Proceedings Paper:

Zaidi, SAR, Ghogho, M orcid.org/0000-0002-0055-7867 and McLernon, DC (2015) Information Centric Modeling for Two-tier Cache Enabled Cellular Networks. In: 2015 IEEE International Conference on Communication Workshop, ICCW 2015. 2015 IEEE International Conference on Communication Workshop, ICCW 2015, 08 Jul - 12 Jun 2015, London, UK. IEEE , pp. 80-86. ISBN 9781467363051

https://doi.org/10.1109/ICCW.2015.7247159

Information Centric Modeling for Two-tier Cache Enabled Cellular Networks

Syed Ali Raza Zaidi, *Member, IEEE,* Mounir Ghogho, *Senior Member,IEEE,* Desmond C. McLernon, *Member, IEEE*

***Abstract*—In this article, we introduce a new metric called 'information centric coverage probability' to characterize the performance of a two-tier cache enabled cellular network. The proposed metric unifies the dynamics of in-network caching and heterogeneous networking to provide a unified performance measure. Specifically, it quantifies the probability that a mobile user (MU) is covered at a desired rate when a certain content is requested from a global content library. In other words, it quantifies the percentage of time when an MU can be served locally without paying the traffic penalties at backhaul, fronthaul and core networks. Caching dynamics are modeled by considering that the content which is least recently used (LRU) is evicted while the requested content is stored in the cache. The considered two-tier cellular model leverages coordination between the macro base-station (MBS) and the small cell base-stations (SBSs) to maximize the resource efficiency. More specifically, coordination between macro and small cells enables an arbitrary SBS to exploit the caches at other SBSs in the neighborhood. Thus reducing the requirement for huge and expensive memory modules at individual SBSs. The spatial dynamics of cellular network are modeled by borrowing well established tools from stochastic geometry. Propagation uncertainties are explicitly factored in characterization by considering the small scale Rayleigh fading and the large scale power-law path-loss model. It is shown that the information centric coverage probability is a function of (i) the size of caches at the SBSs and the MBS; (ii) the content eviction strategy; (iii) the underlying popularity law for referenced objects; (iv) the size of the global content library; (v) desired downlink transmission rate; (vi) the amount of spectrum allocated to each tier; (vii) path-loss exponent; and (viii) the deployment density of the SBSs and the MBSs. Our analysis reveals that significant performance gains can be harnessed with appropriate dimensioning of both cache sizes and deployment density. Finally, identification of memory limited vs. QoS limited operational regime for two-tier cellular networks is considered.**

***Index Terms*—Cache, small cells, LRU, two-tier, Poisson process, coverage.**

I. Introduction

A. Motivation

In recent times, the demand for ubiquitous wireless connectivity has intensified. The exponential growth in capacity requirements can be attributed to the increasing popularity of multimedia infotainment applications and the enormous penetration of smart platforms facilitating their execution. According to recent statistics [1], by the end of year 2019 mobile broadband subscriptions are expected to reach 7.6 billion, accounting for 80% of all mobile subscriptions, compared to around just 30% in 2013. The demand for mobile data traffic is expected to grow at a compound annual growth rate (CAGR) of 45% between 2013-2019. Consequently, it is predicted that while the voice traffic will maintain its current trend, the data traffic will grow 10 fold by the end of 2019.

These formidable capacity demands have lead to so called '$1000\times$ mobile data challenge' introduced by Qualcomm. More specifically, the $1000\times$ challenges dictates that fifth generation (5G) wireless networks (which are expected to roll out by early 2020) should be designed to be 1000 times more efficient than existing networks. In order to enable such a high level of efficiency, the architecture has to leverage three vital building blocks: (i) network densification; (ii) spectral agility; and (iii) higher energy efficiency. While network densification has received significant attention in the recent past, the prime objective has been to explore the design space from the quality-of-service (QoS) perspective (see [2] and references therein). An alternative yet powerful design perspective is to employ the content centric approach towards network densification. In other words, treating networks as intelligent reconfigurable platforms where proactive decision making can be employed to opportunistically enhance the quality-of-experience (QoE) with better spectral utilization and a smaller CO$_2$ footprint. It has been predicted by Cisco that the aggregate traffic generated by all forms of video traffic, i.e., TV, video on demand (VoD), Internet, and P2P will be in the range of $80-90\%$ of global consumer traffic by 2018. From the content/information centric perspective, intelligently designed networks can reduce this demand from such a high share of traffic by opportunistically utilizing network resources in conjunction with proactive in-network content caching.

The idea of caching in the IP networks has evolved from the underlying principles earlier exploited to empower computers with higher performance. In this context the main memory access was a key bottleneck and introduction of an intermediate on board faster memory (such as L1/L2 cache) yielded a several fold gain in computing performance. Consequently, cache memory has been central ingredient of computer architecture for several decades. The design principles of caching have evolved to a whole new level with the emerging 'information centric networking' paradigm. In particular, the information centric design of a network departs from the traditional end-to-end host centric architecture. The information centric design is inspired by the fact that the named information/content is the key consumable commodity from the end-user's perspective. Making data independent from the geographical location, application, storage and means of transportation through in-network caching and cognitive replication is envisaged to bring several fold gains in terms of spectral efficiency and scalability of networks. The information centric networking paradigm has also enabled content providers (such as Akamai, Level3 Communications and Limelight) to enhance QoE by reducing access delays through moving desired content closer to the user, i.e., towards the edge of the network.

B. Related Work

Investigation of cache empowered wireless networks is much more recent than the traditional computer networking paradigm. There are two key obstacles as regards exploring the design space:

- Performance of the caching algorithms on its own has been quite intricate and solution generally rely on sophisticated time consuming simulations. Thus characterizing the performance of the cache enabled cellular networks considering spatio-temporal dynamics of the cellular network is a non-trivial task

S. A. R. Zaidi, M. Ghogho and Des. C. McLernon are with the School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, United Kingdom, e-mail: {s.a.zaidi,m.ghogho,d.c.mclernon}@leeds.ac.uk. M. Ghogho is also affiliated with International University of Rabat, Morocco.

when compared to the traditional computer network with static topology, user association, etc.

- Modeling the link level performance of the cellular network itself has been quite a challenge until very recently. In the recent past [3], it has been recognized that borrowing well established mathematical tools from stochastic geometry can circumvent certain modeling issues. In particular, modeling the locations of the BSs as a Poisson point process significantly enhances the tractability while providing a lower bound on the actual performance experienced in a practical large scale deployment [3].

In the light of recent advances, the authors in [4] studied the performance of the cache-enabled small cell networks. Optimal cooperation between caching empowered device-to-device (D2D) communication nodes is investigated in [5], [6]. The authors in [5], [6], assume idealistic propagation conditions, i.e, the impact of multipath fading is completely ignored and an ideal protocol interference model is adapted. A detailed survey of the literature on cache-enabled D2D communication is beyond the scope of this current article. However, interested readers may refer to [6]–[9] and references therein. In [10], authors treated small cell base-stations (SBS) as helpers and formulated the problem of which files to cache when each user has access to multiple helpers. Since the focus in [10] is to derive an optimal caching strategy, authors do not consider the dynamics of the connectivity and the network topology. Moreover, most of the existing studies also ignore the fact that the cellular macro base-station (MBS) can also control the dynamics of information storage/retrieval in cache enabled networks.

The closest study to our work is [4], where the authors investigated the design space of the cache enabled small cell networks by adapting the tractable approach for characterizing coverage probability from [3]. To the best of our knowledge, it is the only study which considers both the spatial and the propagation dynamics of the cache enabled cellular network. Nevertheless, in [4] only a single tier network is considered, i.e., the existence and participation of the MBSs in both service and storage processes is completely ignored. Furthermore, the dynamics of the cache content replacement algorithm are not incorporated into the analysis. So, in this article, we consider a two-tier cache enabled small cellular network and explore its dimensioning under a well known least recently used (LRU) content eviction policy (see Section II for details).

C. *Contributions & Organization*

The contribution of this article is two fold:

1) We first introduce a new metric for a two-tier cache-enabled cellular network which unifies several networking aspects into a single quantitative measure. The introduced metric is termed as an 'information centric coverage probability' (refer to Section IV) which quantifies the probability that in a two-tier network a certain requested content can be retrieved locally and can be successfully transmitted to the desired MU. In contrast to [4], we consider that both the MBS and the SBS can perform caching while operating under the LRU based content eviction policy (see Section II & III). Thus, at each request for a certain content if it is either not located at the serving SBS, the MBS can arbitrate by retrieving the content from other SBSs or from its own cache. MBS can also serve the MU if the desired QoS cannot be met by its serving SBS. Such dynamic arbitration capitalizes on the fact that memory in proximity of a serving SBS is also an exploitable resource. Optimal exploitation of such a resource may bring significant gains by reducing both (i) backhaul/core network utilization; and (ii) memory requirements at individual SBSs.

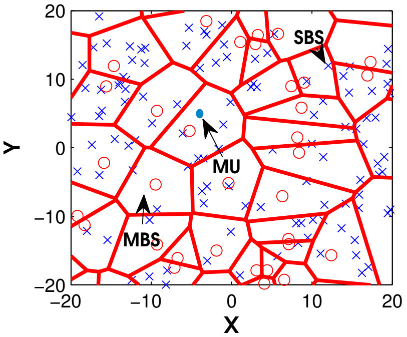


Figure 1. A realization of the proposed two-tier cellular network.

2) We characterize the information centric coverage probability by employing point process framework from stochastic geometry and also by using some recent results on the performance evaluation of the LRU cache hit rates. In particular, we revisit a well known "Che approximation" [11] for LRU cache hit probability, when the content is independently referenced and follows the Zipf like popularity law (details are deferred until Section III). We demonstrate how an alternative approximation based on the central limit theorem (CLT), and developed in [12], can be employed to quantify the cache hit probability. We then develop an analytical framework which combines the hit probability approximation with the spatial and channel dynamics of the cellular network under the proposed association model (see Section II) to characterize the information centric coverage probability. Finally, the impact of various parametric variations on the coverage probability are investigated in Section V. It is shown that employing the proposed metric provides interesting insights for the design of a large scale two-tier cellular network.

To the best of our knowledge, the system architecture considered in this article and the proposed metric have not been studied in any of the existing literature.

D. *Notations*

Throughout the paper, a particular realization of a random variable Z is denoted by a corresponding lower-case letter z and its probability density function (PDF) by $f_Z(.)$. The boldface lower-case letter (e.g. $\boldsymbol{x}$) is employed to denote a vector in $\mathbb{R}^2$. For sake of compactness, we employ $\boldsymbol{x}$ to refer to both the vector itself and also its location. The symbol $\setminus$ denotes the set subtraction and $\|\boldsymbol{x}\|$ denotes the Euclidean norm of the vector $\boldsymbol{x}$. The symbol $b(\boldsymbol{x}, r)$ denotes a ball of radius r centered at point $\boldsymbol{x}$. The symbol $\in$ denotes set membership and Π is generally employed to denote the point process. The point process is also used as a counting measure by using the notation $\Pi(\mathcal{A})$ which returns the number of points in Π which lie inside $\mathcal{A} \in \mathbb{R}^2$. Finally, $Z \sim \mathcal{E}(\mu)$ is used to represent an exponential random variable with mean μ.

II. System Model

A. *Spatial Model*

We consider a two-tier HetNet such that the first tier is formed by the macro cellular BSs. The subsequent tier is formed by SBSs such as femto cells. The spatial configuration of the BSs in both tiers is modeled by employing two independent homogeneous Poisson point processes (HPPPs), Π_m and Π_s, with intensity λ_m and λ_s respectively (see Fig. 1). Specifically, the probability of finding $n \in \mathbb{N}$ BSs inside a typical area foot-print $\mathcal{A} \subseteq \mathbb{R}^2$ follows the Poisson law with mean measure $\Lambda_i(\mathcal{A}) = \lambda_i v_2(\mathcal{A})$, $i \in \{s, m\}$. The mean measure is characterized by both the average number of BSs per unit area (i.e., λ_i) and the Lebesgue measure [13] $v_2(\mathcal{A}) = \int_{\mathcal{A}} d\boldsymbol{x}$ on $\mathbb{R}^2$, where if $\mathcal{A}$ is a disc of radius r then $v_2(\mathcal{A}) = \pi r^2$ is the area of the disc. Given $n \in \mathbb{N}$, the BSs are uniformly distributed in $\mathcal{A} \subseteq \mathbb{R}^2$. Notice that by virtue of superposition theorem for HPPPs (see [13]) the combined spatial distribution of BSs also forms a HPPP, i.e., $\Pi = \Pi_m \cup \Pi_s$.

B. Channel Model

The channel between an MU ($\boldsymbol{o}$), located at origin and a BS $\boldsymbol{x} \in \Pi_i$ $i \in \{s, m\}$ is modeled by the composite random variable $H_{\boldsymbol{o},\boldsymbol{x}} l(\|\boldsymbol{x}\|)$. Here $H_{\boldsymbol{o},\boldsymbol{x}} \sim \mathcal{E}(1)$ is a unit mean exponential random variable which captures the impact of the Rayleigh fading channel between a BS and a MU. The small-scale Rayleigh fading is complemented by a large-scale path-loss modeled by $l(\|\boldsymbol{x}\|) = G \|\boldsymbol{x}\|^{-\alpha}$. Here $\|\boldsymbol{x}\|$ is distance between the BS ($\boldsymbol{x}$) and the MU ($\boldsymbol{o}$); G is a frequency dependent constant and $\alpha \geq 2$ is an environment/terrain dependent path-loss exponent. The fading channel gains are assumed to be mutually independent and identically distributed (i.i.d.).

C. Content Access & Cellular Association Model

By virtue of the stationarity of the HPPP Π, it follows that the downlink performance of a typical MU can be treated as a proxy for the attainable performance of any MU (see [14]). Without any loss of generality, we consider a typical reference MU (denoted by $\boldsymbol{o}$) located at the origin. The MU requests a certain content n from a global library $\mathbb{S}$ of size N. It is assumed that each BS ($\boldsymbol{x} \in \Pi$) maintains a cache of size $C_{\boldsymbol{x}}$ [objects][1] where objects are stored upon their request. Whenever object $n \notin C_{\boldsymbol{x}}$, it is then retrieved and stored in the cache while the least recently accessed object is discarded. This eviction strategy is frequently known as an LRU strategy.

In this paper, we consider that upon the MU is associated with the nearest SBS and the nearest MBS. In subsequent discussion, the nearest SBS and MBS will be frequently referred to as the serving SBS and the serving MBS respectively. The serving cell of MBS $\boldsymbol{z} \in \Pi_m$ is defined as:

$$\mathcal{C}_{\boldsymbol{z}} = \left\{ \boldsymbol{x} \in \mathbb{R}^2 \middle| \, \|\boldsymbol{z} - \boldsymbol{x}\| \leq \|\boldsymbol{y} - \boldsymbol{x}\| \; \forall \boldsymbol{y} \in \Pi_m \text{ s.t.} \boldsymbol{y} \neq \boldsymbol{z} \right\}\}, \tag{1}$$

i.e., $\mathcal{C}_{\boldsymbol{z}}$ is the Voronoi cell of $\boldsymbol{z}$. The serving region of the SBS is defined in a similar manner. The key features of the downlink service for an MU are as follows:

- The MU request for the n^{th} item and the nearest SBS ($\boldsymbol{x} \in \Pi_s$) first searches its cache to locate the requested content. It is assumed that depending on the type of content, it is required to be transmitted to the MU at a certain desired rate $\mathcal{T}$ [bps/Hz]. If the content cannot be found at the serving SBS it redirects its request to the serving MBS. The serving MBS tries to locate the requested content locally at its own cache or at the caches furnished to the SBSs which are inside its serving region. For an LTE-A cellular network, the intra-tier coordination is attained over the X2 interface.
- If the requested content is located, it is served by either the SBS or the serving MBS depending on which one out of the two can satisfy the desired rate requirement.
- In the event that content cannot be locally located, a request is forwarded to serving gateway (S-GW) which retrieves the content from the source through the core network. Notice that such retrieval incurs a two-level cost which is essentially due to the backhaul transmission. The cost can be modeled in terms of the shared capacity, the energy penalty or the backhaul outage. Modeling such a cost, essentially empowers designers to explore the design space of HetNets in terms cache-ability of content. Due to space restrictions, we refrain from presenting the characterization of such costs.

D. Spectrum Allocation

In this paper, we assume that the small cells are deployed with the macro cellular network in an overlay mode. The choice is inspired by the recent proposals for the LTE Release 12 (expected in March 2015) where C/U plane split has been advocated by several industrial leaders. We assume that the MU is mainly associated on C-Plane (control plane) with the serving MBS while the U-plane (user plane) services are rendered by the small cells. The spectrum allocation across the macro and small cell tier is non-overlapping. In other words, small cells are deployed in non co-channel mode and the available spectral resources are split between both tiers. Consequently, if a unit bandwidth is available to operator, its fraction β is assigned to macro tier, while $(1-\beta)$ is allocated to small cell tier [2].

III. Cache Hit Probability under LRU Content Eviction Policy

In [11] Che et al. developed the fundamental design principles for the hierarchical web caching systems. The key contribution of [11] was to develop an approximation for estimating the cache hit probability under LRU content eviction policy. The "Che approximation" proved to be extremely accurate, even in the scenario's where an intuitive explanation for its applicability was quite vague. In a recent article [12], Fricker et al. revisited the Che approximation to unveil its remarkable success through an alternative mathematical explanation for a wide spectrum of scenario, beyond the specific conditions anticipated in [11]. In this article, we expand Fricker's alternative framework for the Che approximation for two-tier cellular networks. To this end, the key objectives of this section are:

1) To explicitly detail the traffic model, the content popularity laws and other assumptions which are central for quantifying the LRU cache hit probability by employing the Che approximation.
2) To recap the Che approximation and develop the Gaussian approximation for hit probability as in [12].

The developed approximations for the hit probabilities are later employed to quantify the proposed information centric coverage probability for an arbitrary MU in two tier cellular network with the limited shared backhaul.

A. The Independent Reference Model (IRM)

The independent reference model (IRM) has been frequently employed in literature (see [11], [12], [15] and references therein), to model how a particular object is referenced in an arbitrary traffic stream. Under the IRM model, requests for the objects/content occur in an infinite stream. The content/file demanded on the i^{th} request, for $i > 0$, is an independent random variable on $\mathbb{S} \subset \mathbb{Z}^{+}$with a common probability distribution. It is assumed that users request items from a global content library of size N, i.e., $|\mathbb{S}| = N$. The probability that the content $n \in \mathbb{S}$ is accessed is governed by the underlying content popularity distribution $p(n)$, where $p(n)$ is frequently referred to as the "popularity law" which is dependent upon the type of content, i.e., video, audio, etc.

[1] In this paper, we assume that each object has same size and the cache size is defined in terms of number of objects it can hold. Nevertheless, the analysis can be easily extended to the case where cache size is defined in bytes and each object has a different size (see [15] for details of LRU caches with variable sized objects).

[2] With the nearest neighbor association, it is straight-forward to show that equal allocation (i.e., $\beta = 0.5$) maximizes the performance metric considered in this paper.

Content Type	Popularity Exponent (η)
VoD	$.65 \le \eta \le 1.2$
Websites	$.64 \le \eta \le .83$
P2P Files	$.75 \le \eta \le .82$

Table I
SUMMARY OF CONTENT POPULARITY.

B. Content Popularity Law- A Case for Zipf Distribution

As discussed in the previous sub-section, the content popularity law governs how often a particular object is referenced in an infinite stream of the access requests. Empirical studies [15]–[17] have shown that the Zipf distribution renders itself as a most promising fit for describing the popularity of the content across various references. Under the Zipf popularity distribution $p(i) < p(j)$ for $i > j$. Moreover, the probability that the n^{th} content will be referenced is given by

$$p(n) = \frac{1/n^{\eta}}{\kappa_1}, \text{ where } \kappa_1 = \sum_{n=1}^{N} n^{-\eta}. \tag{2}$$

here κ_1 is normalizing constant, while $\eta > 0$ defines the exponent which characterizes the distribution. The value of η is strongly coupled with the type of content. Table 1, summarizes a few reference values for the various content types.

C. Che & Gaussian Approximations for Cache Hit Probability

Proposition 1 (Che Approximation)**.** *For a cache, say $\boldsymbol{x}$ of size $C_{\boldsymbol{x}}$, the probability that the object can be retrieved from a cache without accessing the content provider library for the requested n^{th} object is given as*

$$q_{\boldsymbol{x}}^{\{H\}}(n) = 1 - \exp\left(-p(n)t_{C_{\boldsymbol{x}}}\right), \tag{3}$$

where $t_{C_{\boldsymbol{x}}}$ is the characteristic time for the cache and is given by the unique root of

$$C_{\boldsymbol{x}} = \sum_{n=1}^{N} q_{\boldsymbol{x}}^{\{H\}}(n). \tag{4}$$

Proof: Please refer to [11]. ■

A detailed discussion on computation of characteristic time is beyond the scope of this article. The interested reader is directed to [11] for a comprehensive analysis. As can be seen from Proposition 1, analysis of the cache hit probability requires the solution of the non-linear equation (Eq. (4)). This can be quite involved for large cache and library sizes. This limitation can be circumvented by employing the Gaussian approximation for the hit probability developed by the Fricker et al. in [12].

Proposition 2 (Gaussian Approximation)**.** *The probability that the n^{th} requested item can be found in the cache of size $C_{\boldsymbol{x}}$ can be approximated by employing the central limit theorem as follows*

$$q_{\boldsymbol{x}}^{\{H\}}(n) = 1 - \frac{1}{2}\int_{0}^{\infty} \operatorname{erfc}\left(\frac{C_{\boldsymbol{x}} - \mu(t)}{\sqrt{2\sigma^2(t)}}\right) p(n)\exp\left(-p(n)t\right)dt. \tag{5}$$

where

$$\mu(t) = \sum_{n=1}^{N}\left(1 - \exp\left(-p(n)t\right)\right). \tag{6}$$

$$\sigma^2(t) = \sum_{n=1}^{N} \exp\left(-p(n)t\right)\left(1 - \exp\left(-p(n)t\right)\right). \tag{7}$$

Proof: For a detailed proof, readers may refer to [12]. ■

Intuitively, the effectiveness of the Gaussian approximation can be understood by observing that the cache hit probability is given by

$$q_{\boldsymbol{x}}^{\{H\}}(n) = \Pr\left\{T_{C_{\boldsymbol{x}}}(n) > \tau_n\right\} \tag{8}$$

where τ_n is the time since the last occurrence of the request for the object n. Notice that this time is exponentially distributed, i.e., the request arrivals are assumed to follow a Poisson point process with the rate proportional to the popularity of the content. Moreover, $T_{C_{\boldsymbol{x}}}(n) = \inf\left\{t > 0 : \sum_{i=1, i\neq n}^{N} \mathbb{1}_{\{\tau<\}} = C_{\boldsymbol{x}}\right\}$, i.e. the time at which cache is completely filled with the number of different objects excluding the requested object n. As $T_{C_{\boldsymbol{x}}}(n)$ is coupled with the sum of the exponential random variables, for a large N then the central limit theorem can be invoked and thus the Gaussian approximation for Eq. (8) can be easily established.

Note that the evaluation of hit probability by employing the Gaussian approximation requires the solution of the integral given Eq. (5). This integral cannot be evaluated in closed-form in general. However, a numerical solution can easily be obtained via standard analysis software such as MATLAB. The computation complexity is significantly reduced when compared to the original Che approximation.

IV. INFORMATION CENTRIC LOCAL COVERAGE PROBABILITY FOR TWO-TIER CELLULAR NETWORKS

The information centric coverage probability for an arbitrary MU ($\boldsymbol{o}$) is defined as the probability of finding the requested content locally from the caches inside the Voronoi cell of its serving MBS. Let $\boldsymbol{z} \in \Pi_m$ be the serving MBS of a typical MU $\boldsymbol{o}$ with its corresponding cell denoted by $\mathcal{C}_{\boldsymbol{z}}$. Moreover, let $n \in \mathbb{S}$ be the requested content in the current request generated by an MU, then the information centric downlink coverage probability is given by

$$\mathbb{P}_{cov}(n) = \mathbb{P}_{cov}^{\{S\}}(n) + \left(1 - \mathbb{P}_{cov}^{\{S\}}(n)\right)\mathbb{P}_{cov}^{\{M\}}(n), \tag{9}$$

where

$$\mathbb{P}_{cov}^{\{M\}}(n) = \Pr\left\{\beta\log_2(1+\Gamma_M) \ge \mathcal{T},\, n \in C_s\right\}, \tag{10}$$

$$\mathbb{P}_{cov}^{\{S\}}(n) = \Pr\left\{(1-\beta)\log_2(1+\Gamma_S) \ge \mathcal{T},\, n \in C_s\right\}, \tag{11}$$

here Γ_S and Γ_M are the received signal-to-interference ratios (SIRs) at an MU and the aggregate network memory in terms of caching is given by

$$\begin{aligned} C_s &= \overbrace{\sum_{\boldsymbol{y}\setminus\boldsymbol{x}\in\Pi_s} C_{\boldsymbol{y}}}^{\text{Caches at all SBSs inside Macro-cell}} + \overbrace{C_{\boldsymbol{x}}}^{\text{Cache at Nearest SBS}} \\ &+ \underbrace{C_{\boldsymbol{z}}}_{\text{Cache at Nearest MBS}} . \end{aligned} \tag{12}$$

Intuitively, an MU can be successfully served iff:

1) The nearest SBS which is serving the MU has a copy of the requested content in its cache and can serve MU at its desired rate;
2) Either the other SBSs or the MBS has a copy of the requested content in its cache and can transfer it to the serving SBS, which in turn can transmit content to the MU at its desired rate;
3) The content can be locally found at the SBSs managed by the serving MBS but the serving SBS cannot satisfy the desired rate requirement and thus the MU has to be served by the MBS directly.

The first two factors contribute towards the first term in Eq. (9), while the third case is captured by the second term. The information centric coverage probability unifies both the desired QoS and the cache dynamics into a single analytically tractable metric.

V. Performance Analysis of Two-tier Cache Enabled Cellular Networks

In the previous section, we briefly outlined the analytical framework which can be employed to study the performance of an LRU cache. In this section, we build on the already developed framework to unify the hit probabilities with the link level QoS to characterize the information centric coverage probability for two tier cellular networks.

Proposition 3 (Small Cell Coverage $\mathbb{P}_{cov}^{\{S\}}(n)$)**.** *The probability that an MU can be served by its serving SBS in the downlink for a certain locally accessible content $n \in \mathbb{S}$ request can be characterized as*

$$\mathbb{P}_{cov}^{\{S\}}(n) = \frac{\Xi(n)}{1 + \left(\delta\gamma_{ths}^{\delta}\mathcal{B}(\delta, 1-\delta) - {}_2\mathcal{F}_1\left(1, \delta; 1+\delta; -1/\gamma_{ths}\right)\right)}, \tag{13}$$

where $\gamma_{ths} = 2^{\mathcal{T}/(1-\beta)} - 1$ and

$$\Xi(n) = 1 - \frac{m^m\left(1 - q_{\boldsymbol{z}}^{\{H\}}(n)\right)}{\left(\lambda_S/\lambda_M q_{\boldsymbol{y}}^{\{H\}}(n) + m\right)^m} \text{ with } m = 3.5\,. \tag{14}$$

Proof: From Eq. (11), it follows that

$$\begin{aligned}
\mathbb{P}_{cov}^{\{S\}}(n) &= \Pr\left\{(1-\beta)\log_2(1+\Gamma_S) \geq \mathcal{T},\, n \in C_s\right\}, \\
&= \Pr\left\{\underbrace{\log_2(1+\Gamma_S)}_{\kappa_2} \geq \underbrace{\mathcal{T}/(1-\beta)}_{\mathcal{T}_2} \,\middle|\, n \in C_s\right\}\Pr\left\{n \in C_s\right\}, \\
&\underset{(a)}{=} \underbrace{\Pr\left\{\kappa_2 \geq \mathcal{T}_2\right\}}_{A_1}\underbrace{\Pr\left\{n \in C_s\right\}}_{\Xi(n)},
\end{aligned} \tag{15}$$

where (a) follows from the mutual independence of the QoS and the cache hit probabilities. Notice that the term A_1 corresponds to the event that the MU can be served at its desired rate by the nearest SBS and can be evaluated as

$$\begin{aligned}
A_1 &= \Pr\left\{\kappa_2 \geq \mathcal{T}_2\right\} = \Pr\left\{\Gamma_S \geq \underbrace{2^{\mathcal{T}_2} - 1}_{\gamma_{ths}}\right\}, \\
&= \Pr\left\{h_{\boldsymbol{ox}} \geq \frac{\gamma_{ths} I_r}{l(r)}\right\} = \mathbb{E}_R\left(\mathcal{L}_{I_r}(s)\big|_{s=\gamma_{ths}r^{\alpha}}\right).
\end{aligned} \tag{16}$$

The random variable $I_r = \sum_{\boldsymbol{x}\in\Pi_s\setminus b(\boldsymbol{o},r)} h_{\boldsymbol{x}} l(\|\boldsymbol{x}\|)$ models the other-cell interference experienced at a typical MU $\boldsymbol{o}$. Now, following the steps similar to [3] with several mathematical manipulations, the Laplace transform of the aggregate interference experienced by the MU can be quantified as

$$\begin{aligned}
\mathcal{L}_{I_r}(\gamma_{th}r^{\alpha}) &= \exp\left(-\lambda_s\pi\delta\gamma_{ths}^{\delta}r^2\left(\mathcal{B}(\delta, 1-\delta)\right.\right. \\
&- \left.\left.\left(\delta\gamma_{ths}^{\delta}\right)^{-1}{}_2\mathcal{F}_1\left(1, \delta; 1+\delta; -1/\gamma_{ths}\right)\right)\right)
\end{aligned} \tag{17}$$

where $\delta = \frac{2}{\alpha}$, $\mathcal{B}(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the standard Beta function and ${}_2\mathcal{F}_1(a, b; c; z) = \sum_{n=0}^{\infty}\frac{a^{(n)}b^{(n)}}{c^{(n)}}\frac{z^n}{n!}$ is the Gauss Hypergeometric function[3]. Finally employing the distance distribution for nearest SBS, Eq. (16) can be evaluated as

$$\begin{aligned}
A_1 &= \int_0^{\infty} 2\pi\lambda_s r\exp\left(-\lambda_s\pi r^2\zeta\right)dr, \\
&= \frac{1}{1 + \left(\delta\gamma_{ths}^{\delta}\mathcal{B}(\delta, 1-\delta) - {}_2\mathcal{F}_1\left(1, \delta; 1+\delta; -1/\gamma_{ths}\right)\right)}.
\end{aligned} \tag{18}$$

Now, $\Xi(n) = \Pr\{n \in C_s\}$ can be evaluated as follows

$$\begin{aligned}
\Xi(n) &= 1 - \Pr\left\{n \notin C_s\right\}, \\
&\underset{(b)}{=} 1 - \underbrace{\Pr\left\{n \notin C_{\boldsymbol{z}}\right\}}_{A_2}\underbrace{\prod_{\boldsymbol{y}\in\mathcal{C}_{\boldsymbol{Z}}\cap\Pi_s}\Pr\left\{n \notin C_{\boldsymbol{y}}\right\}}_{A_3},
\end{aligned} \tag{19}$$

where (b) can be deducted from Eq. (12). Moreover, the term A_2 can be computed as

$$A_2 = 1 - q_{\boldsymbol{z}}^{\{H\}}(n). \tag{20}$$

The term A_3 can be written as

$$\begin{aligned}
A_3 &\underset{(c)}{=} \mathbb{E}_A\left[\sum_{k=0}^{\infty}\left(1 - q_{\boldsymbol{y}}^{\{H\}}(n)\right)^k\frac{(\lambda_s a)^k}{k!}\right. \\
&\times \left.\exp\left(-\lambda_s a\right)\right], \\
&\underset{(d)}{=} \mathbb{E}_A\left[\exp\left(-\lambda_s q_{\boldsymbol{y}}^{\{H\}}(n)a\right)\right],
\end{aligned} \tag{21}$$

where (c) follows from the Poisson law for Π_s and (d) can be obtained by using the fact that $\exp(x) = \sum_{k=0}^{\infty} x^k/k!$. The Evaluation of Eq. (21), requires the distribution for the area of a typical Voronoi cell. From [17] the distribution of the normalized area is given by

$$f_X(x) = \frac{m^m}{\Gamma(m)}x^{m-1}\exp(-mx), \tag{22}$$

where $m = 3.5$ and X is a random variable that denotes the size of the typical Voronoi cell normalized by the value $1/\lambda_M$. Employing Eq. (22), A_3 can be simplified to

$$A_3 = \frac{m^m}{\left(\lambda_S/\lambda_M q_{\boldsymbol{y}}^{\{H\}}(n) + m\right)^m}. \tag{23}$$

Substituting A_2 and A_3in Eq. (19), we obtain

$$\Xi(n) = 1 - \frac{m^m\left(1 - q_{\boldsymbol{z}}^{\{H\}}(n)\right)}{\left(\lambda_S/\lambda_M q_{\boldsymbol{y}}^{\{H\}}(n) + m\right)^m}. \tag{24}$$

Remarks ■

- From Eq. (18) it is easy to see that the coverage probability in terms of the desired QoS is a function of: (i) the fraction of spectrum (β) allocated to the small cell tier; (ii) the required rate for content transfer ($\mathcal{T}$) and (iii) the path-loss exponent (α). It is independent from the density of small cells and their transmit powers. This can be credited to the fact that while increasing the density of small cells reduces average link distance by $\Theta(1/\sqrt{\lambda_s})$ the distance between nearest interferer also scales under the same law. Consequently, the gain in terms of signal power is offset with the increase in aggregate interference.
- Noting that the probability of being covered, given the fact that content can be locally retrieved is independent of the density and the transmit power, then it follows that

$$\mathbb{P}_{cov}^{\{M\}}(n) = \frac{\Xi(n)}{1 + \left(\delta\gamma_{thm}^{\delta}\mathcal{B}(\delta,1-\delta) - {}_2\mathcal{F}_1(1,\delta;1+\delta;-1/\gamma_{thm})\right)}. \tag{25}$$

where $\gamma_{ths} = 2^{\mathcal{T}/\beta} - 1$ and $\Xi(n)$ is defined in Eq. (13).

Proposition 4. *Consider an arbitrary MU which generates a request for n^{th} object from the global content library of size N. The probability that the MU is covered locally in two-tier small cell networks under LRU based content eviction policy can be quantified as*

$$\mathbb{P}_{cov}(n) = \Delta(1-\beta) + (1 - \Delta(1-\beta))\Delta(\beta), \tag{26}$$

[3] Here $a^{(n)} = a \times (a+1) \times \ldots.(a+n-1)$ is the rising Pochhammer Symbol.

$$\Delta(a) = \frac{1 - \frac{m^m\left(\left(\frac{1}{2}\int_0^\infty \mathrm{erfc}\left(\frac{C_{\boldsymbol{z}}-\mu(t)}{\sqrt{2\sigma^2(t)}}\right)p(n)\exp(-p(n)t)dt\right)\right)}{\left({}^{\lambda_S}/_{\lambda_M}\left(1-\frac{1}{2}\int_0^\infty \mathrm{erfc}\left(\frac{C_{\boldsymbol{y}}-\mu(t)}{\sqrt{2\sigma^2(t)}}\right)p(n)\exp(-p(n)t)dt\right)+m\right)^m}}{1+\left(\delta\left(2^{\mathcal{T}/a}-1\right)^{\delta}\,{}^{\pi}/_{\sin(\pi\delta)} - {}_2\mathcal{F}_1\left(1,\delta;1+\delta;-{}^{1}/_{\left(2^{\mathcal{T}/a}-1\right)}\right)\right)}. \quad (27)$$

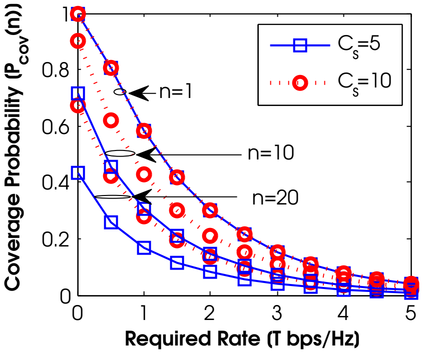


Figure 2. Information centric coverage probability ($\mathbb{P}_{cov}(n)$) as a function of varying the required transmission rate ($\mathcal{T}$) for various requested object indices (n) and small cell cache memory (C_s) with $C_m = 10$, $\alpha = 4$, $\lambda_m = 10^{-3}$, $\lambda_s = 10^{-2}$, $\beta = 0.5$, $N = 10^4$ and $\eta = 1.2$ (see Eq.(26)).

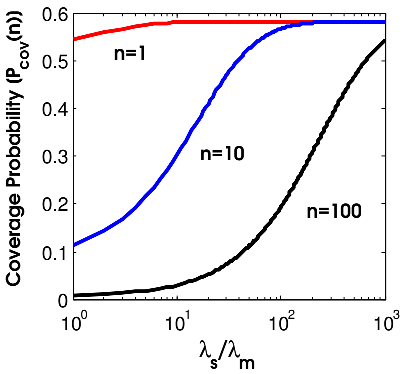


Figure 3. Information centric coverage probability ($\mathbb{P}_{cov}(n)$) as a function of varying the ratio of deployment between small cells and macro cells (${}^{\lambda_S}/_{\lambda_M}$) for various requested object indices (n) with $\mathcal{T} = 1$, $C_m = 10$, $C_s = 5$, $\alpha = 4$, $\beta = 0.5$, $N = 10^4$ and $\eta = 1.2$ (see Eq.(26)).

where $\Delta(a)$ is defined in Eq. (27).

Proof: Employing the definition of information centric probability of being locally covered from Eq. (9) in conjunction with Eqs. (13) and (25) completes the proof. ■

Remarks

- From Proposition 4, it can be easily observed that unlike the traditional coverage probability metric, the information centric coverage probability is strongly dependent upon the density of deployment for each tier. More specifically, the ratio between the density of SBSs and MBSs dictate the overall performance of the scheduled downlink MU.
- Another interesting observation from Proposition 4 is that the information centric coverage probability is dependent upon which content is requested (i.e. object index n) and also the cache sizes deployed across both tiers. Optimal dimensioning of the allocated memory at each tier can be employed to maximize the downlink performance. We defer the dimensioning problem for the future work.

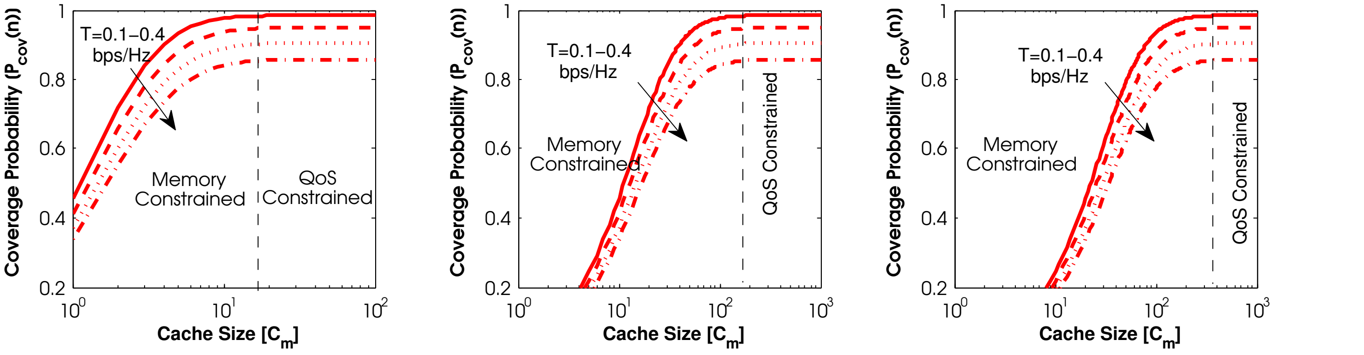
VI. Discussion & Results

Fig. 2 investigates the impact of the desired bit rate, the popularity of requested object and the size of SBS cache on information centric coverage probability[4]. As expected, the probability of being locally covered decreases with an increase in the desired rate. Intuitively, this can be attributed to the co-channel interference which limits the maximum attainable rate in the downlink transmission for any scheduled MU. For a fixed desired rate, the information centric coverage probability is strongly dependent on how popular the requested content is. The content which is less frequently demanded by the users has a lower probability of being in the cache and thus the probability of retrieving such a object from the serving macro or small cell BSs is also low. Thus the information centric coverage probability decreases with an increase in the requested object index n, when objects are sorted in accordance with their popularity (such as in case of the Zipf distribution). As shown in Fig. 2 the coverage probability for the objects which are less frequently requested can be significantly improved by increasing the size of the cache available at the SBSs. For instance, when $\mathcal{T} = 1$ bps/Hz is required for the downlink transmission, the coverage probability can be improved by 50% for $n = 10-20$. For the popular content, deploying more memory does not effect the coverage probability as transmissions are already constrained by the attainable QoS performance for the the downlink MU and not on the cache size. Note that the coverage probability increases with an increase in η (the exponent of Zipf distribution). Due to the space constraints, we refrain from presenting the straightforward results.

Fig. 3 depicts the impact of increasing the average number of small cells per macro cells (i.e., the ratio of deployment density ${}^{\lambda_S}/_{\lambda_M}$) on the information centric coverage probability. As is obvious from the figure, the coverage probability for the less frequently referenced items improves significantly through the network densification. Following, the widely observed $10-90$ law, i.e., 10% of the content is requested 90% of the time, it is easy to see that for a library of size $N = 10^4$, most referenced items are from $n = 1-100$. Moreover, as depicted by Fig. 3 the local coverage probability for $n = 100$ can be increased by a factor of 6× with ultra dense deployment. However, the coverage probability considering only attainable QoS, does not depend on the deployment density λ_s (see Eq. (18) and subsequent remarks). Thus instead of deploying more SBSs, simply increasing the cache memory at each SBS will have a similar impact.

Fig. 3 explores the behavior of the information centric coverage probability against the deployed cache memory ($C_s = C_m$) at both the small and macro BSs for various values of the desired rate ($\mathcal{T}$). An interesting observation from Fig. 3 is that for any referenced object, there exists two distinct operation regimes, i.e., (i) a memory constrained regime; and (ii) a QoS constrained regime. In a memory constrained regime, the information centric coverage probability can be significantly improved by increasing the size of the cache installed on the BSs. However, after a certain value of C_m, say C_m^*, the benefits of adding more memory saturate. In other words, the information centric coverage probability saturates to a certain value which is dictated by the path-loss exponent,

[4] Throughout this section, the size of cache at SBSs is denoted by (i.e., $C_{\boldsymbol{y}} = C_s$) while the size of cache at MBS is denoted by $C_z = C_m$.



(a) $n = 1$ (b) $n = 10$ (c) $n = 20$

Figure 4. Information centric coverage probability ($\mathbb{P}_{cov}\,(n)$) as a function of cache size ($C_s = C_m$) at small cells and macro cells (λ_S/λ_M) for various requested object indices (n) and desired rates ($\mathcal{T}$) with , $\alpha = 4$, $\beta = 0.5$, $\lambda_s/\lambda_m = 2$, $N = 10^4$ and $\eta = 1.2$ (see Eq.(26)).

the required downlink transmission rate and the allocated fraction of spectrum at each tier. Thus, the downlink performance of the MU becomes QoS constrained rather than memory constrained. Furthermore, it is observed that as expected C_m^* increases with an increase in n. Effectively, sporadically referenced items have an expanded memory-constrained operational regime as compared to the popular items. Thus the improvement by adding the extra cache memory at the BSs strongly depends on the maximum content index say n^*, for which the network information centric coverage needs to be maximized.

VII. Conclusion

In this article, we presented a comprehensive framework for the information centric modeling of a two-tier heterogeneous cellular network. We introduce a new metric called 'information centric local coverage probability' which quantifies the probability that a downlink MU can retrieve its desired content from the content caches deployed at the nearest macro base-station or from the overlaid small cells. We demonstrated that unlike a traditional coverage metric, which is completely characterized by the distribution of the signal-to-interference ratio (SIR), the information centric coverage probability is coupled with: (i) the size of caches; (ii) the content eviction strategy; (iii) the underlying popularity law for referenced objects; (iv) the size of global content library; (v) the desired downlink transmission rate; and (vi) the amount of spectrum allocated to each tier. Moreover, under the SIR based coverage model the probability of an MU being covered is independent of the density of the deployed base stations and only depends on desired rate threshold, allocated fraction of spectrum and the path-loss exponent. However, the information centric coverage probability is strongly coupled with the deployment density of base-stations. It is demonstrated that when a least recently used content is evicted from the cache, the network densification can bring gains of the order of 6× for a moderately accessed object. It is demonstrated that for a certain desired rate, the information centric coverage probability has two distinct operational regimes with respect to the cache size, i.e., (i) memory limited; and (ii) cache limited. It is shown that the performance can be maximized in a memory limited regime by increasing the cache capacity. However, there exists a threshold capacity (dependent upon the underlying popularity law and considered referenced object) after which additional memory does not bring any benefit in terms of performance. The operation beyond this threshold is strongly dependent on the required rate. In summary, it is shown that the information centric coverage probability is a strong function of the content type and thus content aware design can bring significant performance gains for the 5G cellular networks.

References

[1] A. Ericsson, "Traffic and market data report: On the pulse of the networked society," *White Paper*, 2014.

[2] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 996–1019, 2013.

[3] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, 2011.

[4] E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking, (In Press)*, 2014.

[5] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Wireless device-to-device communications with distributed caching," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*. IEEE, 2012, pp. 2781–2785.

[6] N. Golrezaei, A. Dimakis, and A. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4286–4298, July 2014.

[7] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *arXiv preprint arXiv:1305.5216*, 2013.

[8] X. Wang, M. Chen, T. Kwon, L. Jin, and V. Leung, "Mobile traffic offloading by exploiting social network services and leveraging opportunistic device-to-device sharing," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 28–36, June 2014.

[9] Y. Zhang, E. Pan, L. Song, W. Saad, Z. Dawy, and Z. Han, "Social network aware device-to-device communication in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 177–190, Jan 2015.

[10] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, 2013.

[11] 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, 2002.

[12] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for lru cache performance," in *Proceedings of the 24th International Teletraffic Congress*. International Teletraffic Congress, 2012, p. 8.

[13] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[14] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1029–1046, 2009.

[15] C. Fricker, P. Robert, J. Roberts, and N. Sbihi, "Impact of traffic mix on caching performance in a content-centric network," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2012, pp. 310–315.

[16] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," in *ACM SIGOPS Operating Systems Review*, vol. 40, no. 4. ACM, 2006, pp. 333–344.

[17] S. M. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," in *11th International Symposium on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt)*. IEEE, 2013, pp. 119–124.