the per node capacity as $\lambda$. Since each node receives $\lambda$ bits per second, there are $n$ nodes, and each bit needs to be relayed on average by $\bar{d}(n,B,C)$ hops, the total amount of bandwidth requirement is at least $n\lambda \bar{d}(n,B,C)$. The total amount of bandwidth that can be supported by
the entire network is proportional to $1/r^2(n) = n/ \log n$. Therefore, we have

\[ n \lambda \tilde{d}(n, B, C) = O \left( \frac{n}{\log n} \right) \]

\[ \lambda = O \left( \max \left( \frac{C}{B} \frac{1}{\sqrt{n \log n}} \right) \right) \]

VI. Conclusion

In this paper we study the capacity of cache-enabled content distribution wireless ad hoc networks. We first establish the capacity of content distribution wireless ad hoc networks without caching as a baseline for investigating the benefit of caching. We then obtain the scaling laws of the capacity for the two content access schemes. Based on the results we further explore their design and performance implications. Our results show that the capacity exhibits distinct scaling behaviors under different scaling regimes of the network parameters, namely, the number of unique contents and the cache-size of each node. Under certain conditions increasing the cache size of the nodes can effectively improve the capacity while under other conditions the improvement may be negligible. The characterizations of the capacity allows us to identify the bottleneck of the content access capacity for given network scenarios and choose effective approaches to improve the capacity. For future work we plan to study the capacity when there is infrastructure support (well connected base stations) in the network.

REFERENCES


