The Effect of Caching in Sustainability of Large Wireless Networks

G. S. Paschos, S. Gitzenis, and L. Tassiulas,

***Abstract*—We study the scalability of multihop wireless communications, a major concern in networking, for the case that users access content cached across the nodes. In contrast to the standard paradigm of randomly selected communicating pairs, content replication is efficient for certain regimes of content volume and popularity, cache and network size. Assuming the Zipf popularity law, and investigating on the relative ways that the number of files, the cache size and the network nodes can all jointly scale to infinity, we derive asymptotic laws on required link capacity, which range from $O\left(\sqrt{N}\right)$ down to $O(1)$, and identify regimes of network operation.**

I. Introduction

The proliferation of video applications and the advent of new paradigms like multiview and 3D video, as well as other demanding applications push the operation of networks to their physical limits. To overcome these challenges, the networking community devises new technologies and architectures like the Peer-to-Peer (P2P) communication paradigm and the Content Centric Networking in an effort to improve the scalability and the efficiency for the Internet of the future.

In this landscape, wireless networks are considered to hold an important role, supporting mobility of users, extending network connectivity and promoting ubiquitous computing. According to [1], traffic from wireless devices will exceed traffic from wired ones by 2015.

Despite their worldwide deployment, wireless networks are mostly confined to one-hop access from the wired backhaul. Multihop operation of wireless networks is limited to specific applications like sensor networks where the supported communication rates are low, or engineered fixed point-to-point links with directional antennas. In the seminal work of Gupta-Kumar [2], the traffic-carrying capacity of a planar wireless network was shown to be $O\left(\frac{1}{\sqrt{N}}\right)$, where N is the number of nodes. This implies that large multihop networks cannot sustain throughput among random pairs due to the increasing hop number between the source and the destination.

There is anecdotal evidence that the average P2P video file travels back and forth on the same optical link multiple times causing detrimental effects to the network efficiency. This effect has a very large volume if one considers that video data account for almost 50% of the overall network traffic today, [1]. In this context, *network caching* has a key role, as it can mitigate these inefficiencies by storing the data close to the customer and avoiding excess traffic, thereby increasing significantly the efficiency of the network.

In the wireless networks of the future, one can envisage the direct participation of a myriad of computing devices with variable cache capabilities. These devices are expected to form new paradigms of networks based on multihop wireless connectivity, and deliver high quality services as the ones already described. Moreover, due to the ongoing research on memory and storage technologies, the size of these caches is expected to increase geometrically over time[1]. The question we study is *whether the trend of increasing cache size is sufficient to bring a measurable improvement in the operation of wireless networks*, and, in particular, to change the asymptotic law of the wireless network capacity.

In this work, we depart from the random-pairs communicating paradigm of [2] and important follow-ups [4]–[8], by defining a content-based communication paradigm, where nodes request content replicated inside the network, in the caches of other nodes. This paradigm gives rise to the joint problem of *replication* and *routing*. Assuming the symmetric topology of the square grid (a well-accepted model for various planar wireless networks), we set up a replication problem whose optimal solution results in the same order to the complex combinatorial joint problem. Then, we use this solution to identify the asymptotic laws of the required link capacity that can sustain the associated network traffic.

In contrast to our prior works' perspective of [9], [10], the cache size is assumed to increase to infinity. In such a study, the statistics of the applications are quite important. Therefore, in this work, we assume that the requested messages have a popularity described by the Zipf law with parameter τ, a well-known approach for modeling file popularity.

Due to space constraints, the proofs are omitted from this version of the paper, and will be appearing in an upcoming extended version (to assist the reviewer, we keep the appendix in this document). Table I provides the definitions on the asymptotic notation that we use throughout this work.

II. Basic Definitions and the Density Problem

Assume a square lattice with N nodes, with N being the square of an integer; the set of nodes is indexed by $n \in \mathcal{N} \triangleq \{1, 2, \ldots, N\}$. Each node is connected to its four neighbors that lie next to it on the same row or column with undirected links. By keeping the node density fixed and increasing the network size N, we obtain a scaling network similar to [2]. Moreover, for simplicity, we consider a toroidal structure as in [11] to avoid boundary effects.

[1]After a period of doubling of the areal density of hard disks per year, the growth rate has dropped to doubling every three years. Similarly, DRAM capacity quadruples every three years [3].

TABLE I
DEFINITION OF ASYMPTOTIC NOTATION (f and g are positive functions).

$f = o(g)$	For any $k > 0$, there exists $\hat{x}$:	$x \geq \hat{x} \Rightarrow f(x) \leq kg(x)$
$f = O(g)$	There exists $k > 0$ and $\hat{x}$:	
$f \overset{\lim}{\leq} kg$	There exists $\hat{x}$:	
$f \overset{\lim}{<} k'g$	For any $0<k<k'$, there exists $\hat{x}$:	
$f = \omega(g)$	For any $k > 0$, there exists $\hat{x}$:	$x \geq \hat{x} \Rightarrow f(x) \geq kg(x)$
$f = \Omega(g)$	There exists $k > 0$ and $\hat{x}$:	
$f \overset{\lim}{\geq} kg$	There exists $\hat{x}$:	
$f \overset{\lim}{>} k'g$	For any $k > k'$, there exists $\hat{x}$,	
$f = \Theta(g)$	Iff $f = O(g)$ and $f = \Omega(g)$	

Nodes (or users located therein) generate requests to access files/data, indexed by $m \in \mathcal{M} \triangleq \{1, 2, \ldots, M\}$. Each node n is equipped with a cache/buffer, whose contents are denoted by the set $\mathcal{B}_n$, a subset of $\mathcal{M}$. If a request at node n regards a file m that lies in $\mathcal{B}_n$, then it is served locally. Due to the limited buffer capacity, m will often be not available locally, thus, node n will have to retrieve m over the network from some other node w that keeps m in its cache. Thus for each n, m pair, a route (or set of routes) $\mathcal{R}_{n,m}$ should be decided to specify the path(s) followed from n to w in accessing m.

Let, moreover, K be the storage capacity of nodes' cache measured in the number of files it can store. This means that all M files are of the same (unit) size, placing a constraint on the cardinality of cache contents $|\mathcal{B}_n| \leq K$. The generalization to variable sized files can be still captured in this framework by splitting each large file into multiple unit segments, and then treating its segments as separate, independent files.

For the problem of replication not to be trivial, it should be

$$K < M, \tag{1}$$

which implies that each node has to select the files to buffer in its cache. Moreover, for the network to have sufficient memory to store each file at least once, it should be

$$KN \geq M. \tag{2}$$

Last, let each node $n \in \mathcal{N}$ generate requests for data at rate λ_n. In this work, we focus on the symmetric node request rate, that is, $\lambda_n = \lambda = 1$. Each request regards a particular file $m \in \mathcal{M}$, depending on the file m's *popularity* p_m. In essence, $[p_m]$ is a probability distribution, i.e., sets the probability of a request for a given file. Clearly, replication should be governed by the popularity: storing the popular files densely will maximize the gain of caching on the network traffic.

A. General Replication-Delivery Problem

Assuming a $[\mathcal{B}_n]$ replication and delivery routes $[\mathcal{R}_{n,m}]$, it is easy to compute the traffic load C_ℓ at each link ℓ of the network. The associated replication-delivery problem regards minimizing over the worst (or, in a relaxed form, average) traffic C_ℓ as in [9], [10] given the constraints of (i) node capacity and (ii) storing at least one copy of each file over the network. The resulting C_ℓ, then, sets the minimum *capacity* of each link so that the network operates properly (i.e., is stable).

Obviously, the entanglement between the positions $[\mathcal{B}_n]$ where each file is stored and the delivery paths $[\mathcal{R}_{n,m}]$ calls for a joint optimization. However, this is clearly a combinatorial complexity problem, and thus not amenable to an easy to compute solution. Therefore, we should seek suitable simplifications and approximations to derive a suboptimal but efficient solution. For the needs of our study, this translates to an order-optimal solution, i.e., the proposed suboptimal solution lies within a constant to the optimal (but hard to compute) solution.

B. Replication Density-based Problem

Assuming a solution on the general problem, we can define a particularly important quantity, the frequency of occurrence of each file m in the caches, or *replication density* d_m as the fraction of nodes that store file m in the network:

$$d_m = \frac{1}{N} \sum_{n \in \mathcal{N}} \mathbb{1}_{\{m \in \mathcal{B}_n\}}. \tag{3}$$

Based on this metric, we can define a much simpler problem based on the the file densities:

PROBLEM 1: *Minimize* $C \triangleq \sum_{m \in \mathcal{M}} \left[\frac{1}{\sqrt{d_m}} - 1 \right] p_m$, *s.t.*

1) *For any* $m \in \mathcal{M}$, $\frac{1}{N} \leq d_m \leq 1$,
2) $\sum_{m \in \mathcal{M}} d_m \leq K$.

In the above, the optimization variables are the densities d_m, which express the fraction of caches containing file m. In the objective, $d_m^{-\frac{1}{2}} - 1$ approximates the average hop count from a random node to a cache containing m. Weighted by the probability p_m of requests on m, the summation expresses the average link load per request.

This optimization is shown in [10] to be a relaxed version of the actual general problem, and, moreover, whose optimal solution C is of the same order to the solution of the original problem; in particular, [10] presents an algorithm to assign the node cache contents $[\mathcal{B}_n]$ from the densities d_m and uses shortest path routing for the delivery paths $[\mathcal{R}_{n,m}]$. Thus, the asymptotic laws of the original problem and of C coincide, and, therefore, it suffices to study C's scaling.

It should be noted that a similar optimization is formulated in [12], without, however, the $\frac{1}{N} \leq d_m \leq 1$ constraints. As seen next, these inequalities have a major impact on the solution, and, consequently, in the asymptotics.

C. Density Problem Solution

As explained in [9], [10], and easily seen from the functional form, the density problem admits a unique solution using the Karush-Kuhn-Tucker (KTT) conditions, and a computationally efficient algorithm which finds the solution in polynomial time. With regard to the constraints on d_m about its minimum and maximum value, either one of them can be an equality, or none. This causes the partition of $\mathcal{M}$ into three subsets, one containing files of unit replication density (i.e., stored at every node) $\mathcal{M}_{\uparrow} = \{m : d_m = 1\}$, one containing files stored in just one node $\mathcal{M}_{\downarrow} = \{m : d_m = \frac{1}{N}\}$, and the complementary

$\mathcal{M}_{\updownarrow} = \mathcal{M} \setminus (\mathcal{M}_{\uparrow} \cup \mathcal{M}_{\downarrow})$. When p_m is in decreasing order, these sets become ordered, too; then, we use variables l and r to identify the boundaries of these sets, as follows: $\mathcal{M}_{\uparrow} = \{1, 2, \ldots, l-1\}$, $\mathcal{M}_{\updownarrow} = \{l, l+1, \ldots, r-1\}$, and $\mathcal{M}_{\downarrow} = \{r, r+2, \ldots, M\}$, where l and r are integers with $1 \leq l \leq r \leq M+1$. Given these, the solution d_m takes the form of

$$d_m = \begin{cases} 1, & m \in \mathcal{M}_{\uparrow}, \quad (4a) \\ \dfrac{K - l + 1 - \frac{M-r+1}{N}}{\sum_{j=l}^{r-1} p_j^{\frac{2}{3}}} \, p_m^{\frac{2}{3}}, & m \in \mathcal{M}_{\updownarrow}, \quad (4b) \\ \dfrac{1}{N}, & m \in \mathcal{M}_{\downarrow}. \quad (4c) \end{cases}$$

III. Asymptotic Laws for Zipf Popularity

To derive concrete asymptotics of the link rate, we consider the Zipf law, a distribution well-known for the Internet's traffic.

A. Zipf Law and Approximations

The Zipf distribution is defined as follows:

$$p_m = \frac{1}{H_\tau(M)} m^{-\tau}, \quad (5)$$

where τ is the power law parameter, indicating the rate of popularity decline as m increases, and $H_\tau(n) \triangleq \sum_{j=1}^{n} j^{-\tau}$ is the truncated (at n) zeta function evaluated at τ (also called the n^{th} *τ-order generalized harmonic number*). The limit $H_\tau \triangleq \lim_{n\to\infty} H_\tau(n)$ is the Riemann zeta function, which converges when $\tau > 1$. We derive an approximation for $H_\tau(n)$ by bounding the sum: for $n \geq m \geq 0$,

$$\int_m^n (x+1)^{-\tau} dx \leq H_\tau(n) - H_\tau(m) \leq 1 + \int_{m+1}^{n} x^{-\tau} dx, \Rightarrow$$
$$\begin{cases} \frac{(n+1)^{1-\tau} - (m+1)^{1-\tau}}{1-\tau} \leq H_\tau(n) - H_\tau(m) \leq \frac{n^{1-\tau} - (m+1)^{1-\tau}}{1-\tau} + 1, & \text{if } \tau \neq 1, \\ \ln \frac{n+1}{m+1} \leq H_\tau(n) - H_\tau(m) \leq \ln \frac{n+1}{m+2}, & \text{if } \tau = 1. \end{cases} \quad (6)$$

For any $m < n$ such that $n \sim m$, we will make use of the following approximation derived by counting the sum terms

$$H_\tau(n) - H_\tau(m) = \sum_{j=m}^{n} j^{-\tau} \sim n^{-\tau}(n-m). \quad (7)$$

Substituting the solution (4) and plugging in the Zipf distribution into C, it follows that

$$C \triangleq \sum_{m \in \mathcal{M}} \left(d_m^{-\frac{1}{2}} - 1 \right) p_m = C_{\updownarrow} + C_{\downarrow} - \sum_{j=l}^{M} p_m, \quad (8)$$

where $\sum_{j=l}^{M} p_m = O(1)$ (as it lies always in $[0, 1]$), and

$$C_{\updownarrow} \triangleq \sum_{m \in \mathcal{M}_{\updownarrow}} \frac{p_m}{\sqrt{d_m}} \overset{(5)}{=} \frac{\left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1) \right]^{\frac{3}{2}}}{\sqrt{K - l + 1 - \frac{M-r+1}{N}} \, H_\tau(M)}, \quad (9)$$

$$C_{\downarrow} \triangleq \sum_{m \in \mathcal{M}_{\downarrow}} \frac{p_m}{\sqrt{d_m}} \overset{(5)}{=} \sqrt{N} \, \frac{H_\tau(M) - H_\tau(r-1)}{H_\tau(M)}. \quad (10)$$

B. Estimation of l and r

As indices l, r are not provided in a closed form, we derive approximations in order to find the scaling of C.

1) Estimation of l: note that $l-1$ represents the number of files cached in all nodes. If $\mathcal{M}_{\updownarrow} \cup \mathcal{M}_{\downarrow}$ is not empty, $d_l < 1 \overset{(4b)}{\Leftrightarrow}$

$$K - l + 1 - \frac{M-r+1}{N} < l^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1) \right]. \quad (11)$$

If, moreover, the first set $\mathcal{M}_{\uparrow}$ is not empty, i.e., $l > 1$, then $d_{l-1} = 1$. This means that if we attempted to decrease index l by 1, this would violate the density constraints, and result in (4b) a number greater than 1 for d_{l-1}:

$$K - l - \frac{M-r+1}{N} \geq (l-1)^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-2) \right]. \quad (12)$$

Thus, provided $l > 1$, it can be uniquely determined as the lowest integer that satifisfies (11)-(12). An approximation for l can be computed treating (11) as an approximate equality (as $d_{l-1} = 1$ and $d_l < 1$):

$$K - l + 1 - \frac{M-r+1}{N} \cong l^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1) \right]. \quad (13)$$

2) Estimation of r: If $\mathcal{M}_{\updownarrow} \cup \mathcal{M}_{\downarrow}$ is not empty, $d_{r-1} > \frac{1}{N} \Leftrightarrow$

$$(K - l + 1)N - M + r - 1 > (r-1)^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1) \right]. \quad (14)$$

Again, if set $\mathcal{M}_{\downarrow}$ is not empty, i.e., $r \leq M$, then $d_r = N^{-1}$. Thus, if we attempted increasing index r by one, (4b) would violate the constraint resulting in a density less than N^{-1}:

$$(K - l + 1)N - M + r \leq r^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r) - H_{\frac{2\tau}{3}}(l-1) \right]. \quad (15)$$

As before, (14) is an approximate equality, i.e.,

$$(K - l + 1)N - M + r - 1 \cong (r-1)^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1) \right]. \quad (16)$$

3) Estimation of $\frac{l}{r}$: For all l, r, it is $N > \frac{d_l}{d_{r-1}} = \left(\frac{r-1}{l} \right)^{\frac{2\tau}{3}}$. As before, whenever l and r are not equal to the extremes, i.e., $1 < l \leq r < M+1$, it holds $d_{l-1}/d_r = N$. Hence,

$$l \cong r N^{-\frac{3}{2\tau}}. \quad (17)$$

Up to this point, we have summarized the analysis of [9], [10]. Departing from them, we investigate the behavior of l and r as N, M and K go to infinity.

C. Almost empty $\mathcal{M}_{\downarrow}$

The first case of interest is when the solution results in an almost empty set $\mathcal{M}_{\downarrow}$. Formally, $\mathcal{M}_{\downarrow} \approx \emptyset$ iff $|\mathcal{M}_{\downarrow}| = o(M)$, i.e., the number of last set's elements over the total files is negligible; thus, $\mathcal{M}_{\downarrow} = \emptyset$ is a special case of $\mathcal{M}_{\downarrow} \approx \emptyset$.

For $\mathcal{M}_{\downarrow} \approx \emptyset$, M should increase at a slow pace in respect to N and K, so that the constraint $d_m \geq N^{-1}$ is satisfied for almost all (i.e., $M - o(M)$) files. Since $|\mathcal{M}_{\downarrow}| = M - r + 1$, this condition is equivalent to $M - r = o(M)$.

Lemma 1 [l, r and conditions for almost Empty $\mathcal{M}_{\downarrow}$]: *If $\mathcal{M}_{\downarrow} \approx \emptyset$, then $r \sim M$ and*

- *for $\tau < 3/2$, it is*

$$\begin{cases} l \to 1, & \text{if } K \overset{\lim}{<} M^{1-\frac{2\tau}{3}} \\ l \sim \left(\frac{3-2\tau}{3}\right)^{\frac{3}{2\tau}} \frac{K^{\frac{3}{2\tau}}}{M^{\frac{3}{2\tau}-1}}, & \text{if } M^{1-\frac{2\tau}{3}} \overset{\lim}{\leq} K = o(M), \\ l \sim \alpha K, & \text{if } K \sim \beta_\tau M, \end{cases}$$

where $\alpha \in (0,1]$, $\beta_{\alpha,\tau} \triangleq \alpha^{\frac{2\tau}{3-2\tau}} \left[\frac{3-2\tau(1-\alpha)}{3}\right]^{\frac{-3}{3-2\tau}}$ *and*

$$\begin{cases} \mathcal{M}_{\downarrow} = \emptyset, & \text{if } \omega(K) = M \overset{\lim}{<} \frac{3-2\tau}{3} KN, \\ \mathcal{M}_{\downarrow} \approx \emptyset, & \text{if } \omega(K) = M \overset{\lim}{\leq} \frac{3-2\tau}{3} KN, \\ \mathcal{M}_{\downarrow} = \emptyset, & \text{if } K = \Theta(M); \end{cases}$$

- *for $\tau = 3/2$, it is*

$$\begin{cases} l \to 1 & \text{if } K \overset{\lim}{\leq} \ln M, \\ l \sim \frac{K}{\ln M} & \text{if } \ln M \overset{\lim}{<} K = o(M), \\ l \sim \alpha K, & \text{if } K \sim \gamma_\alpha M, \end{cases}$$

where $\alpha \in (0,1]$, $\gamma_\alpha \triangleq \frac{1}{\alpha} e^{\frac{\alpha-1}{\alpha}}$, *and*

$$\begin{cases} \mathcal{M}_{\downarrow} = \emptyset, & \text{if } M = \omega(K), \text{ and } M \ln M \overset{\lim}{<} KN, \\ \mathcal{M}_{\downarrow} \approx \emptyset, & \text{if } M = \omega(K), \text{ and } M \ln M \overset{\lim}{\leq} KN, \\ \mathcal{M}_{\downarrow} = \emptyset, & \text{if } K = \Theta(M); \end{cases}$$

- *for $\tau > 3/2$, it is*

$$\begin{cases} l \sim \frac{2\tau-3}{2\tau} K, & \text{if } K = o(M), \\ l \sim \alpha K, & \text{if } K \sim \delta_{\alpha,\tau} M, \end{cases}$$

where $\alpha \in \left(\frac{2\tau-3}{2\tau}, 1\right]$, $\delta_{\alpha,\tau} \triangleq \alpha^{\frac{2\tau}{3-2\tau}} \left(\frac{3-2\tau(1-\alpha)}{3}\right)^{\frac{-3}{3-2\tau}}$ *and*

$$\begin{cases} \mathcal{M}_{\downarrow} = \emptyset, & \text{if } \omega(K) = M \overset{\lim}{<} \frac{2\tau-3}{2\tau} KN^{\frac{3}{2\tau}}, \\ \mathcal{M}_{\downarrow} \approx \emptyset, & \text{if } \omega(K) = M \overset{\lim}{\leq} \frac{2\tau-3}{2\tau} KN^{\frac{3}{2\tau}}, \\ \mathcal{M}_{\downarrow} = \emptyset, & \text{if } K = \Theta(M). \end{cases}$$

Note, that we have $K - l + 1 = \Theta(K)$ if $K \overset{\lim}{<} M$. If, however $K \sim M$, then $l \sim K \sim M$, in which case the majority of files are stored locally.

D. Non-empty $\mathcal{M}_{\downarrow}$

If $\mathcal{M}_{\downarrow}$ is non-empty, then $M - r = \Theta(M)$.

LEMMA 2 [l AND r FOR NON-EMPTY $\mathcal{M}_{\downarrow}$]: *If $\mathcal{M}_{\downarrow} \neq \emptyset$, and $KN - M = O(1)$, then $l \to 1$ and $r = \Theta(1)$; in particular,*

$$\begin{cases} r \approx 1 + \frac{3-2\tau}{2\tau}(KN - M), & \text{if } \tau < 3/2, \\ (r-1)\ln(r-1) \approx KN - M, & \text{if } \tau = 3/2, \\ r \approx 1 + \frac{2\tau-3}{2\tau} \frac{KN-M}{N^{1-\frac{3}{2\tau}}} & \text{if } \tau > 3/2. \end{cases}$$

Else, if $\mathcal{M}_{\downarrow} \neq \emptyset$, and $KN - M = \omega(1)$, then

- *for $\tau < 3/2$,*

$$\begin{cases} \text{if } KN - M \overset{\lim}{\leq} \frac{2\tau}{3-2\tau} N^{\frac{3}{2\tau}}, \\ \quad \text{then } l \to 1, & r \sim \frac{3-2\tau}{2\tau}(KN - M), \\ \text{if } KN - M \overset{\lim}{>} \frac{2\tau}{3-2\tau} N^{\frac{3}{2\tau}}, \\ \quad \text{then } l \sim \frac{3-2\tau}{2\tau} \frac{KN-M}{N^{\frac{3}{2\tau}}}, & r \sim \frac{3-2\tau}{2\tau}(KN - M), \end{cases}$$

- *for $\tau = 3/2$,*

$$\begin{cases} \text{if } KN - M \overset{\lim}{\leq} N \ln N, \\ \quad \text{then } l \to 1, & r \ln r \sim KN - M, \\ \text{if } KN - M \overset{\lim}{>} N \ln N, \\ \quad \text{then } l \sim \frac{KN-M}{N \ln N}, & r \sim \frac{KN-M}{\ln N}, \end{cases}$$

- *for $\tau > 3/2$*

$$\begin{cases} \text{if } KN - M \overset{\lim}{\leq} \frac{2\tau}{2\tau-3} N, \\ \quad \text{then } l \to 1, & r \sim \left(\frac{2\tau-3}{2\tau}\right)^{\frac{3}{2\tau}} (KN - M)^{\frac{3}{2\tau}}, \\ \text{if } KN - M \overset{\lim}{>} \frac{2\tau}{2\tau-3} N, \\ \quad \text{then } l \sim \frac{2\tau-3}{2\tau}\left(K - \frac{M}{N}\right), & r \sim \frac{2\tau-3}{2\tau} \frac{KN-M}{N^{1-\frac{3}{2\tau}}}, \end{cases}$$

Note, that for all the cases we have $K - l + 1 = \Theta(K)$.

E. Capacity scaling

We proceed to the asymptotic behavior of the system on the rate C, which regards the case of the number of nodes N, the number of files M and the size of caches K, all increasing to infinity in various relative rates.

First, we establish the Gupta-Kumar rate [2] $O\left(\sqrt{N}\right)$ as an upper bound. This is intuitive: if replication is ineffective (e.g., due to large number of files or small size of caches), then the system and its performance essentially reduce to [2].

LEMMA 3 [UPPER BOUND ON C]: $C = O\left(\sqrt{N}\right)$.

Next, we begin the asymptotic analysis, partitioning the space of M, N, K parameters to whether they produce single replicated files (non-empty $\mathcal{M}_{\downarrow}$) or not. To study the asymptotics of C, we use the results for l and r obtained in Lemmas 1 and 2. The almost empty $\mathcal{M}_{\downarrow}$ implies that $M - r = o(M)$ and there are very few files stored once in the network in which case the required link capacity law depends on $C_{\wr}$.

THEOREM 4 [CAPACITY FOR ALMOST EMPTY $\mathcal{M}_{\downarrow}$]: *if $K \sim M$, then $C = o(1)$. Otherwise,*

- *if $\tau < 1$,* $C = \Theta\left(\frac{\sqrt{M}}{\sqrt{K}}\right)$,
- *if $\tau = 1$,* $C = \Theta\left(\frac{\sqrt{M}}{\sqrt{K}\log M}\right)$,
- *if $1 < \tau < \frac{3}{2}$,* $C = \Theta\left(\frac{M^{\frac{3}{2}-\tau}}{\sqrt{K}}\right)$,
- *if $\tau = \frac{3}{2}$, and $K = \Theta(M)$,* $C = \Theta\left(\frac{1}{\sqrt{K}}\right)$,

- *if* $\tau = \frac{3}{2}$, *and* $K = o(M)$, $\quad C = \Theta\left(\frac{\log^{3/2} M}{\sqrt{K}}\right)$,
- *if* $\tau > \frac{3}{2}$, $\quad C = \Theta\left(\frac{1}{\sqrt{K}}\right)$.

THEOREM 5 [CAPACITY FOR NON-EMPTY $\mathcal{M}_{\downarrow}$]:

- *If* $\tau < 1$, $\quad C = \Theta\left(\sqrt{N}\right)$,
- *if* $\tau = 1$, *and* $M \overset{\lim}{<} KN$, $\quad C = \Theta\left(\frac{\sqrt{N}}{\log M}\right)$,
- *if* $\tau = 1$, *and* $M \sim KN$, $\quad C = \Theta\left(\sqrt{N}\right)$,
- *if* $1 < \tau < \frac{3}{2}$, $\quad C = \Theta\left(\frac{\sqrt{N}}{(KN-M)^{\tau-1}}\right)$,
- *if* $\tau = \frac{3}{2}$, $\quad C = \Theta\left(\frac{\log^{\frac{3}{2}} r}{\sqrt{K-\frac{M}{N}}}\right)$,
- *if* $\tau > \frac{3}{2}$, *and* $KN - M \overset{\lim}{\leq} \frac{2\tau}{2\tau-3} N$, $C = \Theta\left(\frac{\sqrt{N}}{\sqrt{NK-M}}\right)$,
- *if* $\tau > \frac{3}{2}$, *and* $KN - M \overset{\lim}{>} \frac{2\tau}{2\tau-3} N$, $C = \Theta\left(\frac{N^{\tau-1}}{(NK-M)^{\tau-1}}\right)$.

IV. DISCUSSION ON ASYMPTOTIC LAWS

The main result of the asymptotic laws regards the minimum link rate (of the bottleneck link) required to sustain a constant request rate from each node. As a preliminary comment, the link rates are subject to the information theory, e.g. Shannon's capacity law. Thus, a rate C that scales to infinity should be rather interpreted as the inverse of the sustainable request rate λ, e.g., a result $C = \Theta(\sqrt{N})$ for $\lambda = 1$ is equivalent to $C = \Theta(1)$ for the Gupta-Kumar law of $\lambda = \Theta\left(\frac{1}{\sqrt{N}}\right)$.

Power law parameter τ sets two phase transition points for the values of 1 and $\frac{3}{2}$, leading to distinct asymptotics: the higher τ, the more uneven the popularity of files, and thus, the more advantageous the caching (i.e. lower rate C). As summarized in Table II, $C = O(1)$ on $\tau > \frac{3}{2}$, or, i.e., the wireless network is sustainable (it corresponds to a traffic carrying capacity of $O(1)$ in the Gupta-Kumar setup).In real systems, the Zipf parameter ranges typically from 0.5 [13] to 3 [14] depending on the application: low values are typical in routers, intermediate values in proxies and higher values in mobile applications [15], [16]–see also references therein.

More common are the cases of low and intermediate values of τ (representative also of the whole file population without any application bias), which flatten the popularity distribution towards the uniform. Replication is less effective, ending up to the $\Theta\left(\sqrt{N}\right)$ law for $\tau < 1$, a synonym for the Gupta-Kumar law. When $M \overset{\lim}{\leq} \frac{3-2\tau}{3} KN$, then only a few files are cached once (the condition $\mathcal{M}_{\downarrow} \approx \emptyset$) in which case there is an improvement over [2], see Theorem 4.

Comparing our results to [9], [10], we note the differences due to node cache capacity going to infinity as well. First, when $\mathcal{M}_{\downarrow} \approx \emptyset$, the improvement is significant. More precisely, the term $\frac{1}{\sqrt{K}}$ multiplies the asymptotic law, suggesting that the required link rate of the bottleneck can be partially mitigated by investment in caching. On the other hand, when $\mathcal{M}_{\downarrow} \neq \emptyset$, the term K appears only for $\tau > 1$, in the form of $KN - M$. Note, however, that the condition for $\mathcal{M}_{\downarrow} \approx \emptyset$ depends itself on K, see Lemma 1. In particular, a sufficient increase in K can guarantee this condition. Thus, if $\tau \leq 1$, investment in cache size makes sense only if it suffices to guarantee $\mathcal{M}_{\downarrow} \approx \emptyset$.

V. CONCLUSIONS & FUTURE WORK

In this work, we investigated the effect of caching in the asymptotic capacity of wireless networks under the paradigm of content replication and delivery. We showed that depending on the file popularity distribution, there exist regimes of network expansion where caching can be effective tool in mitigating the problem of multihop wireless networks sustainability. More precisely, in the regime $\mathcal{M}_{\downarrow} \approx \emptyset$, increasing the cache size brings a $\frac{1}{\sqrt{K}}$ multiplicative term in the required capacity of the bottleneck link. Also, if $\mathcal{M}_{\downarrow} \neq \emptyset$, increasing the cache size is helpful if $\tau > 1$.

A future extension of this work will focus on establishing the result on non-symmetric topologies and arrival conditions. Also, we are interested to investigate the effect of in-network caching in medium-sized wireless networks.

REFERENCES

[1] Cisco, "*Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20102015*," White Paper, 2011, Tech. Rep.

[2] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, pp. 388–404, Mar. 2000.

[3] J. Hennesy and D. Patterson, *Computer Architecture: A Quantitative Approach*. San Francisco, CA, USA: Morgan-Kauffman publishers, 4th edition, 2007.

[4] A. Zemlianov and G. de Veciana, "Capacity of ad hoc wireless networks with infrastructure support," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 657–667, Mar. 2005.

[5] A. Özgür, O. Lévêque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3549–3572, Oct. 2007.

[6] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inf. Theory*, vol. 53, pp. 1009–1018, Mar. 2007.

[7] S. Toumpis, "Asymptotic capacity bounds for wireless networks with non-uniform traffic patterns," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2231–2242, Jun. 2008.

[8] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: information-theoretic and physical limits," *IEEE Trans. Inf. Theory*, vol. 55, pp. 3413–3424, Aug. 2009.

[9] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for content replication and delivery in wireless networks," in *Proc. of INFOCOM*, Orlando, FL, USA, Mar. 2011.

[10] ——, "*Asymptotic Laws for Joint Content Replication and Delivery in Wireless Networks*," arXiv:1201.3095v1 [cs.NI], Tech. Rep.

[11] M. Franceschetti and R. Meester, *Random Networks for Communication*. New York, NY, USA: Cambridge University Press, Series: Cambridge Series in Statistical and Probabilistic Mathematics (No. 24), 2007.

[12] S. Jin and L. Wang, "Content and service replication strategies in multihop wireless mesh networks," in *MSWiM '05: Proceedings of the 8th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, Montréal, Quebec, Canada, Oct. 2005, pp. 79–86.

[13] J. Chu, K. Labonte, and B. N. Levine, "Availability and popularity measurements of peer-to-peer file systems," in *Proceedings of SPIE*, Boston, MA, USA, Jul. 2002.

[14] T. Yamakami, "A Zipf-like distribution of popularity and hits in the mobile web pages with short life time," in *Proc. of Parallel and Distributed Computing, Applications and Technologies, PDCAT '06*, Taipei, ROC, Dec. 2006, pp. 240–243.

[15] L. Breslau, P. Cue, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. of INFOCOM*, New York, NY, USA, Mar. 1999, pp. 126–134.

[16] C. R. Cunha, A. Bestavros, and M. E. Crovella, "Characteristics of WWW Client-based Traces," in *View on NCSTRL*, Boston University, MA, USA, Jul. 1995.

TABLE II

(a) The Cases of $\tau < 1$, $\tau = 1$ and $1 < \tau < \frac{3}{2}$.

$M \div K \div N$		$K \sim M$	$K \sim \beta_{\alpha,\tau} M$	$M^{1-\frac{2\tau}{3}} \overset{\lim}{\leq} K$	$M^{1-\frac{2\tau}{3}} \overset{\lim}{>} K$	$M \overset{\lim}{<} \frac{3-2\tau}{3} KN$	$M \sim \frac{3-2\tau}{3} KN$	$M \overset{\lim}{>} \frac{3-2\tau}{3} KN$ and $KN-M \overset{\lim}{>} \frac{2\tau}{3-2\tau} N^{\frac{3}{2\tau}}$	$M \overset{\lim}{>} \frac{3-2\tau}{3} KN$ and $\frac{2\tau}{3-2\tau} N^{\frac{3}{2\tau}} \overset{\lim}{\geq} KN-M = \omega(1)$	$M \overset{\lim}{>} \frac{3-2\tau}{3} KN$ and $KN-M = O(1)$
l		$\sim K$	$\sim \alpha K$	$\sim \left(\frac{3-2\tau}{3}\right)^{\frac{3}{2\tau}} \frac{K^{\frac{3}{2\tau}}}{M^{\frac{3}{2\tau}-1}}$	$\to 1$			$\sim \frac{3-2\tau}{2\tau} \frac{KN-M}{N^{\frac{3}{2\tau}}}$	$\to 1$	
r		$M+1$				$M+1$	$\sim M$	$\sim \left(\frac{3-2\tau}{2\tau}\right)(KN-M)$		
$\mathcal{M}_{\downarrow}$		empty				empty	almost empty	non-empty		
C	$\tau<1$	$o(1)$	$\Theta\left(\sqrt{\frac{M}{K}}\right)$					$\Theta\left(\sqrt{N}\right)$		
	$\tau=1$	$o(1)$	$\Theta\left(\frac{\sqrt{M}}{\sqrt{K}\log M}\right)$					$\Theta\left(\frac{\sqrt{N}}{\log M}\right)$	$\Theta\left(\sqrt{N}\right)$	
	$1<\tau<\frac{3}{2}$	$o(1)$	$\Theta\left(\frac{M^{\frac{3}{2}-\tau}}{\sqrt{K}}\right)$					$\Theta\left(\frac{\sqrt{N}}{(KN-M)^{\tau-1}}\right)$		

(b) The Case of $\tau = \frac{3}{2}$.

$M \div K \div N$	$K \sim M$	$K \sim \gamma_\alpha M$	$\ln M \overset{\lim}{<} K$	$\ln M \overset{\lim}{\geq} K$	$M \ln M \overset{\lim}{<} KN$	$M \ln M \sim KN$	$M \ln M \overset{\lim}{>} KN$ and $KN-M \overset{\lim}{>} N \ln N$	$M \ln M \overset{\lim}{>} KN$ and $N \ln N \overset{\lim}{\geq} KN-M$
l	$\sim K$	$\sim \alpha K$	$\sim \frac{K}{\ln M}$	$\to 1$			$\sim \frac{KN-M}{N \ln N}$	$\to 1$
r	$M+1$				$M+1$	$\sim M$	$\sim \frac{KN-M}{\ln N}$	$r \ln r \sim KN-M$
$\mathcal{M}_{\downarrow}$	empty				empty	almost empty	non-empty	
C	$o\left(\frac{1}{\sqrt{K}}\right)$	$\Theta\left(\frac{1}{\sqrt{K}}\right)$	$\Theta\left(\frac{\log^{\frac{3}{2}} M}{\sqrt{K}}\right)$				$\Theta\left(\log^{\frac{3}{2}} r \sqrt{\frac{N}{KN-M}}\right)$	

(c) The Case of $\tau > \frac{3}{2}$.

$M \div K \div N$	$K \sim M$	$K \sim \delta_{\alpha,\tau} M$	$K = o(M)$	$M \overset{\lim}{<} \frac{2\tau-3}{2\tau} KN^{\frac{3}{2\tau}}$	$M \sim \frac{2\tau-3}{2\tau} KN^{\frac{3}{2\tau}}$	$M \overset{\lim}{>} \frac{2\tau-3}{2\tau} KN^{\frac{3}{2\tau}} N$ and $KN-M \overset{\lim}{>} \frac{2\tau}{2\tau-3} N$	$M \overset{\lim}{>} \frac{2\tau-3}{2\tau} KN^{\frac{3}{2\tau}} N$ and $\frac{2\tau}{2\tau-3} N \overset{\lim}{\geq} KN-M$
l	$\sim K$	$\sim \alpha K$	$\sim \frac{2\tau-3}{2\tau} K$			$\sim \frac{2\tau-3}{2\tau}\left(K - \frac{M}{N}\right)$	$\to 1$
r	$M+1$			$M+1$	$\sim M$	$\sim \frac{2\tau-3}{2\tau} \frac{KN-M}{N^{1-\frac{3}{2\tau}}}$	$\sim \left(\frac{2\tau-3}{2\tau}\right)^{\frac{3}{2\tau}} (KN-M)^{\frac{3}{2\tau}}$
$\mathcal{M}_{\downarrow}$	empty			empty	almost empty	non-empty	
C	$o\left(\frac{1}{\sqrt{K}}\right)$	$\Theta\left(\frac{1}{\sqrt{K}}\right)$				$\Theta\left(\frac{N^{\tau-1}}{(KN-M)^{\tau-1}}\right)$	$\Theta\left(\sqrt{\frac{N}{KN-M}}\right)$