On the Delay Scaling Laws of Cache Networks

Boram Jin
KAIST
boramjin@lanada.kaist.ac.kr

Daewoo Kim
KAIST
daewookim@kaist.ac.kr

Se-Young Yun
KAIST
yunseyoung@gmail.com

Jinwoo Shin
KAIST
jinwoos@kaist.ac.kr

Seongik Hong
Amazon Web Services
seongikh@amazon.com

Byoung-Joon (BJ) Lee
Creatrix Design Group
bjlee@creatrix.ca

Yung Yi
KAIST
yiyung@kaist.edu

ABSTRACT

The Internet is becoming more and more content-oriented. CDN (Content Distribution Networks) has been a popular architecture compatible with the current Internet, and a new revolutionary paradigm such as ICN (Information Centric Networking) has studied. One of the main components in both CDN and ICN is considering cache on network. Despite a surge of extensive use of cache in the current and future Internet architectures, analysis on the performance of general cache networks are still quite limited due to complex inter-plays among various components and thus analytical intractability. Due to mathematical tractability, we consider 'static' cache policies and study asymptotic delay performance of those policies in cache networks, in particular, focusing on the impact of heterogeneous content popularities and nodes' geographical 'importances' in caching policies. Furthermore, our simulation results suggest that they perform quite similarly as popular 'dynamic' policies such as LFU (Least-Frequently-Used) and LRU (Least-Recently-Used). We believe that our theoretical findings provide useful engineering implications such as when and how various factors have impact on caching performance.

CCS CONCEPTS

•Networks → Network performance modeling;

ACM Reference format:
Boram Jin, Daewoo Kim, Se-Young Yun, Jinwoo Shin, Seongik Hong, Byoung-Joon (BJ) Lee, and Yung Yi. 2017. On the Delay Scaling Laws of Cache Networks. In *Proceedings of CFI'17, Fukuoka, Japan, June 14-16, 2017,* 6 pages.
DOI: 10.1145/3095786.3095789

1 INTRODUCTION

Due to a recent shift that the Internet has increasingly become content-delivery oriented, Internet researchers constantly seek for ways of adapting the Internet to such a shift, e.g., advancing CDN technologies as an evolutionary approach, or proposing revolutionary architectures such as ICN and CCN (Content Centric Networking) [12, 18]. IP (Internet Protocol) was designed simply for host-to-host conversation, giving rise to a mismatch with content-based delivery, regarded as the root cause of several fundamental problems of the current Internet, e.g., security and mobility. ICN/CCN proposes to change the basic abstraction of the Internet from "end-to-end delivery of packets to named hosts" to "secure dissemination of named contents."

In a content-oriented architecture (whether it is evolutionary or revolutionary), content caching seems to be a crucial component to reduce delay of fetching contents and/or overall traffic transport cost, often forming a group of large-scale caches, namely a cache network. While networked caches have already appeared in the past, e.g., web caches [5, 6], they were mainly small-scale ones based on simple topological structures (e.g., hierarchical). Despite an array of recent research interests in content search, scalable architecture, and performance analysis in cache networks (see Section 2 for details), analytically understanding networked caches is known to be a daunting task in general, which leaves much to be researched in the upcoming years. The main challenge comes from complex inter-plays among several components such as content request routing, network topology, and heterogeneous per-cache budget, and dynamic cache replacement policies such as LFU (Least-Frequently-Used) and LRU (Least-Recently-Used).

1.1 Our Contribution

In this paper, we perform asymptotic delay analysis of large-scale cache networks, where quantitatively understands the relation between content popularity and delay performance as well as the impact of heterogeneity in terms of "nodes geometric importance" (i.e., caching more contents at caches with larger accessibility).

- ***Homogeneous per-node cache size.*** We observe the asymptotic delay of cache networks under homogeneous per-node cache sizes, i.e., cache sizes are uniform among nodes. We develop an analysis module (highly versatile 'black-box' tool) that provides the expected delay for a given routing distance between a content requester and the original content server, independent from the details of cache network topology and request routing policy. Our results reveal precise asymptotic performances of cache networks in terms of content popularities, content placement

CFI'17, Fukuoka, Japan

DOI: 10.1145/3095786.3095789

policies, cache sizes and number of contents, which guides to design an efficient cache network in real-world scenarios.

- ***Heterogeneous per-node cache size.*** Second, we analyze how we allocate limited cache budget, i.e., heterogeneous per-node cache size. Due to the technical hardness coming from non-trivial geometric coupling between heterogeneous per-node cache sizes and content popularities, we consider a cache network where the request routing consists of shortest-paths on a regular spanning tree under a simple cache sizing policy that has more cache sizes at geometrically important nodes. Even under this simple cache sizing policy, we could analytically achieve order-optimal delay and log-order delay reduction compared to homogeneous caching sizes.

We consider *static cache policies*, where contents are placed without replacement over time. We remark that such a static policy assumption has also been adopted in other cache network analysis research [4, 11, 16, 19] to remove the coupling of cache replacements which is regarded as one of the most challenging parts in cache network analysis. However, it does not incur too much loss of generality since static cache policies can be regarded as steady-state regimes of dynamic, general ones. We also provide simulation results to validate our theoretical results in homogeneous and heterogeneous cache size compared with dynamic cache replacement policies: random, LFU, and LRU, and observe that the delay performance of the static policies studied in this paper is similar to that of LFU and LRU in simulations (see Section 6). We believe that our theoretical findings are not limited for analytic purposes, and provide useful insights on how various factors have impact on caching performance.

2 RELATED WORK

Analyzing cache performance started from a single-scale case [8, 13, 14], where the main focus was on deriving asymptotic or approximate closed-form of cache hit probabilities for well-known cache policies such as LRU, LFU, and FIFO, often on the assumption of IRM (Independence Reference Model) (no temporal correlations in the stream of requests) for tractability. A network of cache, in spite of only for limited topology, has been studied for web caches. The work [5, 6] adopted a fluid model for analyzing a hierarchical caching (i.e., caching at clients, institutions, regions, etc.), and proposed a new, analysis-inspired caching scheme. The authors in [20] studied tradeoffs between hierarchical and distributed caching (i.e., caching at only institutional edges), and proposed a hybrid scheme that has lower latency. Recently, the work [10] mathematically explained the intuitions in [6].

The asymptotic analysis of cache networks has been studied only in wireless (multi-hop) networks, to the best of our knowledge. In [11], it was proved the required link capacity decreases from $O(\sqrt{n})$ to down to $O(1)$ using caches, where n is the number of nodes. This is due to the reduction of wireless interference induced by the decrease in necessary transmissions in presence of caching. In [2], a dynamic content change at caches was modeled by abstracting cache dynamics with limited lifetime of cached content. They showed that maximum throughput becomes $1/\sqrt{n}$ and $1/\log n$ for grid and random networks, compared to $1/n$ and $1/\sqrt{n \log n}$ for non-cache network.

3 MODEL AND PROBLEM STATEMENT

3.1 Model

Network and content servers. We consider a sequence of graphs $\mathcal{G}_n = (\mathcal{V}_n, \mathcal{E}_n)$, where $\mathcal{V}_n$ is the set of nodes or caches with $|\mathcal{V}_n| = n$ and $\mathcal{E}_n \subset \mathcal{V}_n \times \mathcal{V}_n$ describes neighboring relations between caches. In addition, we let $\mathcal{C}_n$ be the set of contents, and $\mathcal{S}_n$ be the set of servers containing the original contents. For notational simplicity, we will drop the subscript n for all quantities that depends on n, unless confusion arises. We assume that contents are of equal size and each content $c \in \mathcal{C}$ is stored in a single server, say s_c, and each server $s_c \in \mathcal{S}$ is attached to a randomly chosen node $v_c := v_{s_c} \in \mathcal{V}$. In this paper, we consider the case of one server per one content, but we remark that our results can be easily extended to multiple-server cases.

Content requests and routing. We assume that there are exogenous requests for contents at each cache, and locations of servers and content requesting nodes are uniformly at random in $\mathcal{V}$. In this environment, we assume that network capacity is large enough to ignore the negligible values such as waiting time at caches, and only consider the service time, denoted as the delay (see its formation definition in Section 3.2). Additionally, we assume each request is independent with others and the request rate of content $c_i \in \mathcal{C} = \{c_1, c_2, \ldots\}$ is proportional to a Zipf-like distribution with parameter $\alpha > 0$:

$$p_i = \frac{K}{i^\alpha}, \tag{1}$$

where the normalizing constant K is such that $\frac{1}{K} = \sum_{i=1}^{|\mathcal{C}|} 1/i^\alpha$. For a higher value α, we sometimes say that a cache network is with *higher popularity bias*, i.e., content popularity difference is high in that network. When a request for content $c \in \mathcal{C}$ arrives at cache v, it is forwarded along the path given by some routing algorithm, e.g., the shortest path routing in $\mathcal{G}$, from the cache v to the server v_c. The request generates HIT at cache v if c is located at v on the routing path or MISS otherwise. In the MISS event, the request keeps being forwarded to a next cache in the routing path until reaching to the server v_c.

Caches and policies. Each cache $v \in \mathcal{V}$ stores a set of contents $\mathcal{C}_v \subset \mathcal{C}$ independently, and has the cache size $b_v := |\mathcal{C}_v| \geq 0$ with the network-wide cache budget $B = \sum_{v \in \mathcal{V}} b_v$. The primary goal of this paper is to choose appropriate $[b_v]_{v \in \mathcal{V}}$ and $[\mathcal{C}_v]_{v \in \mathcal{V}}$ for the high performance of the cache networked system. Clearly, how many and which contents can be stored in each cache is governed by *content placements* and *cache sizing policies*. For content placements, the following rules are studied in this paper:

- ***URP (Uniformly Random Policy).*** Each cache contains b_v contents which are chosen from $\mathcal{C}$ uniformly at random.
- ***PPP (Pure Popularity-based Policy).*** Each cache contains b_v contents following the 'pure' (or 'exact') content popularity distribution, i.e., for a popularity parameter α.
- ***TPP (Tilted Popularity-based Policy).*** We also study more generalized popularity-based policies: each contains b_v contents following the Zipf-like distribution with parameter β (which may not be equal to α). Since we found that the choice $\beta = \alpha/2$ is optimal in some sense over many scenarios, we only focus on such a choice (see Section 4.2 for details).

- ***TPP-C (TPP with Cutting).*** This policy is a variant of TPP under some threshold for un-cached popularity ranking. Specifically, for given average routing distance $\bar{d}$ is given, we could compute the content index $\hat{i} = \min\{s \cdot \bar{d}, |C|\}$ (cf, s is the per-node cache size) and only the contents in $\{c_1, c_2, \ldots, c_{\hat{i}}\} \subset C$ are randomly cached based on Zipf-like distribution with parameter $\beta = \alpha/2$. This policy provides the best delay scaling laws among static ones studied in this paper, and we observe through simulations (see Section 6) that its delay is quite similar to those of dynamic ones such as LFU and LRU.

  For cache sizing policies, we separately study the following:

  - ***Homogeneous per-node cache size.*** All caches have the same size such that $b_v = \frac{B}{n}$ for all $v \in \mathcal{V}$ (see Section 4).
  - ***Heterogeneous per-node cache size.*** We also consider a setting concentrating cache budgets on more influential nodes such as having high 'betweenness centrality' that quantifies the fraction of shortest paths that pass through a node, i.e., cache budgets $[b_v]$ are different among nodes (see Section 5).

3.2 Performance Metric

In this section, we introduce the performance metric of our interest for cache networks. We define the delay of a content request as the number of (expected) hops until it finds the desired content, i.e., HIT occurs. Formally, let random variable X_i be the delay of the i-th request for some content in the (entire) cache network. Then, the asymptotic average delay Δ of the cache network is defined as follows:

$$\Delta \triangleq \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] \tag{2}$$

where it is not hard to check that the limit always exists given system setups $\mathcal{G}, [C_v], \alpha$, and a routing policy.

For a fixed $d > 0$, let $\Delta(d)$ be the "expected" (or average) delay when the routing distance between a content requesting node and the server is d, where the expectation is taken with respect to the randomness in the requested content, content placement policy and cache sizing policy. More formally, for a given distance d,

$$\Delta(d) = \sum_{i=1}^{|C|} p_i \xi_i(d), \tag{3}$$

where $\xi_i(d)$ denotes the expected delay of contents c_i for a given distance d under a (fixed) routing policy. However, note that d is a also random variable, when a randomly chosen content requesting node is assumed. Thus, the actual average delay Δ is given by, from (3),

$$\Delta = \mathbb{E}[\Delta(d))] = \sum_{d} f_d \Delta(d) = \sum_{d} f_d \sum_{i=1}^{|C|} p_i \xi_i(d), \tag{4}$$

where the expectation is taken over the distribution of random variable d and f_d is its probability which relies on the underlying topology of $\mathcal{G}$ and a given routing policy. Our objective is to study the asymptotic order of Δ under various setups, by studying $\Delta(d)$, which are analyzed in Sections 4 and 5. This study will asymptotically quantify the fundamental performance gains generated by the network of caches, which is expected to give practical implications into how we should design a cache network.

4 HOMOGENEOUS PER-NODE CACHE SIZE

4.1 Approach and LBND policy

In this section, we first focus on the case when each node has equal cache budget $s = s_v = B/n$ for all caches v. Content placement policies considered in our paper are mostly all identical random ones and do not differentiate particular caches. Hence, for a given routing path with distance d, the average delay $\xi_i(d)$ for content i depends only on the distance d and the cache hit probability h_{c_i} of content i at any arbitrary node, which is simply given by:

$$\xi_i(d) = \sum_{l=1}^{d-1} l \cdot h_{c_i} \cdot (1 - h_{c_i})^{l-1} + d \cdot (1 - h_{c_i})^{d-1}. \tag{5}$$

In addition to four content placement policies introduced in Section 3, we also consider an unrealistic ideal policy, which we call **LBND** (Lower BouND), that provides delay lower bounds on $\Delta(d)$, i.e., any policy cannot beat it. In LBND, for any routing path between a content requesting node and the server, contents are assumed to be placed on caches with descending order of popularity from the most popular contents such as $\{c_1, \ldots, c_s\}, \{c_{s+1}, \ldots, c_{2s}\}, \ldots$. Clearly, this is unrealistic because such a popularity-based descending ordering for *any* requesting node and server is impossible. Note that in LBND, $\xi_i(d)$, the average delay of content c_i for given distance d, is $\min\{\lceil i/s \rceil, d\}$ where $\lceil x \rceil$ indicates the minimum integer satisfying $\lceil x \rceil \geq x$.

4.2 Main Result

Theorem 4.1. *For a given routing distance d and the average routing distance $\bar{d}$ between an arbitrary pair of content requester and content server, the average delay $\Delta(d)$ scales as those in Table 1 under homogeneous per-node cache size.*

The proofs of URP and LBND in Theorem 4.1 are in Section 4.4, and remained proofs are in our technical report [15].

Here, we first provide interpretations of Theorem 4.1. For ease of explanation, we assume that $s = s_n = O(1)$, which is the most interesting case, because our natural interest lies in whether there is a delay reduction via a small amount of cache budget. For a constant per-node cache size, the results in Table 1 can be conveniently explained by diving the regimes into (i) $C \ll d$ and (ii) $C \gg d$ (in the asymptotic sense).

(a) As expected, the caching gains of popularity-based policies such as PPP, TPP, TPP-C increase as content popularity bias parameter α grows. For $\alpha > 2$ (very high popularity bias), TPP and TPP-C are order-optimal.

(b) In case of $|C| \ll d$, TPP and TPP-C outperforms PPP, and TPP/TPP-C is very close to even LBND. This is because when $|C| \ll d$, there is a large number of caching places from the requester to the corresponding server, so that TPP/TPP-C are efficient to reduce delay because of relatively high probability of un-popular contents rather than PPP.

(c) However, in case of $|C| \gg d$, the opposite occurs, i.e., due to lack of caches in the routing path, to reduce delay, more popular

Table 1: Homogeneous cache size: Delay of five static content placement policies. TPP-C-AVG corresponds to TPP-C with the average routing distance (and thus delay upper-bound from Jensen's inequality)

	URP	PPP	LBND	TPP and TPP-C ($s \cdot \bar{d} \geq |C|$)	TPP-C ($s \cdot \bar{d} < |C|$)	TPP-C-AVG
$2 < \alpha$	$\Theta\Big(\min[d, \frac{|C|}{s}]\Big)$	$O\Big(\min[\frac{(ds)^{1/\alpha}}{s}, \frac{|C|}{s}]\Big)$	$\Theta(1)$	$\Theta(1)$	$O\Big(\frac{d}{(s\bar{d})^{\alpha-1}}\Big)$	$\Theta(1)$
$\alpha = 2$	$\Theta\Big(\min[d, \frac{|C|}{s}]\Big)$	$O\Big(\min[\sqrt{\frac{d}{s}}, \frac{|C|}{s}]\Big)$	$\Theta\Big(\frac{\log(\min[s\cdot d, |C|])}{s}\Big)$	$O\Big(\min[d, \frac{\log^2(|C|)}{s}, \frac{\log(|C|)\log(s\cdot d)}{s}]\Big)$	$O\Big(\frac{1}{s}\max[\log^2 \bar{d}, \frac{d}{\bar{d}}]\Big)$	$O\Big(\frac{\log^2(\min[s\cdot\bar{d}, |C|])}{s}\Big)$
$1 < \alpha < 2$	$\Theta\Big(\min[d, \frac{|C|}{s}]\Big)$	$O\Big(\min[\frac{(ds)^{1/\alpha}}{s}, \frac{|C|}{s}]\Big)$	$\Theta\Big(\frac{(\min[s\cdot d, |C|])^{2-\alpha}}{s}\Big)$	$O\Big(\min[d, \frac{|C|^{2-\alpha}}{s}, \frac{|C|^{(2-\alpha)}\frac{\alpha-1}{\alpha} d^{\frac{2}{\alpha}-1}}{s^{2-2/\alpha}}]\Big)$	$O\Big((s\bar{d})^{1-\alpha}\max[\bar{d}, d]\Big)$	$O\Big(\frac{(\min[s\cdot\bar{d}, |C|])^{2-\alpha}}{s}\Big)$
$\alpha = 1$	$\Theta\Big(\min[d, \frac{|C|}{s}]\Big)$	$\Theta\Big(\min[d, \frac{|C|}{s}]\Big)$	$\Theta\Big(\min[d, \frac{|C|}{s\cdot\log|C|}]\Big)$	$O\Big(\min[d, \frac{|C|}{s\cdot\log|C|}]\Big)$	$O\Big(\max[\frac{\bar{d}}{\log|C|}, d]\Big)$	$O\Big(\min[\bar{d}, \frac{|C|}{s\cdot\log|C|}]\Big)$
$0 < \alpha < 1$	$\Theta\Big(\min[d, \frac{|C|}{s}]\Big)$	$\Theta\Big(\min[d, \frac{|C|}{s}]\Big)$	$\Theta\Big(\min[d, \frac{|C|}{s}]\Big)$	$O\Big(\min[d, \frac{|C|}{s}]\Big)$	$O\Big(\max[\bar{d}\cdot(\frac{s\cdot\bar{d}}{|C|})^{1-\alpha}, d]\Big)$	$O\Big(\min\{\bar{d}, \frac{|C|}{s}]\Big)$

contents should be cached with high probability. Thus, PPP outperforms TPP.

(d) TPP-C can be regarded as an *adaptive* policy that works well for both cases, because it tends to cache more kinds of caches when $|C| \ll d$, and focus on more popular contents when $|C| \gg d$, by adaptively determining the contents that should not be cached.

(e) As presented in (4), our analytical result $\Delta(d)$ in Theorem 4.1 can be plugged into the equation $\Delta = \sum_d f_d \Delta(d)$ to obtain the final average delay, once the distribution of routing distance f_d is known. However, in case when only *average* routing distance is available, our result is of great use, because from Jensen's inequality and concavity of Δ,

$$\Delta = \mathbb{E}[\Delta(d)] \leq \Delta(\mathbb{E}[d]), \tag{6}$$

and by replacing d in Table 1 by the average routing distance $\bar{d} = \mathbb{E}[d]$, at least delay upper-bounds can be computed. In fact, we present this for TPP-C, named TPP-C-AVG in Table 2, which shows delay performance being very close to LBND.

Why square root in tilting (TPP and TPP-C)? As mentioned earlier, TPP is the policy that provides more chances for less popular contents to be cached than PPP, and TPP-C is based on TPP with cutting the contents with "very low" popularity. Just for simplicity of exposition, assume $s = 1$, i.e., each node can cache only one content, and also assume that the routing distance d is extremely large, just like the regime $d \gg C$. Now consider a cache placement policy under which content c_i is cached in each cache with probability q_i. Note that a special case when $q_i = p_i$ corresponds to PPP (because PPP directly applies the content popularity distribution to the cache placement distribution) Then, the expected delay $\Delta(d)$ becomes:

$$\Delta(d) = \sum_{i=1}^{|C|} p_i \cdot \frac{1}{q_i} = \Big(\sum_{i=1}^{|C|} p_i \frac{1}{q_i}\Big)\Big(\sum_{i=1}^{|C|} q_i\Big) \geq \Big(\sum_{i=1}^{|C|} {p_i}^{\frac{1}{2}}\Big)^2,$$

where the last inequality comes from the Cauchy-Schwarz inequality. In Cauchy-Schwarz inequality, it is widely known that the equality holds if and only if there is some constant k such that $p_i \frac{1}{q_i} = k \cdot q_i$ for all i. Therefore, $\Delta(d)$ is minimized when $q_i \propto i^{-\frac{\alpha}{2}}$, and the minimum value is $\Big(\sum_{i=1}^{|C|} {p_i}^{\frac{1}{2}}\Big)^2$. This is why $\alpha/2$ is selected for TPP/TPP-C. Note that a special case when $qi = pi$ corresponds to the case utilizing content popularity distribution directly for the cache placement, and $\Delta(d) = |C|$.

4.3 Application to Power-law and Erdös–Rényi graphs

As case studies, we now apply Theorem 4.1 to popular random graphs: Power-law (PL) and Erdös–Rényi (ER) graphs, where we assume a shortest-path based request routing algorithm, $s = \Theta(1)$ and $|C| = \Theta(n)$. In the PL graph, the fraction of nodes with degree

Table 2: Orders of delay with URP, PPP, TPP, TPP-C, and LBND policies for average distance $\bar{d}$ in case study

	URP	PPP	TPP	TPP-C	LBND
$2 < \alpha$	$O(\bar{d})$	$O((\bar{d})^{1/\alpha})$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$\alpha = 2$	$O(\bar{d})$	$O(\sqrt{\bar{d}})$	$O(\bar{d})$	$O(\log^2(\bar{d}))$	$O(\log\bar{d})$
$1 < \alpha < 2$	$O(\bar{d})$	$O((\bar{d})^{1/\alpha})$	$O(\bar{d})$	$O((\bar{d})^{2-\alpha})$	$O((\bar{d})^{2-\alpha})$
$0 < \alpha \leq 1$	$O(\bar{d})$	$O(\bar{d})$	$O(\bar{d})$	$O(\bar{d})$	$O(\bar{d})$

i is proportional to $1/i^\gamma$ for some constant $\gamma > 0$. If the average degree is strictly greater than 1, and $2 < \gamma < 3$, it is known that the average routing distance under the shortest path routing is $\bar{d} = \Theta(\log n/\log\log n)$ [7]. The ER-graph is constructed by randomly connecting two nodes with some probability, say p. If np is of order $\log n$, then the graph almost surely contains a giant component of size of order n connected with high probability, and it is known in [9] that the average routing distance under the shortest path routing is $\bar{d} = \Theta\Big(\frac{\log n}{\log np}\Big)$. Using those facts about the average routing distances under two example random graphs and applying $\bar{d}$ for upper-bounds from Jensen's inequality (as in TPP-C-AVG), we obtain the delay orders for various content placement policies, shown in Table 2. TPP-C policy outperforms other policies except the case $0 < \alpha < 1$, and for $\alpha > 2$, TPP and TPP-C have $\Theta(1)$.

4.4 Proof of Theorem 4.1

Proof for URP. Since each cache has s spaces for $|C|$ contents, $\binom{|C|}{s}$ possible configurations are located over all caches uniformly at random, and $\binom{|C|-1}{s-1}$ configurations include a content c_i. Thus hit probability of content c_i becomes $\frac{\binom{|C|-1}{s-1}}{\binom{|C|}{s}} = \frac{s}{|C|}$, and

$$\Delta(d) = \xi_i(d) = \frac{1-(1-h_{c_i})^d}{h_{c_i}} = \Theta\Big(\frac{|C|}{s}\Big(1-e^{-\frac{ds}{|C|}}\Big)\Big).$$

Thus, $\Delta(d) = \Theta(d)$ for $d \le \frac{|C|}{s}$, and $\Theta(\frac{|C|}{s})$ for $d \ge \frac{|C|}{s}$. This completes the proof.

Proof for LBND. Under LBND, $\xi_i(d) = \min\{\lceil i/s \rceil, d\}$, and

$$\Delta(d) = \sum_{i=1}^{s} p_i \cdot \min\{d,1\} + \sum_{i=s+1}^{2s} p_i \cdot \min\{d,2\} + \cdots + \sum_{i=(\lceil \frac{|C|}{s} \rceil -1)s+1}^{|C|} p_i \cdot \min\left\{d, \left\lceil \frac{|C|}{s} \right\rceil\right\} \le K\left(1 + \int_0^{d-1}\int_{y\cdots+1}^{|C|} \frac{1}{x^\alpha} dx dy\right), \quad (7)$$

where $K = K(\alpha)$ is such that:

$$\frac{1}{K} = \sum_{i=1}^{|C|} \frac{1}{i^\alpha} = \begin{cases} \Theta(1) & \text{if} \quad \alpha > 1, \\ \Theta(\log |C|) & \text{if} \quad \alpha = 1, \\ \Theta(|C|^{1-\alpha/2}) & \text{if} \quad 1 > \alpha > 0. \end{cases}$$

Using K above and a simple calculation of the integral in the last term of (7) for various values α, the result follows.

5 HETEROGENEOUS PER-NODE CACHE SIZE

5.1 Motivation, Challenges and Model

The study in Section 4 enables us to purely focus on the impact of content popularity based caching on delay under the assumption of equal per-node cache size. However, it may be possible to gain more benefits by caching more contents at the caches that has more geometric importance. Examples include the policy that assign more cache budgets at the nodes with, e.g., high degrees or high access in request routing. This section is devoted to quantifying such an impact of heterogeneous cache sizing on delay.

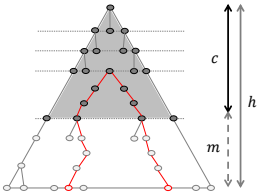


Figure 1: Regular spanning tree topology and BoW (Black or White) cache sizing policy, where the red line represents the shortest path between two nodes located in the bottom of the tree.

Regular tree and BoW (Black or White) sizing policy. In this paper, we consider a cache network whose topology is a $(r+1)$-*regular spanning tree,* and *shortest-path* based routing, as illustrated in Fig. 1. The tree has total n nodes and h layers, and each node has $r+1$ neighbors. This enables us to cover a large class of popular topologies, ranging from a line network to a star network, by simply changing r (e.g., a line for $r = 1$). We note that such tree topologies have popularly been used in P2P streaming systems [17, 21] and several content routing proposals in ICN [3]. We comment that we assume 'perfect regularity' because of simplicity in analysis, and our work can be readily extended to non-regular spanning trees.

There may be a large number of candidate cache sizing polices, out of which we consider a very simple policy, called *BoW (Black or White)*, which partitions the entire nodes into black (cacheable) and white (non-cacheable) ones. In other words, the system-wide cache budget is divided only among black nodes, and especially in BoW, the nodes only up-to the c-th layer become black, as seen in Fig. 1, where c should be carefully chosen to achieve low delay. Again, although not optimal, this simple policy provides a lower bound on the gain from heterogeneous cache sizing.

5.2 Main Results

Theorem 5.1. *The average delay Δ scales as those in following table for BoW cache sizing policy and shortest-path request routing under the $(r+1)$-regular spanning tree topology when $B = \Theta(n)$, where $X = \min[\log_r n, |C|]$. $\Leftarrow$ means that the corresponding value is same as its left one. Similar meaning for $\Uparrow$.*

	Homogeneous size		Heterogeneous size	
	TPP-C	*LBND*	*TPP-C*	*LBND*
$2 < \alpha$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$
$\alpha = 2$	$O(\log_r^2(X))$	$O(\log_r(X))$	$O(\log_r \log_r(X))$	$\Leftarrow$
$1 < \alpha < 2$	$O(X^{2-\alpha})$	$\Leftarrow$	$O(\log_r(X))$	$\Leftarrow$
$\alpha = 1$	$O(\min[\log_r n, \frac{\lvert C\rvert}{\log_r \lvert C\rvert}])$	$\Leftarrow$	$O(\log_r(\min[n, \lvert C\rvert]))$	$\Leftarrow$
$0 < \alpha < 1$	$O(\min[\log_r n, \lvert C\rvert])$	$\Leftarrow$	$\Uparrow$	$\Leftarrow$

Due to space limitation, we omit the proof, which is in our technical report [15]. Here, we summarize the key proof techniques and the interpretations of Theorem 5.1.

(a) In this topology, $h = \Theta(\log_r n)$ and for any fixed $c \ge 1$, the per-node cache size b_v for each black node v (i.e., nodes up-to the c-th layer) is $\Theta(\frac{B}{r^c})$. Let $m \triangleq h - c$. The key lies in how to choose c for small delay, as explained in what follows: First, note that the delay is bounded by: $\Delta \le 2m + 2\Delta_{\text{black}}(m)$, where Δ_{black} denotes the expected delay experienced in the "black region" whose bound can be computed by Theorem 4.1. The best $m^\star$ should be chosen satisfying $m^\star = \Theta(\Delta_{\text{black}})(m^\star)$.

(b) TPP-C and LBND in heterogeneous cache sizing achieve same order and approximately log-order delay reduction over those in homogeneous cache sizing.

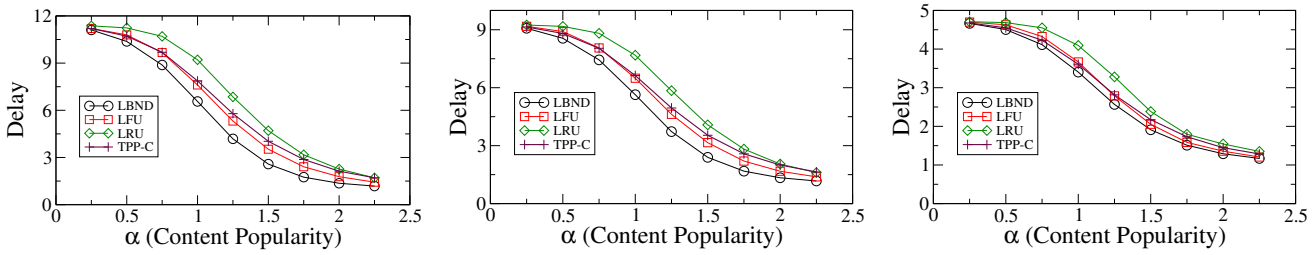
(c) Caching gains increase with the degree r of tree except the case $\alpha \le 1$ (i.e., low popularity bias) and $|C| > \log_r n$, where $\log_r n$ is the (worst-case) routing distance order in our tree topology.

(d) Recall that $m^\star = \Theta(\Delta_{\text{black}})(m^\star)$ in *(a)*. From the results that the delay order decreases as α increases, we can conclude that all cache nodes becomes useful, as the distribution of the content popularity is skewed more.

6 SIMULATION RESULTS

In this section, we present simulation results for three AS topologies: Cogent (USA-Europe), Colt Telecom (Europe), and TW Telecom (USA) from [1], as seen in Fig. 3. We conducted an off-line processing to understand three topologies, presented in Table 3, whose features are sufficiently heterogeneous.

We construct the simulation environment such that $s \cdot \bar{d} \ll |C|$, based on the recent trend of explosive increase in the number of contents. The main purpose of our simulations lies in figuring out how well the stationary policies approximate the dynamic policies in practice, where we plot only TPP-C for a stationary policy. To get simulation results, we perform 10 times of random instances



(a) Cogent (b) Colt (c) TW

Figure 2: Delay performance for three AS topologies

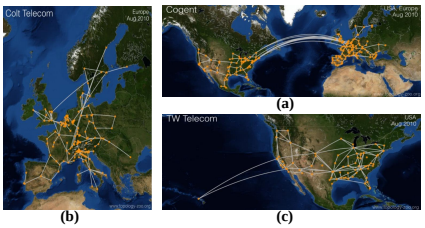


Figure 3: AS topologies of (a) Cogent (Europe-USA), (b) Colt Telecom (Europe), and (c) TW Telecom (USA) [1]

during 100000 slots. For each test, we first place content servers uniformly at random, and a content request arrives at a cache with probability 0.5 at the beginning of the time slot, and unresolved requests are forwarded to the next cache under the shortest path routing.

Table 3: Topology features and simulation environments

Topology	Cogent	Colt Tel.	TW Tel.
N	197	153	76
Avg. degree, $\bar{d}$	2.49, 10.4	2.50, 8.24	3.08, 3.21
$\|C\|$, s	3000, 5		
$\hat{i} = \min[s \cdot \bar{d}, \|C\|]$	50	40	15

Figure 2 shows the absolute (average) delay performances of two dynamic cache replacement policies, LRU and LFU, compared to the ideal policy and a static cache policy TPP-C, for three graph topologies. As done in the analysis earlier, the delay in y-axis corresponds to the number of hops. We observe that TPP-C's delay, which is a static cache policy smartly considering contents' popularity, has good match in those of LFU and LRU (on average, about 5.2% and 9.6% differences for LFU and LRU, respectively). Note that LFU is known to perform better than LRU with higher implementation complexity. From our simulation results, our analysis considering static cache policies can be practically used to predict the large-scale cache networks' delay performance.

7 CONCLUSION

In this paper, we performed asymptotic analysis of the delay performance of large-scale cache networks. We focused on quantitatively understanding the relation between content popularity and delay performance as well as the impact of heterogeneity in terms of "node importance". We studied the asymptotic delay performance of cache networks under homogeneous and heterogeneous per-node cache budget where caching gain incurred by heterogeneous cache sizing based on nodes' geometric importances increases as optimal.

ACKNOWLEDGEMENTS

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (No.B0717-17-0034, Versatile Network System Architecture for Multi-dimensional Diversity)

REFERENCES

[1] The Internet topology zoo. http://www.topology-zoo.org/dataset.html.
[2] B. Azimdoost, C. Westphal, and H. R. Sadjadpour. Fundamental limits on throughput capacity in information-centric networks. *IEEE Transactions on Communications*, 64(12):5037–5049, 2016.
[3] M. Bari, S. Rahman Chowdhury, R. Ahmed, R. Boutaba, and B. Mathieu. A survey of naming and routing in information-centric networks. *IEEE Communications Magazine*, 50(12):44–53, 2012.
[4] S. Borst, V. Gupta, and A. Walid. Distributed caching algorithms for content distribution networks. In *Proc. IEEE Infocom*, 2010.
[5] H. Che, Y. Tung, and Z. Wang. Hierarchical Web caching systems: modeling, design and experimental results. *IEEE Journal on Selected Areas in Communications*, 20(7):1305–1314, 2002.
[6] H. Che, Z. Wang, and Y. Tung. Analysis and design of hierarchical Web caching systems. In *Proc. Infocom*, 2001.
[7] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proc. the National Academy of Sciences*, 99(25):15879–15882, 2002.
[8] A. Dan and D. Towsley. An approximate analysis of the LRU and FIFO buffer replacement schemes. *Performance Evaluation Review*, 18(1):143–152, 1990.
[9] M. Draief and L. Massouli. *Epidemics and rumours in complex networks*. Cambridge University Press, 2010.
[10] C. Fricker, P. Robert, and J. Roberts. A versatile and accurate approximation for LRU cache performance. In *Proc. ITC*, 2012.
[11] S. Gitzenis, G. S. Paschos, and L. Tassiulas. Asymptotic laws for content replication and delivery in wireless networks. In *Proc. IEEE Infocom*, 2012.
[12] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard. Networking named content. In *Proc. ACM CoNext*, 2009.
[13] P. Jelenković. Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities. *The Annals of Applied Probability*, 9(2):430–464, 1999.
[14] P. R. Jelenkovic, X. Kang, and A. Radovanovic. Near optimality of the discrete persistent access caching algorithm. In *Proc. International Conference on Analysis of Algorithms*, 2005.
[15] B. Jin, S. Yun, D. Kim, J. Shin, Y. Yi, S. Hong, and B.-J. B. Lee. On the delay scaling laws of cache networks. *arXiv preprint arXiv:1310.0572*, 2013.
[16] K. Lee, H. Zhang, Z. Shao, M. Chen, A. Parekh, and K. Ramchandran. An optimized distributed video-on-demand streaming system: Theory and design. In *Proc. Allerton*, 2012.
[17] S. Liu, M. Chen, S. Sengupta, M. Chiang, J. Li, and P. A. Chou. P2P streaming capacity under node degree bound. In *Proc. IEEE ICDCS*, 2010.
[18] N. Niebert, S. Baucke, I. El-Khayat, M. Johnsson, B. Ohlman, H. Abramowicz, K. Wuenstel, H. Woesner, J. Quittek, and L. Correia. The way 4WARD to the creation of a future Internet. In *Proc. IEEE PIMRC*, 2008.
[19] I. Psaras, R. G. Clegg, R. Landa, W. K. Chai, and G. Pavlou. Modelling and evaluation of CCN-caching trees. In *Proc. NETWORKING*, pages 78–91. Springer, 2011.
[20] P. Rodriguez, C. Spanner, and E. W. Biersack. Analysis of Web caching architectures: hierarchical and distributed caching. *IEEE Transactions on Networking*, 9(4):404–418, 2001.
[21] T. Xu, J. Chen, W. Li, S. Lu, Y. Guo, and M. Hamdi. Supporting VCR-like operations in derivative tree-based P2P streaming systems. In *Proc. ICC*, 2009.