Title

Scaling Laws of the Throughput Capacity and Latency in Information-Centric Networks

Authors

Azimdoost, Bita
Westphal, Cedric
Sadjadpour, Hamid R

Publication Date

2012-10-03

Peer reviewed

Scaling Laws of the Throughput Capacity and Latency in Information-Centric Networks

Bita Azimdoost†, Cedric Westphal‡ *, and Hamid R. Sadjadpour†
†Department of Electrical Engineering and ‡Computer Engineering
University of California Santa Cruz, Santa Cruz, CA 95064, USA
{bazimdoost,cedric,hamid}@soe.ucsc.edu
* Huawei Innovation Center, Santa Clara, CA 95050, USA
cedric.westphal@huawei.com

***Abstract*—Wireless information-centric networks consider storage as one of the network primitives, and propose to cache data within the network in order to improve latency and reduce bandwidth consumption. We study the throughput capacity and delay in an information-centric network when the data cached in each node has a limited lifetime. The results show that with some fixed request and cache expiration rates, the order of the data access time does not change with network growth, and the maximum throughput order is inversely proportional to the square root and logarithm of the network size n in cases of grid and random networks, respectively. Comparing these values with the corresponding throughput and latency with no cache capability (throughput inversely proportional to the network size, and latency of order $\sqrt{n}$ and $\sqrt{\frac{n}{\log n}}$ in grid and random networks, respectively), we can actually quantify the asymptotic advantage of caching. Moreover, we compare these scaling laws for different content discovery mechanisms and illustrate that not much gain is lost when a simple path search is used.**

I. Introduction

In today's networking situations, users are mostly interested in accessing content regardless of which host is providing this content. They are looking for a fast and secure access to data in a whole range of situations: wired or wireless; heterogeneous technologies; in a fixed location or when moving. The dynamic characteristics of the network users makes the host-centric networking paradigm inefficient. Information-centric networking (ICN) is a new networking architecture where content is accessed based upon its name, and independently of the location of the hosts [1]–[4]. In most ICN architectures, data is allowed to be stored in the nodes and routers within the network in addition to the content publisher's servers. This reduces the burden on the servers and on the network operator, and shortens the access time to the desired content.

Combining content routing with in-network-storage for the information is intuitively attractive, but there has been few works considering the impact of such architecture on the capacity of the network in a formal or analytical manner. In this work we study a wireless information-centric network where nodes can both route and cache content[1]. We also assume that a node will keep a copy of the content only for a finite period of time, that is until it runs out of memory space in its cache and has to rotate content, or until it ceases to serve a specific content.

The nodes issue some queries for content that is not locally available. We suppose that there exists a server which permanently keeps all the contents. This means that the content is always provided at least by its publisher, in addition to the potential copies distributed throughout the network. Therefore, at least one replica of each content always exists in the network and if a node requests a piece of information, this data will be provided either by its original server or by a cache containing the desired data. When the customer receives the content, it will store the content and share it with the other nodes if needed.

The present paper thus investigates the access time and throughput capacity in such content-centric networks and addresses the following questions:

1) Looking at the throughput capacity and latency, can we quantify the performance improvement brought about by a content-centric network architecture over networks with no content sharing capability?
2) How does the content discovery mechanism affect the performance? More specifically, does selecting the nearest copy of the content improve the scaling of the capacity and access time compared to selecting the nearest copy in the direction of original server?
3) How does the caching policy, and in particular, the length of time each piece of content spends in the cache's memory, affect the performance?

We state our results in three Theorems; Theorem 1 formulates the throughput capacity in a grid network which uses the shortest path to the server content discovery mechanism considering different content availability in different caches, and Theorem 2 and 3 will answer the above questions studying two different network models (grid and random network) and two content discovery scenarios (shortest path to the server and shortest path to the closest copy of the content) when the information exists in all caches with the same probability. These Theorems demonstrate that adding the content sharing capability to the nodes can significantly increase the capacity.

The rest of the paper is organized as follows. After a brief

Bita Azimdoost was with Huawei Innovation Center, Santa Clara, CA 95050, USA, as an intern while working on this paper.

[1]A preliminary version of this paper has appeared at ITC25 [5]

review of the related work in Section II, the network models, the content discovery algorithms used in the current work, and the content distribution in steady-state are introduced in Section III. The main Theorems are stated and proved in Section IV. We will discuss the results and study some simple examples in Section V. Finally the paper is concluded and some possible directions for the future work will be introduced in section VI.

II. Related Work

Information Centric Networks have recently received considerable attention. While our work presents an analytical abstraction, it is based upon the principles described in some ICN architectures, such as CCN [4], NetInf [6], PURSUIT [2], or DONA [7], where nodes can cache content, and requests for content can be routed to the nearest copy. Papers surveying the landscape of ICN [3] [8] show the dearth of theoretical results underlying these architectures.

Caching, one of the main concepts in ICN networks, has been studied in prior works [3]. [9] computes the performance of a LRU cache taking into account the dynamical nature of the content catalog. Some performance metrics like miss ratio in the cache, or the average number of hops each request travels to locate the content have been studied in [10], [11], and the benefit of cooperative caching has been investigated in [12].

Optimal cache locations [13] and cache replacement techniques [14] are two other aspects most commonly investigated. And an analytical framework for investigating properties of these networks like fairness of cache usage is proposed in [15]. [16] considered information being cached for a limited amount of time at each node, as we do here, but focused on flooding mechanism to locate the content, not on the capacity of the network. [17] investigates the routing in such networks in order to minimize the average access delay.

However, to the best of our knowledge, there are just a few works focusing on the achievable data rates in such networks. Calculating the asymptotic throughput capacity of wireless networks with no cache has been solved in [18] and many subsequent works [19] [20]. Some work has studied the capacity of wireless networks with caching [21] [22] [23] . There, caching is used to buffer data at a relay node which will physically move to deliver the content to its destination, whereas we follow the ICN assumption that caching is triggered by the node requesting the content. [24] uses a network simulation model and evaluates the performance (file transfer delay) in a cache-and-forward system with no request for the data. [25] proposes an analytical model for single cache miss probability and stationary throughput in cascade and binary tree topologies. [26] considers a general problem of delivering content cached in a wireless network and provides some bounds on the caching capacity region from an information-theoretic point of view. Some scaling regimes for the required link capacity is computed in [27] for a static cache placement in a multihop wireless network.

III. Preliminaries

A. Network Model

Two network models are studied in this work.

1) Grid Network: Assume that the network consists of n nodes $V = \{v_1, v_2, ..., v_n\}$ each with a local cache of size L located on a grid (Figure 1). The distance between two adjacent nodes equals to the transmission range of each node, so the packets sent from a node are only received by four adjacent nodes. There are m different contents, $F = \{f_1, f_2, ..., f_m\}$ with sizes $B_i,\ i = 1, ..., m$, for which each node v_j may issue a query. Based on the content discovery algorithms which will be explained later in this section, the query will be transmitted in the network to discover a node containing the desired content locally. v_j then downloads b bits of data with rate γ in a hop-by-hop manner through the path P_{xj} from either a node $(v_i, x = i)$ containing it locally $(f \in v_i)$ or the server $(x = s)$. When the download is completed, the data is cached and shared with other nodes either by all the nodes on the delivery path, or only by the end node. In the paper we consider both options.

P_{js} denotes the nodes on the path from v_j to server. Without loss of generality, we assume that the server is attached to the node located at the middle of the network, as changing the location of the server does not affect the scaling laws. Using the protocol model and according to [28], the transport capacity in such network is upper bounded by $\Theta(W\sqrt{n})$. This is the model studied in 1 and the first two scenarios of Theorem 2.

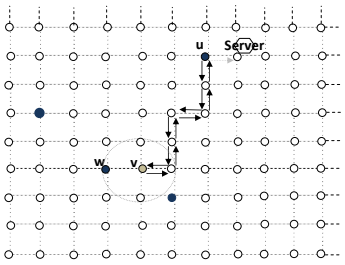

Fig. 1. The transmission range of node v contains four surrounding nodes. The black vertices contain the content in their local caches. The arrow lines demonstrate a possible discovery and receive path in scenario i, where node v downloads the required information from u. In scenario ii, v will download the data from w instead.

2) Random Network: The next network studied in Theorem 2 is a more general network model where the nodes are randomly distributed over a unit square area according to a uniform distribution. We use the same model used in [28] (section 5) and divide the network area into square cells each with side-length proportional to the transmission range $r(n)$, which is selected to be at least in the order of $\sqrt{\frac{\log n}{n}}$ to guarantee the connectivity of the network [29]. According to the protocol model [28], if the cells are far enough they can transmit data at the same time with no interference; we assume that there are M^2 non-interfering groups which take turn to transmit at the corresponding time-slot in a round robin fashion. Again, without loss of generality the server is assumed to be located at the middle of the network. In this model the maximum number of simultaneous feasible transmissions will be in the order of $\frac{1}{r^2(n)}$ as each transmission consumes an

area proportional to $r^2(n)$.

All other assumptions are similar to the grid network.

B. Content Discovery Algorithm

1) Path-wise Discovery: To discover the location of the desired content, the request is sent through the shortest path toward the server containing the requested content. If an intermediate node has the data in its local cache, it does not forward the request toward the server anymore and the requester will start downloading from the discovered cache. Otherwise, the request will go all the way toward the server and the content is obtained from the main source. In case of the random network when a node needs a piece of information, it will send a request to its neighbors toward the server, i.e. the nodes in the same cell and one adjacent cell in the path toward the server, if any copy of the data is found it will be downloaded. If not, just one node in the adjacent cell will forward the request to the next cell toward the server.

2) Expanding Ring Search: In this algorithm the request for the information is sent to all the nodes in the transmission range of the requester. If a node receiving the request contains the required data in its local cache, it notifies the requester and then downloading from the discovered cache is started. Otherwise, all the nodes that receive the request will broadcast the request to their own neighbors. This process continues until the content is discovered in a cache and the downloading follows after that. This will return the nearest copy from the requester.

C. Content Distribution in Steady-State

The time diagram of data access process in a cache is illustrated in Figure 2. When a query for content f_i is initiated, the content is available at the requester's cache after a wait time (T_3) which is a function of the distance between the user and the data source (server or an intermediate cache), the data size, and the download speed. An expiration timer will be set upon receiving the data, and this data will be finally dropped after a holding time (T_1) with distribution $\mathfrak{f}_{1_i}$ and mean $1/\mu_i$. During this time, the cached data can be shared with the other users if needed. The same user may re-issue a query for that data after some random time (T_2) with distribution $\mathfrak{f}_{2_i}$ and mean $1/\lambda_i$. Note that a node will send out a request for a content only if it does not have it in its local cache, otherwise, its request will be served locally and no request is sent to the other nodes. The solid lines in this diagram denote the portions of time that the data is available at local cache.

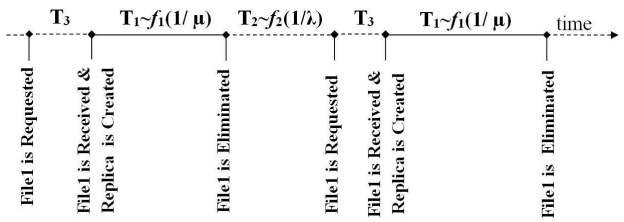

Fig. 2. Data access process time diagram in a cache network

In this work we assume identical content sizes $B_i = B$, and assume all the contents have the same popularity leading to similar request rates $\lambda_i = \lambda$, and the same holding times $\mu_i = \mu$. As the requests for different contents are supposed to be independent and holding times are set for each content independent of the others, we can do the calculations for one single content. If the total number of contents is not a function of the network size, this will not change the capacity order. Suppose that B is much larger than the request packet size, so we ignore the overhead of the discovery phase in our calculations. Furthermore, if the information sizes are the same and the download rates are also the same, the download time will be a function of the number of hops (h) between the source and the customer; $T_3 = Bh/\gamma$. In the steady-state analysis, we ignore this constant time.

The average portion of time that each node contains a content in its local cache is

$$\rho(n) = \frac{1/\mu}{1/\mu + 1/\lambda} = \frac{\lambda}{\lambda + \mu}, \tag{1}$$

which is the average probability that a node contains the data at steady-state. λ is the rate of requests for a data from each user in case of the data not being available, and μ is the rate of the data being expunged from the cache. Both these parameters are strongly dependent on the total number of users, or the topology and configuration of the network or the cache characteristics like size and replacement policy.

IV. Theorem Statements and Proofs

Theorem 1. *Consider a grid wireless network consisting of n nodes. Each node can transmit over a common wireless channel, with bandwidth W bits per second, shared by all nodes. Assume that there is a server which contains all the information. Without loss of generality we assume that this server is located in the middle of the network. Each node contains some information in its local cache. Assume that the probability of the information being in all the caches with the same distance (j hops) from the server is the same ($\rho_j(n)$). The maximum achievable throughput capacity order[2] (γ_{max}) in such network when the nodes use the nearest copy of the required content on the shortest path toward the server is given by*

$$\gamma_{max} \equiv \frac{W\sqrt{n}}{\sum_{i=1}^{\sqrt{n}} i \sum_{j=0}^{i-1}(i-j)\rho_j(n)\prod_{k=j+1}^{i}(1-\rho_k(n))},$$

where $\rho_0(n) = 1$, which means that the server always contains the information.

Proof: A request initiated by a user v_i in i-hop distance from the server (located in level $i = 1, .., \sqrt{n}$) is served by cache u_j located in level j, $1 \le j \le i$ on the shortest path from v_i to the server if no caches before u_j, including v_i, on

[2] $f(n) = O(g(n))$ or $f(n) \preceq g(n)$ if $sup_n(f(n)/g(n)) < \infty$. $f(n) = \Omega(g(n))$ or $f(n) \succeq g(n)$ if $g(n) = O(f(n))$. $f(n) = \Theta(g(n))$ or $f(n) \equiv g(n)$ if both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$. $f(n) = o(g(n))$ or $f(n) \prec g(n)$ if $f(n)/g(n) \to 0$. $f(n) = \omega(g(n))$ or $f(n) \succ g(n)$ if $g(n)/f(n) \to 0$.

this path contains the required information, and u_j contains it. This request is served by the server if no copy of it is available on the path. Assuming that the availability of the information in each cache is independent of the contents in the other caches, this probability denoted by $P_{i,j}$ is given by

$$P_{i,j} = (1-\rho_i(n))(1-\rho_{i-1}(n))...(1-\rho_{j+1}(n))\rho_j(n) \quad (2)$$

where $\rho_j(n)$ is the probability of the information being available in a cache in level $j, 1 \leq j \leq \sqrt{n}$, and $j = 0$ shows the server and $\rho_0(n) = 1$. Thus a content requested by v_i is traveling $i-j$ hops with probability $P_{i,j}$. There are $4i$ nodes in level i so the average number of hops ($E[h]$) traveled by each piece of data from the serving cache (or the original server) to the requester is

$$E[h] = \frac{1}{n}\sum_{i=1}^{\sqrt{n}} 4i \sum_{j=0}^{i-1}(i-j)P_{i,j},$$

$$= \frac{1}{n}\sum_{i=1}^{\sqrt{n}} 4i \sum_{j=0}^{i-1}(i-j)(1-\rho_i(n))...(1-\rho_{j+1}(n))\rho_j(n). \quad (3)$$

Assume that each user is receiving data with rate γ. The transport capacity in this network, which equals to $n\gamma E[h]$, is upper bounded by $\Theta(W\sqrt{n})$. So $\gamma_{max} = \Theta(\frac{W}{E[h]\sqrt{n}})$ and the Theorem is proved. ∎

Theorem 2. *Consider a wireless network consisting of n nodes, with each node containing the information in its local cache with common probability $\rho(n) \nrightarrow 1$.[3] Assume that the request process and cache look up time in each node is not a function of the number of nodes, then*

- *Scenario i- If the nodes are located on a grid and search for the contents just on the shortest path toward the server, the average delay order is*

$$\begin{cases} \Theta(\sqrt{n}) & ,if\ \rho(n) \preceq \frac{1}{\sqrt{n}} \\ \Theta(\frac{1}{\rho(n)}) & ,if\ \rho(n) \succeq \frac{1}{\sqrt{n}} \end{cases}$$

- *Scenario ii- If the nodes are located on a grid and use ring expansion as their content search algorithm, the average delay order is*

$$\begin{cases} \Theta(\sqrt{n}) & ,if\ \rho(n) \preceq \frac{1}{n} \\ \Theta(\frac{1}{\sqrt{\rho(n)}}) & ,if\ \rho(n) \succeq \frac{1}{n} \end{cases}$$

- *Scenario iii- If the nodes are randomly distributed over a unit square area and use path-wise content discovery algorithm, the average delay order is*

$$\begin{cases} \Theta(\sqrt{\frac{n}{\log n}}) & ,if\ \rho(n) \preceq \frac{1}{\sqrt{n\log n}} \\ \Theta(\frac{1}{\rho(n)\log n}) & ,if\ \frac{1}{\sqrt{n\log n}} \preceq \rho(n) \preceq \frac{1}{\log n} \\ \Theta(1) & ,if\ \rho(n) \succeq \frac{1}{\log n} \end{cases}$$

Here we prove Theorem 2 by utilizing some Lemmas.

[3]Note that for $\rho(n) \to 1$, the request is served locally and no data is transferred between the nodes.

Lemma 1. Consider the wireless networks described in Theorem 2. The average number of hops between the customer and the serving node (a cache or original server) is

- Scenario i

$$E[h] \equiv \tfrac{1}{n}\textstyle\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^i + \tfrac{\rho(n)}{n}\sum_{i=1}^{\sqrt{n}} i \sum_{k=1}^{i-1} k(1-\rho(n))^k \quad (4)$$

- Scenario ii

$$E[h] \equiv \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^{2i^2-2i+1} + \frac{1}{n}\sum_{i=2}^{\sqrt{n}} i \sum_{k=1}^{i-1} k(1-\rho(n))^{2k^2-2k+1}(1-(1-\rho(n))^{4k}) \quad (5)$$

- Scenario iii

$$E[h] \equiv \frac{\log n}{n}\sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i^2(1-\rho(n))^{i\log n} + \frac{\log n(1-(1-\rho(n))^{\log n})}{n}\sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i \sum_{k=1}^{i-1} k(1-\rho(n))^{k\log n} \quad (6)$$

Proof: Let h, d_{sr}, and d_{max} denote the number of hops between the customer and the serving node (cache or original server), the number of hops between the customer and the original server, and the maximum value of d_{sr}, respectively. The average number of hops between the customer and the serving node ($E[h]$) is given by

$$E[h] = \sum_{i=1}^{d_{max}} E[h|d_{sr}=i]Pr(d_{sr}=i) \quad (7)$$

Scenario i- This case can be considered as a special case of the network studied in theorem 1, where $\rho_i(n)$ is the same for all i. Thus we can drop the index i and let $\rho(n)$ denote the common value of this probability. Using equation 3 we will have

$$E[h] \equiv \frac{4}{n}\sum_{i=1}^{\sqrt{n}} i\{i(1-\rho(n))^i + \sum_{j=1}^{i-1}(i-j)(1-\rho(n))^{i-j}\rho(n)\} \quad (8)$$

The constant factor 4 does not have any affect on the scaling order, so it can be dropped. Using variable $k = i-j$ then proves the Lemma.

$$E[h] \equiv \frac{1}{n}[\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^i + \sum_{i=1}^{\sqrt{n}} i \sum_{k=1}^{i-1} k(1-\rho(n))^k\rho(n)] \quad (9)$$

Scenario ii - d_{max} in this network is $\Theta(\sqrt{n})$, and there are $4i$ nodes at distance of i hops from the original server.

$$Pr(d_{sr}=i) \equiv \frac{i}{n} \quad (10)$$

Each customer may have the required item in its local cache with probability $\rho(n)$. If the requester is one hop away from

the original server, it gets the required item from the server with probability $1-\rho(n)$. The customers at two hops distance from the server (8 such customers) download the required item from the original server (traveling $h=2$ hops) if no cache in a diamond of two hops diagonals contains it (probability $(1-\rho(n))^2$), and gets it from a cache at distance one hop if one of those caches has the item (probability $(1-\rho(n))(1-(1-\rho(n))^4)$). Using similar reasoning, the customers at distance i from the server get the item from the server (distance $h=i$ hops) with probability $(1-\rho(n))^{1+4(1+2+...+(i-1))}=(1-\rho(n))^{2i^2-2i+1}$, and from a cache at distance $h=k<i$ with probability $(1-\rho(n))^{2k^2-2k+1}(1-(1-\rho(n))^{4k})$ as there are $4k$ nodes at distance of k hops. Therefore, using equations (7) and (3)

$$E[h] \equiv \frac{1}{n}\sum_{i=2}^{\sqrt{n}} i \sum_{k=1}^{i-1} k(1-(1-\rho(n))^{4k})(1-\rho(n))^{2k^2-2k+1} + \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^{2i^2-2i+1} \quad (11)$$

Scenario iii - Each hop is one cell containing $\Theta(\log n)$ caches. d_{max} in this network is of the order of $\sqrt{\frac{n}{\log n}}$ and $\Pr(d_{sr}=i) \equiv \frac{i\log n}{n}$.

Each customer may have the required item in its local cache with probability $\rho(n)$. If the requester is one hop away from the original server ($4\Theta(\log n)$ nodes), it gets the required item from the server with probability $1-\rho(n)$. The customers at two hops distance from the server ($8\Theta(\log n)$ such customers) download the required item from the original server (traveling $h=2$ hops) if no cache in the cell at one hop distance contains it (probability $(1-\rho(n))^{2\log n}$), and gets it from a cache at distance one hop if one of those caches has the item (probability $(1-\rho(n))(1-(1-\rho(n))^{2\log n})$). Using similar reasoning the customers at distance i from the server get the item from the server (distance $h=i$ hops) with probability $(1-\rho(n))^{i\log n}$, and from a cache at distance $h=k<i$ with probability $(1-\rho(n))^{k\log n}(1-(1-\rho(n))^{\log n})$. Therefore, according to equation (7)

$$E[h] \equiv \frac{\log n}{n}(1-\rho(n)) + \frac{\log n}{n}\sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i^2(1-\rho(n))^{i\log n} + \frac{\log n(1-(1-\rho(n))^{\log n})}{n}\sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i\sum_{k=1}^{i-1} k(1-\rho(n))^{k\log n}. \quad (12)$$

Noting that $\frac{\log n}{n}(1-\rho(n))$ is always less than one, and tends to zero for sufficiently large n, the Lemma is proved. ■

Lemma 2. Consider the wireless networks described in Theorem 2. For sufficiently large networks, the average number of hops between the customer and the serving node (a cache or the original server) is

- Scenario i

$$E[h] \equiv \begin{cases} \sqrt{n} & \rho(n) \preceq \frac{1}{\sqrt{n}} \\ \frac{1}{\rho(n)} & \rho(n) \succeq \frac{1}{\sqrt{n}} \end{cases} \quad (13)$$

- Scenario ii

$$E[h] \equiv \begin{cases} \sqrt{n} & \rho(n) \preceq \frac{1}{n} \\ \frac{1}{\sqrt{\rho(n)}} & \rho(n) \succeq \frac{1}{n} \end{cases} \quad (14)$$

- Scenario iii

$$E[h] \equiv \begin{cases} \sqrt{\frac{n}{\log n}} & \rho(n) \preceq \frac{1}{\sqrt{n\log n}} \\ \frac{1}{\rho(n)\log n} & \frac{1}{\sqrt{n\log n}} \preceq \rho(n) \preceq \frac{1}{\log n} \\ 1 & \rho(n) \succeq \frac{1}{\log n} \end{cases} \quad (15)$$

Proof: To prove this Lemma we use the following equation which is true for every N and x.

$$\lim_{N\to\infty}(1-x)^N = \begin{cases} 1 & x=o(\frac{1}{N}) \\ e^{-xN} & x=\Theta(\frac{1}{N}) \\ 0 & x=\omega(\frac{1}{N}) \end{cases} \quad (16)$$

Scenario i - Let's define

$$E_s^i = \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^i, \quad (17)$$

$$E_c^i = \frac{\rho(n)}{n}\sum_{i=1}^{\sqrt{n}} i\sum_{k=1}^{i-1} k(1-\rho(n))^k. \quad (18)$$

Thus equation (4) is written as $E[h]=E_s^i+E_c^i$. First we investigate the value of E_s^i for different ranges of $\rho(n)$. The summation for E_s^i can be decomposed into two summations.

$$E_s^i \equiv \frac{1}{n}\Big(\sum_{i\prec\sqrt{n}} i^2(1-\rho(n))^i + \sum_{i\equiv\sqrt{n}} i^2(1-\rho(n))^i\Big) \quad (19)$$

Assume $\rho(n)\equiv\frac{1}{\sqrt{n}}$, then using first and second region of equation (16) we have

$$E_s^i \equiv \frac{1}{n}\Big(\sum_{i\prec\sqrt{n}} i^2 + \sum_{i\equiv\sqrt{n}} i^2\Big) \equiv \frac{n^{3/2}}{n} \equiv \sqrt{n}. \quad (20)$$

Moreover it can easily be seen that E_s^i is a decreasing function of $\rho(n)$, so for $\rho(n)$ with order less than $\frac{1}{\sqrt{n}}$ it is more than $\sqrt{n}$. Since $d_{max}=\sqrt{n}$, we can say $E_s^i\equiv\sqrt{n}$ for $\rho(n)\preceq\frac{1}{\sqrt{n}}$.

Now we expand the summation to obtain

$$E_s^i = \frac{(1-\rho(n))(2-\rho(n))}{n\rho^3(n)} - \frac{(1-\rho(n))^{\sqrt{n}+1}}{n\rho^3(n)}\times \big(n(1-\rho(n))^2-(1-\rho(n))(2n+2\sqrt{n}-1)+(\sqrt{n}+1)^2\big) \quad (21)$$

when $\rho(n)\succ\frac{1}{\sqrt{n}}$ then using third region in equation 16, $(1-\rho(n))^{\sqrt{n}+1}$ is going to zero exponentially, so $n(1-\rho(n))^{\sqrt{n}+1}\to 0$. Thus, $E_s^i\equiv\frac{1}{n\rho^3(n)}$.

$$E_s^i \equiv \begin{cases} \sqrt{n} & \rho(n)\preceq\frac{1}{\sqrt{n}} \\ \frac{1}{n\rho^3(n)} & \rho(n)\succ\frac{1}{\sqrt{n}} \end{cases} \quad (22)$$

According to equation (22) and since $E[h]=E_s^i+E_c^i$, when $E_s^i\equiv\sqrt{n}$ (for $\rho(n)\preceq\frac{1}{\sqrt{n}}$) which is the maximum possible

order for $E[h]$, then adding E_s^i to $E[h]$ cannot increase its order beyond the maximum possible value. Now to derive the order of $E[h]$ for other values of $\rho(n)$, we decompose the equation of E_c^i to the following summations and investigate their behaviors when $\rho(n) \succ \frac{1}{\sqrt{n}}$.

$$\begin{aligned} E_c^i &= E_c^{i1} + E_c^{i2} \\ E_c^{i1} &= \frac{1}{n}\sum_{i\equiv\sqrt{n}} i\sum_{k=1}^{i-1} k\rho(n)(1-\rho(n))^k \\ E_c^{i2} &= \frac{1}{n}\sum_{i\prec\sqrt{n}} i\sum_{k=1}^{i-1} k\rho(n)(1-\rho(n))^k \end{aligned} \quad (23)$$

The number of $i \equiv \sqrt{n}$ is in the order of $\Theta(1)$. Therefore using the following series $\sum_{x=1}^{n} xa^x = \frac{a^{n+1}(na-n-1)+a}{(a-1)^2}$, we have

$$\begin{aligned} E_c^{i1} &\equiv \frac{1}{\sqrt{n}}\sum_{k=1}^{\sqrt{n}} k\rho(n)(1-\rho(n))^k, \\ &\equiv \frac{1-\rho(n)}{\rho(n)\sqrt{n}}(1-(1-\rho(n))^{\sqrt{n}}(1+\rho(n)\sqrt{n})), \end{aligned}$$

which is equivalent to $\frac{1}{\rho(n)\sqrt{n}}$ when $\rho(n) \succ \frac{1}{\sqrt{n}}$.

Utilizing the same series, the first summation in E_c^{i2} is in the order of $\sqrt{n}$. Hence we arrive at

$$\begin{aligned} E_c^{i2} &\equiv \frac{1-\rho(n)}{\rho(n)n}\sum_{i\prec\sqrt{n}} i[1-\{1-\rho(n)+\rho(n)i\}(1-\rho(n))^{i-1}] \\ &\equiv \frac{1-\rho(n)}{\rho(n)}\times \\ &\quad -\frac{1}{n}\sum_{i\prec\sqrt{n}} i(1-\rho(n))^i \\ &\quad -\frac{1}{n}\sum_{i\prec\sqrt{n}} i^2\rho(n)(1-\rho(n))^{i-1} \\ &\equiv \frac{1-\rho(n)}{\rho(n)} - \frac{(1-\rho(n))^2}{\rho^3(n)n} - \frac{1}{\rho^3(n)n} \\ &\equiv \frac{1}{\rho(n)} \end{aligned} \quad (24)$$

Since $\rho(n) \succ \frac{1}{\sqrt{n}}$, E_c^{i2} is the dominant factor in E_c^i, and also it is dominant factor in $E[h]$. Thus,

$$E[h] \equiv \begin{cases} E_s^i \equiv \sqrt{n} & \rho(n) \preceq \frac{1}{\sqrt{n}} \\ E_c^{i2} \equiv \frac{1}{\sqrt{\rho(n)}} & \rho(n) \succ \frac{1}{\sqrt{n}} \end{cases} \quad (25)$$

Scenario ii - Let's define

$$E_s^{ii} = \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^{2i^2-2i+1},$$

$$E_c^{ii}\frac{1}{n}\sum_{i=2}^{\sqrt{n}} i\sum_{k=1}^{i-1} k(1-\rho(n))^{2k^2-2k+1}(1-(1-\rho(n))^{4k}),$$

$$E[h] = E_s^{ii} + E_c^{ii}. \quad (26)$$

Assume that $\rho(n) \equiv \frac{1}{n}$, then

$$\begin{aligned} E_s^{ii} &\equiv \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\frac{1}{n})^{2i^2-2i+1}, \\ &\equiv \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2 \equiv \sqrt{n}. \end{aligned} \quad (27)$$

Since E_s^{ii} is increasing when $\rho(n)$ is decreasing and its maximum possible order is $\sqrt{n}$, then $E_s^{ii} \equiv \sqrt{n}$ for all $\rho(n) \preceq \frac{1}{n}$.

For $\rho(n) \succ \frac{1}{n}$, we approximate the summation with the integral.

$$\begin{aligned} E_s^{ii} &\equiv \frac{1}{n}\int_{v=1}^{\sqrt{n}} v^2(1-\rho(n))^{2v^2-2v+1} \\ &\equiv \frac{(1-\log(1-\rho(n)))\sqrt{2\pi(1-\rho(n))}erf(\frac{(2v-1)\sqrt{-\log(1-\rho(n))}}{\sqrt{2}})}{n\log^{3/2}(1-\rho(n))} \\ &\quad + \frac{-2\sqrt{-\log(1-\rho(n))}(2v+1)(1-\rho(n))^{2v^2-2v+1}}{n\log^{3/2}(1-\rho(n))}|_{v=1}^{\sqrt{n}} \end{aligned} \quad (28)$$

where erf is the error function which is always limited by $[-1,1]$ and is zero at zero. If $\rho(n) \to 1$, then it is obvious that $E_s^{ii} \to 0$. For other values of $\rho(n) \succ \frac{1}{n}$ we use the third approximation in equation (16), and also[4] $-\log(1-\rho(n)) \equiv \rho(n)$ for $\rho(n) \to 0$ and $-\log(1-\rho(n)) \equiv 1$ for $\rho(n) \nrightarrow 0$ to obtain

$$E_s^{ii} \equiv \begin{cases} \sqrt{n} & \rho(n) \preceq \frac{1}{n} \\ \frac{1}{n\rho^{3/2}(n)} & \rho(n) \succ \frac{1}{n} \end{cases} \quad (29)$$

Since for $\rho(n) \preceq \frac{1}{n}$ the E_s^{ii} reaches the maximum $E[h]$, therefore E_c^{ii} cannot increase the scaling value of $E[h]$ anymore. For $\rho \succ \frac{1}{n}$ we have

$$E_c^{ii} \equiv \sqrt{\frac{1}{\rho(n)}}. \quad (30)$$

Thus it can easily be verified that

$$E[h] \equiv \begin{cases} E_s^{ii} \equiv \sqrt{n}, & \rho(n) \preceq \frac{1}{n} \\ E_c^{ii} \equiv \sqrt{\frac{1}{\rho(n)}}. & \rho(n) \succ \frac{1}{n} \end{cases} \quad (31)$$

Scenario iii - Let's define

$$E_s^{iii} = \frac{\log n}{n}\sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i^2(1-\rho(n))^{i\log n}$$

$$E_c^{iii} = \frac{\log(n)(1-(1-\rho(n))^{\log n})}{n}\sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i\sum_{k=1}^{i-1} k(1-\rho(n))^{k\log n}$$

$$E[h] \equiv E_s^{iii} + E_c^{iii} \quad (32)$$

[4]This is true when $\rho(n)$ tends to zero while n approaches infinity.

First we check the behavior of E_s^{iii} when $\rho(n) \equiv \frac{1}{\sqrt{n \log n}}$. Using second region in equation (16) we will have $E_s^{iii} \equiv \sqrt{\frac{n}{\log n}}$. E_s^{iii} is increasing when $\rho(n)$ is decreasing and the maximum possible value for the number of hops is $\sqrt{\frac{n}{\log n}}$, then $E_s^{iii} \equiv \sqrt{\frac{n}{\log n}}$ for all $\rho(n) \preceq \frac{1}{\sqrt{n \log n}}$.

By approximating the summation with integral, we arrive at

$$E_s^{iii} \equiv \frac{\log n}{n} \int_2^{\sqrt{\frac{n}{\log n}}} v^2 (1-\rho(n))^{v \log n},$$
$$\equiv \{ \frac{\log(n)(1-\rho(n))^{v \log n}}{n \log^3 (1-\rho(n))^{\log n}} \times$$
$$(v^2 \log^2 (1-\rho(n))^{\log n} - 2v \log (1-\rho(n))^{\log n} + 2)\}|_{v=2}^{\sqrt{\frac{n}{\log n}}}. \quad (33)$$

If $\frac{1}{\sqrt{n \log n}} \preceq \rho(n) \preceq \frac{1}{\log n}$, using equation (16) and the fact $\log (1-\rho(n))^{\log n} \equiv -\rho(n) \log n$, we will have

$$E_s^{iii} \equiv \frac{1}{n \rho^3(n) \log^2 n}. \quad (34)$$

When $\rho(n) \succeq \frac{1}{\log n}$, equation (32) tends to zero.

$$E_s^{iii} \equiv \begin{cases} \sqrt{\frac{n}{\log n}} & \rho(n) \preceq \frac{1}{\sqrt{n \log n}} \\ \frac{1}{n \rho^3(n) \log^2 n} & \frac{1}{\sqrt{n \log n}} \preceq \rho(n) \preceq \frac{1}{\log n} \\ 0 & \rho(n) \succeq \frac{1}{\log n} \end{cases} \quad (35)$$

Using the previous approximations along with $1-(1-\rho(n))^{\log n} \equiv 1$ for $\rho(n) \succeq \frac{1}{\log n}$ and $\rho(n) \log n$ for $\rho(n) \preceq \frac{1}{\log n}$, we can approximate E_c^{iii} as its dominant terms.

$$E_c^{iii} \equiv \frac{1}{n \rho(n)} \sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i \equiv \frac{1}{\rho(n) \log n} \quad (36)$$

When $\rho(n) \succeq \frac{1}{\log n}$, the dominant term is $\Theta(1)$. Thus,

$$E[h] \equiv \begin{cases} E_s^{iii} \equiv \sqrt{\frac{n}{\log n}} & \rho(n) \preceq \frac{1}{\sqrt{n \log n}} \\ E_c^{iii} \equiv \frac{1}{\rho(n) \log n} & \frac{1}{\sqrt{n \log n}} \preceq \rho(n) \preceq \frac{1}{\log n} \\ E_c^{iii} \equiv 1 & \frac{1}{\log n} \preceq \rho(n) \end{cases} \quad (37)$$

It can be seen that for large enough $\rho(n)$ the average number of hops between the nearest content location and the customer is just $\Theta(1)$ hops. This is the result of having $log(n)$ caches in one hop distance for every requester. Each one of these caches can be a potential source for the content. When the network grows, this number will increase and if $\rho(n)$ is large enough ($\frac{1}{\log n} \preceq \rho(n)$) the probability that at least one of these nodes contain the required data will approach 1, i.e., $\lim_{n \to \infty} (1-(1-\rho(n))^{\log n}) = 1$.

■

Theorem 2 is now simply proved using the above Lemmas.

Proof: Assuming that the delay of the request process and cache look up in each node is not increasing when the network size (the number of nodes) increases, and the there is enough bandwidth to avoid congestion, then the delay of getting the data is directly proportional to the average number of hops between the serving node and the customer. Thus, the delay and the average number of hops the data is traveling to reach the customer are of the same order and *Theorem 2* is proved. ■

Theorem 3. *Consider the networks of Theorem 2, and assume each node can transmit over a common wireless channel, with W bits per second bandwidth, shared by all nodes. The maximum achievable throughput capacity order γ_{max} in the three discussed scenarios are*

- *Scenario i-*

$$\begin{cases} \Theta(\frac{W \rho(n)}{\sqrt{n}}) & ,if\ \rho(n) \succeq \frac{1}{\sqrt{n}} \\ \Theta(\frac{W}{n}) & ,if\ \rho(n) \preceq \frac{1}{\sqrt{n}} \end{cases}$$

- *Scenario ii-*

$$\begin{cases} \Theta(W \sqrt{\frac{\rho(n)}{n}}) & ,if\ \rho(n) \succeq \frac{1}{n} \\ \Theta(\frac{W}{n}) & ,if\ \rho(n) \preceq \frac{1}{n} \end{cases}$$

- *Scenario iii-*

$$\begin{cases} \Theta(\frac{W}{\log n}) & ,if\ \rho(n) \succeq \frac{1}{\log n} \\ \Theta(\rho^2(n) \log n W) & ,if\ \frac{1}{\sqrt{n \log n}} \preceq \rho(n) \preceq \frac{1}{\log n} \\ \Theta(\frac{W}{n}) & ,if\ \rho(n) \preceq \frac{1}{\sqrt{n \log n}} \end{cases}$$

To prove Theorem 3 we use Lemma 2, and the following two Lemmas.

Lemma 3. Consider the wireless networks described in Theorem 2. In order not to have interference, the maximum throughput capacity is upper limited by

- Scenario i-

$$\begin{cases} \Theta(\frac{W \rho(n)}{\sqrt{n}}) & ,if\ \rho(n) \succeq \frac{1}{\sqrt{n}} \\ \Theta(\frac{W}{n}) & ,if\ \rho(n) \preceq \frac{1}{\sqrt{n}} \end{cases}$$

- Scenario ii-

$$\begin{cases} \Theta(W \sqrt{\frac{\rho(n)}{n}}) & ,if\ \rho(n) \succeq \frac{1}{n} \\ \Theta(\frac{W}{n}) & ,if\ \rho(n) \preceq \frac{1}{n} \end{cases}$$

- Scenario iii-

$$\begin{cases} \Theta(\frac{W}{\log n}) & ,if\ \rho(n) \succeq \frac{1}{\log n} \\ \Theta(\rho W) & ,if\ \frac{1}{\sqrt{n \log n}} \preceq \rho(n) \preceq \frac{1}{\log n} \\ \Theta(\frac{W}{\sqrt{n \log n}}) & ,if\ \rho(n) \preceq \frac{1}{\sqrt{n \log n}} \end{cases}$$

Proof: Assume that each content is retrieved with rate γ bits/sec. The traffic generated because of one download from a cache (or server) at average distance of $E[h]$ hops from the requester node is $\gamma E[h]$. The total number of requests for a content in the network at any given time is limited by the number of nodes n. Thus the maximum total bandwidth needed to accomplish these downloads will be $nE[h]\gamma$, which is upper limited by $(\Theta(W\sqrt{n}))$ in scenarios i, ii, and $(\Theta(\frac{W}{r^2(n)}) = \Theta(\frac{Wn}{\log n}))$ in scenario iii. Thus,

$$\begin{aligned} nE[h]\gamma &\preceq W\sqrt{n} \\ \gamma_{max} &\equiv \frac{W}{\sqrt{n}E[h]} \end{aligned} \tag{38}$$

in scenarios i, ii, and

$$\begin{aligned} nE[h]\gamma &\preceq \frac{Wn}{\log n} \\ \gamma_{max} &\equiv \frac{W}{\log nE[h]} \end{aligned} \tag{39}$$

in scenarios iii. Therefore the maximum download rate is easily derived using the results of Lemma 2. ∎

In the previous Lemma, the maximum throughput capacity in a wireless network utilizing caches has been calculated such that no interference occurs. Now it is important to verify if this throughput can be supported by each node (cell), i.e. the traffic carried by each node (cell) is not more than what it can support ($\Theta(1)$).

Lemma 4. The throughput capacities of Lemma 3 are supported for all values of $\rho(n)$ in grid topology. The random network can support the obtained throughput capacities just when $\rho(n) \succeq \frac{1}{\log n}$. For smaller values of $\rho(n)$ the maximum supportable throughput capacities are as follows.

$$\gamma_{max} \equiv \begin{cases} \frac{1}{n} & \rho(n) \preceq \frac{1}{\sqrt{n\log n}} \\ \rho^2(n)\log n & \frac{1}{\sqrt{n\log n}} \preceq \rho(n) \prec \frac{1}{\log n} \end{cases} \tag{40}$$

Proof: Each link between two nodes in scenarios i and ii, or two cells in scenario iii can carry at most $\Theta(1)$ bits per second. Here we calculate the maximum traffic passing through a link considering the throughput capacities derived in previous Theorems, and check if any link can be a bottleneck.

Scenario i- Each one of the four links connected to the server will carry all the traffic related to the items not found in the on-path caches. Thus, the total traffic carried by each of those links is $\sum_{i=1}^{\sqrt{n}} \gamma i(1-\rho(n))^i$.

When $\rho(n) \preceq \frac{1}{\sqrt{n}}$, we have $(1-\rho(n))^i \equiv 1$ for all $i \leq \sqrt{n}$. So this traffic is equal to

$$\sum_{i=1}^{\sqrt{n}} \gamma i \equiv n\gamma \preceq n\gamma_{max} \equiv 1. \tag{41}$$

When $\rho(n) \succeq \frac{1}{\sqrt{n}}$, using equation 16 the above summation can be written as

$$\begin{aligned} &\gamma\{\frac{(1-\rho(n))^{\sqrt{n}}(\sqrt{n}\log(1-\rho(n))-1)}{\log^2(1-\rho(n))} \\ &\quad -\frac{(1-\rho(n))(\log(1-\rho(n))-1)}{\log^2(1-\rho(n))}\} \\ &\equiv \gamma\frac{(1-\rho(n))(-\log(1-\rho(n))+1)}{\log^2(1-\rho(n))} \\ &\preceq \gamma_{max}\frac{(1-\rho(n))(-\log(1-\rho(n))+1)}{\log^2(1-\rho(n))} \\ &\equiv \frac{\rho(n)(-\log(1-\rho(n))+1)}{\sqrt{n}\log^2(1-\rho(n))} \preceq 1 \end{aligned} \tag{42}$$

Therefore, the links directly connected to the server will never be a bottleneck. On the other hand, the traffic carried by a node to cache content in level j is $\sum_{i=1}^{\sqrt{n}-j} \gamma i(1-\rho(n))^i \preceq \sum_{i=1}^{\sqrt{n}} \gamma i(1-\rho(n))^i$, so the server links carry the maximum load, and thus the derived capacity is supportable in every link.

Scenario ii- Each one of the four links connected to the server will carry all the traffic related to the items not found in any caches closer to the requester. Thus, the total traffic carried by each of those links is

$$\begin{aligned} &\gamma(1-\rho(n)) + \textstyle\sum_{i=1}^{\sqrt{n}} 4\gamma i(1-\rho(n))^{(1+4\sum_{j=1}^{i} j)} \\ &\equiv \gamma(1-\rho(n)) + \textstyle\sum_{i=1}^{\sqrt{n}} \gamma i(1-\rho(n))^{2i^2+2i+1}, \\ &\equiv \gamma\{(1-\rho(n)) + \\ &\frac{(1-\rho(n))^n-(1-\rho(n))^4}{\log(1-\rho(n))/(1-\rho(n))} + \\ &\frac{\sqrt{-\frac{\log(1-\rho(n))}{1-\rho(n)}}(erf(\sqrt{-n\log(1-\rho(n))})-erf(\sqrt{-\log(1-\rho(n))}))}{\log(1-\rho(n))/(1-\rho(n))}\}. \end{aligned} \tag{43}$$

If $\rho(n) \preceq \frac{1}{n}$, then $(1-\rho(n))^{2i^2+2i+1} \equiv 1$ for all $1 \leq i \leq \sqrt{n}$. Thus the above traffic will be $n\gamma \preceq n\gamma_{max} \equiv 1$.

If $\rho(n) \succeq \frac{1}{n}$ and $\rho(n) \to 0$, then using $\log(1-\rho(n)) \equiv -\rho(n)$ the above equation is equivalent to $\frac{\gamma}{\rho(n)} \preceq \frac{1}{\sqrt{n\rho(n)}}$, which is less than 1 in order.

Finally, if $\rho(n) = \Theta(1)$, then the traffic is equivalent to γ, which is less than $\Theta(1)$. So server links will not be a bottleneck. Using similar reasoning as in scenario ii other links carry less traffic, so the derived capacities are supportable.

Scenario iii- The traffic load between the server cell and each of the four neighbor cells is given by

$$\begin{aligned} &\gamma\log(n)\{(1-\rho(n)) + \sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i(1-\rho(n))^{i\log n}\} \\ \equiv\ &\gamma\log(n)\{(1-\rho(n)) \\ &+\frac{(1-\rho(n))^{\sqrt{n\log n}}(\sqrt{n\log n}\log(1-\rho(n))-1)}{\log^2(1-\rho(n))^{\log n}} \\ &-\frac{(1-\rho(n))^{\log n}(\log(1-\rho(n))^{\log n}-1)}{\log^2(1-\rho(n))^{\log n}}\} \end{aligned} \tag{44}$$

If $\rho(n) \preceq \frac{1}{\sqrt{n\log n}}$, then $(1-\rho(n))^{i\log n} \to 1$ for $2 \leq i \leq \sqrt{\frac{n}{\log n}}$, thus the traffic load equals to $\gamma\log n\sum_{i=2}^{\sqrt{\frac{n}{\log n}}} i \equiv n\gamma \equiv \sqrt{\frac{n}{\log n}} \succ 1$. Therefore, the obtained capacity is not supported for very small $\rho(n)$ ($\preceq \frac{1}{\sqrt{n\log n}}$). The maximum supportable throughput capacity in this case is $\gamma \preceq \frac{1}{n}$.

If $\frac{1}{\sqrt{n\log n}} \preceq \rho(n) \preceq \frac{1}{\log n}$, then the maximum traffic load on a link is

$$\gamma\log n + \gamma\log n\frac{1+2\rho(n)\log n}{\rho^2(n)\log^2 n} \equiv \frac{\gamma}{\rho^2(n)\log n} \tag{45}$$

The maximum throughput capacity obtained for this region is $\Theta(\rho(n))$, which will lead to a traffic load of $\frac{1}{\rho(n)\log n} \succeq 1$,

which means that this bit rate is not supportable. The maximum supportable rate in this region is then $\rho^2(n)\log n$, which is much less than $\rho(n)$.

If $\rho(n) \succeq \frac{1}{\log n}$ and $\rho(n) \to 0$, then equation (44) is equivalent to $\gamma \log n \equiv 1$, which is supportable. If $\rho(n) \succeq \frac{1}{\log n}$ and $\rho(n) \nrightarrow 0$, then the maximum traffic is γ, which is less than 1, and supportable.

Note that if there were no cache in the system, or $\rho(n)$ is very low, less than the stated threshold values, almost all the requests would be served by the server, and the maximum download rate would be $\Theta(\frac{W}{n})$ in case i, ii and $\Theta(\frac{W}{\sqrt{n}\log n})$ in case iii.

■

The maximum throughput capacity is the value which can be supported by all the nodes while no interference is occurred. Thus combining Lemmas 3 and 4, Theorem 3 is proved.

V. Discussion

In this section, we discuss our results based on two examples. The first example is that of a grid wireless network with n caches, and one server, which contains all the items located in the middle of the network. The requesters use the path search to locate the contents. In the second example we study the impact of caching on the maximum capacity order in the grid and random networks where all the caches have the same probability of having each item at any given time. The networks where the received data is stored only at the receivers and then shared with the other nodes as long as the node keeps the content can be considered as an example of such networks.

1) Example 1: Assume that each cache in level i (nodes at i hops away from the server) in a grid network receives requests for a specific document according to a Poisson distribution with rate β from the local user, and with rate $\beta_i'(n)$ from all the other nodes. Note that rate $\beta_i'(n)$ is a function of the individual request rate of users (β) and also the location of the cache in the network. The content discovery mechanism is path-wise discovery, and whenever a copy of the required data is found (in a cache or server), it will be downloaded through the reverse path, and all the nodes on the download path store it in their local caches. Moreover, we assume that receiving the data and also any request for the available cached data by a node in level i refreshes a time-out timer with fixed duration D_i. According to [30], this is a good approximation for caches with Least Recently Used (LRU) replacement policy when the cache size and the total number of documents are reasonably large. We will calculate the average probability of the data being in a cache in level i ($\rho_i(n)$) based on these assumptions and then use Theorem 1 to obtain the throughput capacity.

Let random variable $t_{on}(T)$ denote the total time of the data being available in a cache during constant time T. Assume that $N(T)$ requests are received by each node v_i in level i (i hop distance from the server). The data available time between any two successive requests (internal and external) is D_i if the timer set by the first request is expired before the second one comes, or is equal to the time between these two requests. Let τ_i^{req} denote the time between receiving two successive requests. This process has an exponential distribution with parameter $\beta_i = \beta + \beta_i'$. So the total time of data availability in a level i cache is

$$t_{on}(T) = \sum_{k=0}^{N(T)} \min(\tau_i^{req}, D_i), \tag{46}$$

and the average value of this time is

$$\begin{aligned} E[t_{on}(T)] &= \sum_{m=0}^{\infty} E[\sum_{k=0}^{m} \min(\tau_i^{req}, D_i)] Pr(N(T)=m), \\ &= \sum_{m=0}^{\infty} mE[\min(\tau_i^{req}, D_i)] Pr(N(T)=m), \\ &= E[\min(\tau_i^{req}, D_i)] E[N(T)]. \end{aligned} \tag{47}$$

According to the Poisson arrivals of requests with parameter $\beta + \beta_i'$, $E[N(T)] = (\beta + \beta_i')T$.

$E[\min(\tau_i^{req}, D_i)]$ can be easily calculated and equals to $\frac{1-e^{-D_i(\beta+\beta_i')}}{\beta+\beta_i'}$. Therefore,

$$E[t_{on}(T)] = (1 - e^{-D_i(\beta+\beta_i')})T \tag{48}$$

And finally the probability of an item being available in a level i cache is $\rho_i = \frac{E[t_{on}(T)]}{T} = 1 - e^{-D_i(\beta+\beta_i'(n))}$. Note that $D_0 = \infty$ so that $\rho_0 = 1$.

Now we need to calculate the rate of requests received by each node in level i. We assume that the shortest path from the requester to the server is selected such that all the nodes in level i receive the requests with the same rate. There are $4i$ nodes in level i and $4(i+1)$ nodes in level $i+1$. So the request initiated or forwarded from a node in level $i+1$ will be received by a specific node in level i with probability $\frac{i}{i+1}$ if it is not locally available in that node, so $\beta_i'(n)$ can be expressed as

$$\beta_i' = \frac{(1-\rho_{i+1})(\beta + \beta_{i+1}')(i+1)}{i} \tag{49}$$

Combining equation 49, the relationship between ρ_i and β_i', and the fact that there is no external request coming to the nodes at the edge boundary of the network ($\beta'_{\sqrt{n}} = \beta$), together with the result of Theorem 1 we can obtain the capacity (γ_{max}) in the grid network with path-wise content discovery and on-path storing scheme which is given by

$$\frac{W\sqrt{n}/4}{\sum_{i=1}^{\sqrt{n}} i \sum_{j=0}^{i} e^{-\sum_{k=j+1}^{i} D_k(\beta+\beta_k')}(1 - e^{-D_j(\beta+\beta_j')})} \tag{50}$$

Figure 3 (a) illustrates that the maximum throughput capacity changes with the network size (n) when $D_i\beta$ is the same for all nodes. It can be seen that the this capacity is inversely proportional to $\sqrt{n}$, just like the throughput capacity when no timer refreshing is available and the downloaded data is stored just in the end user's cache.

Figure 3 (b) shows the capacity versus different values for $D_i\beta$ assuming $n = 10^4$ and same timer expiration time for

all nodes. It can be seen that the maximum capacity is very close to $e^{D\beta-1}/\sqrt{n}$. For large $D\beta$ products the probability of the content being available in each and every cache will tend to be one, so all the contents are downloaded from the local cache and no data transfer is needed to be done, therefore the calculated throughput capacity will be very large which means that the all the links are available with their maximum bandwidth.

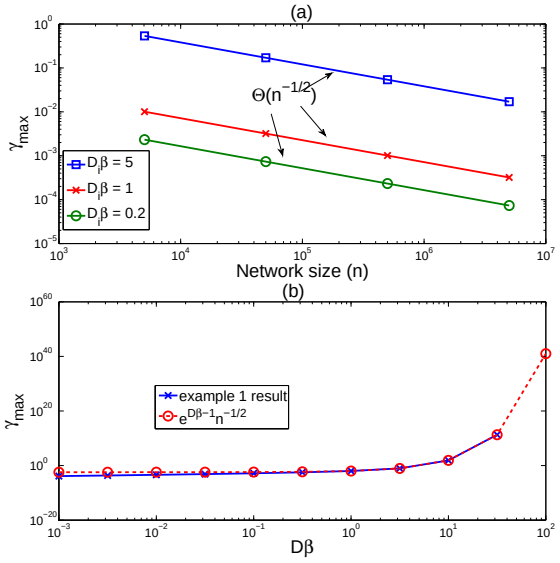

Fig. 3. Maximum throughput capacity (γ_{max}) versus (a) network size (n), (b) Timeout-request rate product (βD).

2) Example 2: As a possible example leading to equal probability of all the caches containing a piece of data, which is the basic assumption of Theorems 2 and 3, assume that receiving a data in the local cache of the requesting user sets a time-out timer with exponentially distributed duration with parameter η and no other event will change the timer until it times-out, meaning that $\mu = \eta$. Considering the request process for each content from each user being a Poisson process with rate β, and using the memoryless property of exponential distribution (internal request inter-arrival times), and assuming that the received data is stored only in the end user's cache (the caches on the download path don't store the downloading data), it can be proved that $\lambda = \beta$. Thus we can write the presence probability of each content in each cache as $\rho(n) = \frac{\beta}{\beta+\eta}$.

Figures 4 (a),(b) respectively illustrate the total request rate and the total traffic generated in a fixed size network in scenario i for different request rates when the time-out rate is fixed. The total request rate in the network is the product of the number of requesting nodes and the rate at which each node is sending the request ($n(1-\rho)\lambda$). The total traffic is the product of the total request rate and the number of hops between source and destination and the content size ($n(1-\rho)\lambda BE[h]$). Small λ means that each node is sending requests with low rate, so fewer caches have the content, and consequently more nodes are sending requests with this low rate. In this case most of the requests are served by the server. The total request rate will increase by increasing the per node request rate. High λ shows that each node is requesting the content with higher rate, so the number of cached content in the network is high, thus fewer nodes are requesting the content with this high rate externally. Here most of the requests are served by the caches. The total request rate then is determined by the content drop rate. So for very large λ, the total request rate is the total number of nodes in the network times the drop rate ($n\mu$) and the total traffic is $n\mu B$. As can be seen there is some request rate at which the traffic reaches its maximum; this happens when there is a balance between the requests served by the server and by the caches, for smaller request rates, most of the requests are served by the server and increasing λ increases the total traffic; for larger λ, on the other hand, most of the requests are served by the caches and increasing the request rate will decrease the distance to the nearest content and decrease total traffic.

Figures 5 (a),(b) respectively illustrate the total request rate and the total traffic generated in a fixed size network in scenario i for different time-out rates when the request rate is fixed. Low $1/\mu$ means high time-out rates or small lifetimes, which means most of the requests are served by the server and caching is not used at all. For large time-out times, all the requests are served by the caches, and the only parameter in determining the total request rate is the time-out rate.

However, when the network grows the traffic in the network will increase and the download rate will decrease. If we assume that the new requests are not issued in the middle of the previous download, the request rate will decrease with network growth. If the holding time of the contents in a cache increases accordingly the total traffic will not change, i.e. if by increasing the network size the requests are issued not as fast as before, and the contents are kept in the caches for longer times, the network will perform similarly.

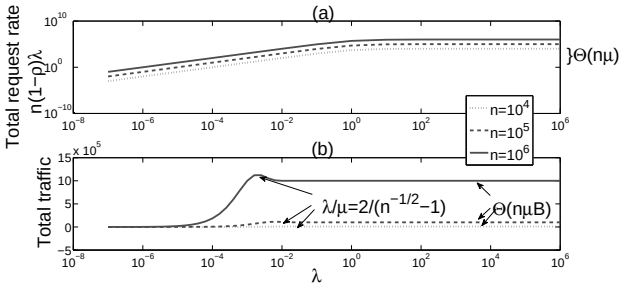

Fig. 4. (a) Total request rate in the network ($\lambda n(1-\rho(n))$), (b) Total traffic in the network ($B\lambda n(1-\rho(n))E[h]$) vs. the request rate ($\lambda$) with fixed time-out rate ($\mu = 1$).

In Figure 6 (a) we assume that the request rate is roughly 7 times the drop rate, so $\rho(n) = 7/8$, and show the maximum throughput order as a function of the network size. According to Theorem 2 and as can be observed from this figure, the maximum throughput capacity of the network in a grid network with the described characteristics is inversely proportional to the square root of the network size if the probability of each item being in each cache is fixed, while in a network with no cache this capacity will be inversely proportional to the network size. Similarly in the random network the maximum throughput is inversely proportional to the logarithm of the network size.

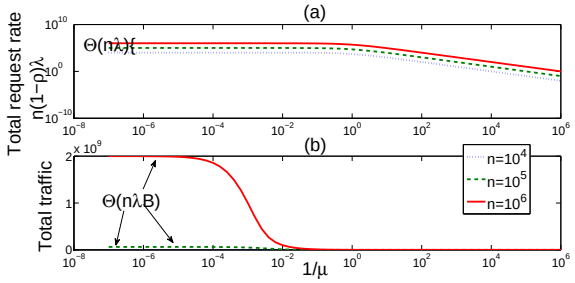

Fig. 5. (a) Total request rate in the network ($\lambda n(1-\rho(n))$), (b) Total traffic in the network ($B\lambda n(1-\rho(n))E[h]$) vs. the inverse of the time-out rate ($1/\mu$) with fixed request ratio ($\lambda = 1$).

Moreover, comparing scenario i with ii, we observe that the throughput capacity in both cases are almost the same; meaning that using the path discovery scheme will lead to almost the same throughput capacity as the expanding ring discovery. Thus, we can conclude that just by knowing the address of a server containing the required data and forwarding the requests through the shortest path toward that server we can achieve the best performance, and increasing the complexity and control traffic to discover the closest copy of the required content does not add much to the capacity.

On the other hand with a fixed network size, if the probability of an item being in each cache is greater than a threshold ($\Theta(\frac{1}{\sqrt{n}})$, $\Theta(\frac{1}{n})$, and $\Theta(\frac{1}{\log n})$ in cases i, ii and iii, respectively), most of the requests will be served by the caches and not the server, so increasing the probability of an intermediate cache having the content reduces the number of hops needed to forward the content to the customer, and consequently increases the throughput (Figure 6 (b), $n = 10^4$). For content presence probability orders less than these thresholds ($\Theta(\frac{1}{\sqrt{n}\log n}$) in cases iii) most of the requests are served by the main server, so the maximum possible number of hops will be traveled by each content to reach the requester and the minimum throughput capacity ($\Theta(\frac{W}{n})$) will be achieved. Note that in random network, the maximum throughput is limited by the maximum supportable load on each link.

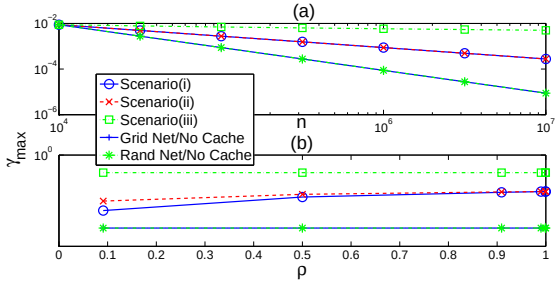

Fig. 6. Maximum download rate (γ_{max}) vs. (a) the number of nodes (n), (b) the content presence probability($\rho(n)$).

As may have been expected and according to our results, the obtained throughput is a function of the probability of each content being available in each cache, which in turn is strongly dependent on the network configuration and cache management policy.

VI. Conclusion And Future Work

We studied the asymptotic throughput capacity and latency of ICNs with limited lifetime cached data at each node. The grid and random networks are two network models we investigated in this work. The results show that with fixed content presence probability in each cache, the network can have the maximum throughput order of $1/\sqrt{n}$ and $1/\log n$ in cases of grid and random networks, respectively, and the number of hops travelled by each data to reach the customer (or latency of obtaining data), can be as small as one hop.

Moreover, we studied the impact of the content discovery mechanism on the performance. It can be observed that looking for the closest cache containing the content will not have much asymptotic advantage over the simple path-wise discovery. Consequently, downloading the nearest available copy on the path toward the server will have the same performance as downloading from the nearest copy. A practical consequence of this result is that routing may not need to be updated with knowledge of local copies, just getting to the source and finding the content opportunistically will yield the same benefit.

Another interesting finding is that whether all the caches on the download path keep the data or just the end user does it, the maximum throughput capacity scale does not change.

In this work, we have made several assumptions to simplify the analysis. For example, we assumed all the contents have the same characteristics (size, popularity). This assumption should be relaxed in future work. We also assumed that the requester downloads the data completely from one content location. However, if the node that needs the data can download each part of it from different nodes and makes a complete content out of the collected parts, achievable capacities may be different. Proposing a caching and downloading scheme that can improve the capacity order is part of our future work.

References

[1] L. Zhang, D. Estrin, J. Bruke, V. Jacobson, J. Thornton, D. Smetters, B. Zhang, G. Tsudik, K. Claffy, D. Krioukov, D. Massey, C. Papadopoulos, T. Abdelzaher, L. Wang, P. Crowley, and E. Yeh, "Named data networking (NDN) project," Oct. 2010.

[2] "PURSUIT: Pursuing a pub/sub internet," http://www.fp7-pursuit.eu/, Sep. 2010.

[3] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *Communications Magazine, IEEE*, vol. 50, no. 7, July 2012.

[4] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *ACM CoNEXT*, 2009, pp. 1–12.

[5] B. Azimdoost, C. Westphal, and H. R. Sadjadpour, "On the throughput capacity of information-centric networks," in *Teletraffic Congress (ITC), 2013 25th International*. IEEE, 2013, pp. 1–9.

[6] B. Ahlgren, M. D'Ambrosio, M. Marchisio, I. Marsh, C. Dannewitz, B. Ohlman, K. Pentikousis, O. Strandberg, R. Rembarz, and V. Vercellone, "Design considerations for a network of information," in *ACM CoNEXT*, 2008, pp. 1–6.

[7] T. Koponen, M. Chawla, B. G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica, "A data-oriented (and beyond) network architecture," in *ACM SIGCOMM*, 2007, pp. 181–192.

[8] A. Ghodsi, T. Koponen, B. Raghavan, S. Shenker, A. Singla, and J. Wilcox, "Information-Centric networking: Seeing the forest for the trees," in *HotNets*, 2011.

[9] F. Olmos, B. Kauffmann, A. Simonian, and Y. Carlinet, "Catalog Dynamics: Impact of Content Publishing and Perishing on the Performance of a LRU Cache," Sep. 2014.

[10] H. Che, Z. Wang, and Y. Tung, "Analysis and design of hierarchical web caching systems," in *IEEE INFOCOM*, 2001, pp. 1416–1424.

[11] E. Rosensweig, J. Kurose, and D. Towsley, "Approximate models for general cache networks," in *IEEE INFOCOM*, 2010, pp. 1–9.

[12] A. Wolman, M. Voelker, N. Sharma, N. Cardwell, A. Karlin, and H. M. Levy, "On the scale and performance of cooperative Web proxy caching," *SIGOPS Oper. Syst. Rev.*, vol. 33, no. 5, pp. 16–31, Dec. 1999.

[13] E. J. Rosensweig and J. Kurose, "Breadcrumbs: Efficient, Best-Effort content location in cag networks," in *IEEE INFOCOM*, 2009, pp. 2631–2635.

[14] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks," *IEEE Transactions on Mobile Computing*, no. 1, pp. 77–89, 2005.

[15] M. Tortelli, I. Cianci, L. A. Grieco, G. Boggia, and P. Camarda, "A fairness analysis of content centric networks," Nov. 2011.

[16] C. Westphal, "On maximizing the lifetime of distributed information in ad-hoc networks with individual constraints," in *ACM MobiHoc*, 2005, pp. 26–33.

[17] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the Complexity of Optimal Routing and Content Caching in Heterogeneous Networks," Dec. 2014.

[18] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, 2000.

[19] J. Li, C. Blake, D. S. De Couto, H. I. Lee, and R. Morris, "Capacity of ad hoc wireless networks," in *MobiCom*, 2001, pp. 61–69.

[20] U. Niesen, P. Gupta, and D. Shah, "On capacity scaling in arbitrary wireless networks," *Information Theory, IEEE Transactions on*, vol. 55, no. 9, pp. 3959–3982, 2009.

[21] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad hoc wireless networks," *Networking, IEEE/ACM Transactions On*, vol. 10, no. 4, pp. 477–486, 2002.

[22] J. D. Herdtner and E. K. Chong, "Throughput-storage tradeoff in ad hoc networks," in *IEEE INFOCOM*, 2005, pp. 2536–2542.

[23] G. Alfano, M. Garetto, and E. Leonardi, "Content-Centric Wireless Networks With Limited Buffers: When Mobility Hurts," *IEEE/ACM Transactions on Networking*, vol. 20, Oct. 2014.

[24] H. Liu, Y. Zhang, and D. Raychaudhuri, "Performance evaluation of the cache-and-forward (CNF) network for mobile content delivery services," in *ICC Workshop*, 2009, pp. 1–5.

[25] G. Carofiglio, M. Gallo, L. Muscariello, and D. Perino, "Modeling data transfer in content-centric networking," in *IEEE Teletraffic Congress (ITC23)*, 2011, pp. 111–118.

[26] U. Niesen, D. Shah, and G. Wornell, "Caching in wireless networks," *IEEE Transactions on Information Theory*, 2011.

[27] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic Laws for Joint Content Replication and Delivery in Wireless Networks," *Information Theory, IEEE Transactions on*, vol. 59, no. 5, pp. 2760–2776, May 2013.

[28] F. Xue and P. Kumar, *Scaling Laws for Ad Hoc Wireless Networks: an Information Theoretic Approach*. Foundations and Trends in Networking, NOW Publishers, 2006.

[29] M. D. Penrose, "The longest edge of the random minimal spanning tree," *The Annals of Applied Probability*, pp. 340–361, 1997.

[30] H. Che, Y. Tung, and Z. Wang, "Hierarchical Web caching systems: modeling, design and experimental results," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305–1314, Sep. 2002.