Performance Evaluation of Information-Centric Networking for Multimedia Services

Geyong Min, Haozhe Wang, Jia Hu, Wang Miao,
College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK
Email: {g.min, hw389, j.hu, wm255}@exeter.ac.uk

***Abstract*—The rapid development in multimedia services has shifted the major function of the current Internet from host-centric communication to service-oriented content dissemination. Motivated by this significant change, Information-Centric Networking (ICN) has emerged as a new networking paradigm, which aims at providing natural support for efficient information retrieval over the Internet. As a crucial characteristic of ICN, in-network caching enables users to efficiently access popular content from ubiquitous caches to improve the Quality-of-Experience (QoE). Therefore, in-network caching for ICN has received considerable attention in recent years and many cache schemes and models have been proposed. However, there is a lack of research into ICN cache models under practical environments such as arbitrary topology and multimedia services exhibiting bursty nature. To bridge the gap, this paper proposes a new analytical model to gain valuable insight into the caching performance of ICN with arbitrary topology and bursty content requests. The accuracy of the proposed model is validated by comparing the analytical results with those obtained from simulation experiments. The analytical model is then used as a cost-efficient tool to investigate the impact of key network and content parameters on the performance of caching in ICN.**

I. Introduction

The current Internet architecture is designed based on the host-to-host model, which is aimed at data exchange and communication. However, with the emerging technologies such as Internet of Things, mobile cloud services and the dramatic growth of various multimedia services [1], there come significant change to the main focus of the Internet, shifting from host-focusing to service-oriented. In fact, the Internet is becoming a content distribution platform, which motivates the origin of Information-Centric Networking (ICN) [2]–[5]. The ICN paradigm treat content as the first class entity in the network architecture and decouple content from host to achieve a naturally service-oriented architecture [6], which is sensible because users do not care where the content is, but are more interested in how fast and reliable the requested content can be obtained [7].

In-network caching is considered as an integral part of ICN to efficiently obtain content, alleviate congestion, reduce network load, and enhance the Quality-of-Experience (QoE). However, in-network caching differs from Web caching in which cache is transparent to applications and content to be cached is finer grained. This poses new challenging issues to be addressed such as cache management and cache placement/replacement strategies. Furthermore, due to the variety of multimedia applications, traffic pattern becomes an important factor when conducting performance evaluation of caching in ICN. The streams generated by multimedia services, which nowadays is the dominant type of network traffic, have been observed to hold the bursty characteristic. Since the data transmission in ICN is a receiver-driven process, i.e., the communication is driven by receivers sending requests, the content request process of multimedia services also exhibits the bursty nature, which needs to be taken into account when evaluating the performance of ICN.

A unified and accurate analytical model of cache networks can be used as a cost-effective tool to analyze the behaviour of ICN caching and further guide the design and optimization of ICN. Among the existing ICN architectures, this paper focuses on the Content-Centric Networking (CCN) architecture [3] because it is seen as a promising global-scale ICN networking paradigm.

The existing analytical models on ICN caching are mainly focused on a single cache node or special cache network topologies, such as cascade topology or tree topology [8]–[11]. These special network topologies simplify the interoperability between cache nodes, and hence simplify the establishment and analysis of cache network models. But in ICN, the realistic topology of cache networks should be represented by arbitrary graphs [12] rather than hierarchical trees. [13] proposed an approximate model to investigate general cache networks under Poisson content request process. However, in today's service-oriented network, several types of multimedia traffic will compete for the same caching space, thus the simplified traffic models cannot be used to accurately quantify its performance measures. Most of the existing in-network cache studies consider the simplified traffic models such as fix arrival rate and Poisson process [8], [9], [14], [15], which fail to capture the bursty nature of content requests in ICN.

To fill in the gap, this paper develops a cost-effective analytical model to investigate ICN caching with arbitrary topology under bursty content requests. To capture the bursty nature of multimedia services, the developed model adopts the Markov-modulated Poisson process (MMPP) to characterize the content requests. The accuracy of the analytical model is validated through comparing the analytical results with those generated from simulation experiments. Moreover, the analytical model is used to explore the impact of the key network and content parameters in terms of cache size, content size and popularity distribution on the performance of ICN caching.

The remainder of the paper is organized as follows. Section II is devoted to the model description, which presents the system parameters and introduces how to capture the bursty content requests. In Section III, a new analytical model is developed. The accuracy of the model is validated in Section IV, and then the proposed model is used to carry out the performance analysis of ICN caching. Finally, Section V concludes the paper.

II. Model Description

This section presents the model description and system parameters of cache networks, followed by the representation of bursty content requests.

A. System parameters

The model description and system parameters of cache networks are introduced in this subsection. Tab. I provides a summary of the notations used in the derivation of the model, and the notations are explained in details below:

(i) The cache network is represented by ICNet $= (V, E)$, where $V = \{v_1, \ldots, v_n\}$ denoting the cache nodes in the network, $E \subseteq V \times V$ denoting the links between two nodes.

(ii) Each node v_n in an ICN caching network contains a cache with the size of C_{v_i} chunks. Chunk is the minimum caching unit. In ICN, contents are segmented into multiple smaller pieces, called chunks, and each chunk is treated as an individually named object, aiming to allow flexible distribution and flow control. The similar idea of segmentation has also been adopted in many other content distribution systems, such as BitTorrent and eMule.

(iii) The cache on each ICN node runs the LRU cache replacement policy. The LRU policy has low complexity and has been used in [8], [9], [16], [17]. Moreover, the caching operations of LRU can be implemented at line speed, which is one important requirement of ICN.

(iv) A total of $\mathbb{O} = \{content_1, \ldots, content_O\}$ different contents are considered in the model. Contents are formed into K sets, with each set denoted as one type of services, thus each set contains $m = O/K$ different contents. Within each service type, the popularity among the m contents is the same.

(v) The popularity of contents belonged to different services follows the Zipf distribution. Zipf distribution is widely used for characterising the content popularity in [8], [15], [17], [18], because it has been pointed out in [19] that the popularity real web content accesses has been observed following the Zipf distribution. As a result, contents in service k (i.e., the k-th most popualr content) are requseted with the probability $q_k = f(\alpha, k) = \frac{1/k^\alpha}{\sum_{i=1}^{K} 1/i^\alpha} = \frac{D}{k^\alpha}$, $k \geq 1$, where $\alpha \geq 1$ is the value of the exponent characterising the distribution, $1/D = \sum_{i=1}^{K} 1/i^\alpha$.

(vi) The size of content, $S(content_i)$ follows the geometrical distribution with an average of F chunks, i.e. $\mathbb{P}(S(content_i) = l) = \frac{1}{F}(1 - \frac{1}{F})^{l-1},\ i = 1, 2, \ldots, O$, and $l > 0$

TABLE I
System parameters investigated in this paper

Parameter	Meaning
V	The set of ICN caching nodes
E	Links between the nodes
$\mathbb{O}$	Total number of different content items
K	Number of different types of services
m	Number of different contents in each type of service
C_{v_i}	Cache size in number of chunks of node v_i
α	Zipf exponent characterizing the distribution
q_k	Probability of requests for contents of service k
F	Average content size in number of chunks
S	File size in number of chunks following geometrically distributed
h_{k,v_n}	Cache hit ratio for a chunk of contents in type k at node v_n
H_k	Mean cache hit ratio for contents in type k service
H	Global cache hit ratio
$\lambda_k(n)$	Mean arrival rate of requests for contains in type k service at node v_n
$\lambda_{k_{tot}}(n)$	Actual content requests rate for type k service at node v_n
$\lambda_{tot}(n)$	Total content request rate for all kinds of services at node v_n
Q_k	Infinitesimal generator of requests for contents in class k
Λ_k	Request rate matrix of contents in class k
N	Number of ICN nodes in the network

B. Bursty content requests

The MMPP is a doubly stochastic process with the arrival rate varying according to an irreducible continuous-time Markov chain [20]. It is capable of modeling the bursty content requests because it can capture the time-varying arrival rate. The arrival process of *Interests* is modeled by a special case of MMPP called Interrupted Poisson Process (IPP). IPP_k with subscript k is adopted to model the request for a content in class k, and is characterized by infinitesimal generator Q_k of the underlying Markov process and the rate matrix Λ_k. Q_k and Λ_k are given by

$$Q_k = \begin{bmatrix} -\sigma_{k1} & \sigma_{k1} \\ \sigma_{k2} & -\sigma_{k2} \end{bmatrix},\ \Lambda_k = \begin{bmatrix} \lambda_{sk} & 0 \\ 0 & 0 \end{bmatrix} \tag{1}$$

where σ_{k1} denotes the transition rate from state 1 to 2, and σ_{k2} is the transition rate from state 2 to 1. λ_{sk} is the request arrival rate when the Markov chain is in state 1. The mean arrival rate for contents in class k, λ_k, is given by

$$\lambda_k = \frac{\sigma_{k1} \times 0 + \sigma_{k2} \times \lambda_{sk}}{\sigma_{k1} + \sigma_{k2}} \tag{2}$$

The global content request is represented by the superposition of the K input IPPs, which is again an MMPP, as the MMPP is closed under the superposition operations. The generator Q and the rate matrix Λ of the composite MMPP are

calculated from the individual generators Q_k and rate matrices Λ_k as follows

$$Q = Q_1 \oplus Q_2 \oplus \cdots \oplus Q_K,$$
$$\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \cdots \oplus \Lambda_K. \quad (3)$$

where $\oplus$ denotes the Kronecker-sum. The composite Q and Λ of the superposed MMPP are $K^2 \times K^2$ matrices and can be written as

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_{12} & \cdots & \sigma_{1K^2} \\ \sigma_{21} & -\sigma_2 & \cdots & \sigma_{2K^2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K^21} & \sigma_{K^22} & \cdots & -\sigma_{K^2} \end{bmatrix}, \ \sigma_i = \sum_{\substack{j=1 \\ j \neq i}}^{K^2} \sigma_i j,$$

$$\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_{K^2}), \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_{K^2})^T \quad (4)$$

The mean arrival rate, λ_{tot}, of the composite MMPP can be derived from the steady-state vector $\boldsymbol{\pi}$ and the arrival rate vector $\boldsymbol{\lambda}$, as

$$\lambda_{tot} = \boldsymbol{\pi}\boldsymbol{\lambda}. \quad (5)$$

where $\boldsymbol{\pi}$ satisfies $\boldsymbol{\pi}Q = 0$, and $\boldsymbol{\pi}\boldsymbol{e} = 1$.

Contents are split into chunks that are uniquely identified by a name and are permanently stored in one or more sources. The arrival *Interests* with rate λ_k for contents in class k are generated according to (Q_k, Λ_k), and the requested content is equally chosen among the m different contents in the given class. A content request yields the request for the first chunk of that content. Once a chunk is received, a new chunk request is sent continuously until the reception of the last chunk of that content.

III. Modelling the Performance of In-network Caching for ICN

An analytical model is developed in this section to investigate the performance of caching network with arbitrary topology under the aforementioned bursty content requests. In the model, *cache hit ratio* is considered as the key performance metric to evaluate the performance of caching, since high cache hit ratio will result in efficiently access of content and better quality of user experience (QoE). Furthermore, it will reduce the traffic load in the network and achieve better energy efficiency.

Considering the miss process of the request for a chunk of $content_i$, due to the property of LRU policy, a miss is generated if more than C_{v_n} different chunks are requested during the interval between the current request and previous request for the same chunk on node v_n. Therefore, the cache hit ratio of the request for a chunk of a content in type k at node v_n is given by

$$h_{k,v_n} = 1 - \mathbb{P}(Req_{v_n}(\tau_{k,n}) \geq C_{v_n}) \quad (6)$$

where $Req(\tau_{k,n})$ denotes that the number of different content requests arrived at node v_n, during the inter-arrival time τ_k, between two subsequent requests of the same chunk in type k. The rest of this section will describe the detail calculation of cache hit ratio.

A. Calculation of the time interval

The inter-arrival time, $\tau_{k,n}$, denotes the interval in which there are more than C_{v_n} requests arrived between two subsequent requests of the same chunk in Type k. Since the number of arrival chunks is larger than the the cache size, the requested chunk has been removed from the cache, thus it generates the event of cache miss.

$\tau_{k,n}$ is determined by content request rate (Λ), the cache size (C_{v_n}), the whole content number (O), the number of types (K), the content size (F) and the Zipf distribution parameter (α), so $\tau_{k,n}$ can be expressed as

$$\tau_{k,n} = f(\Lambda, C_{v_n}, O, K, F, \alpha) \quad (7)$$

In our previous work [21], the mean arrival rate of chunk requests, g and the inter-arrival time τ_k for a single ICN node has been derived as

$$g = \Gamma(1 - \frac{1}{\alpha})^\alpha (\frac{\lambda_{tot} D}{2}) m^{\alpha-1} F^\alpha \quad (8)$$
$$\tau_k = C^\alpha / g \quad (9)$$

However, for a node in the caching network, the rate of content requests arrived at a node is the combination of two streams. The MMPP stream of content requests arrives exogenously, and in the event of miss, the node receives the forwarded requests from its neighbours. So the actual arrival rate at a node of caching network with arbitrary topology can be written as

$$\lambda_{k_{tot}}(n) = \lambda_k(n) + \sum_{\substack{v_{n'}: n \neq n' \\ E_{n,n'} \neq 0}} miss_k(n') \quad (10)$$

The miss rate at an ICN node not only depends on the cache policies, but also relates to the time for moving a copy of the content to the cache after a miss generated. The paper follow the common practice in [13], [22], [23] that the content is downloaded into the cache instantaneously after a miss occurs.

So, the chunk arrival rate g_n and inter-arrival time $\tau_{k,n}$ can be written as

$$g_n = \Gamma(1 - \frac{1}{\alpha})^\alpha (\frac{\lambda_{tot}(n) D}{2}) m^{\alpha-1} F^\alpha$$
$$\tau_{k,n} = C_{v_n}^\alpha / g_n \quad (11)$$

where $\lambda_{tot}(n)$ is the mean arrival rate of all the requests at node v_n, and can be computed by

$$\lambda_{tot}(n) = \sum_{k=1}^{K} \lambda_{k_{tot}}(n)$$
$$= \sum_{k=1}^{K} \lambda_k(n) + \sum_{k=1}^{K} \sum_{\substack{v_{n'}: n \neq n' \\ E_{n,n'} \neq 0}} miss_k(n') \quad (12)$$

The first part of Eq. (12) can be calculated by Eq. (5). To determine the second part, Eq. (6) has to be solved for all types k and nodes v_n.

B. Calculation of the miss rate

To determine the $miss_k(n)$ at at any node in the network, $\mathbb{P}(Req_{v_n}(\tau_{k,n}) \geq C_{v_n})$ needs to be derived first. Since MMPP is used to capture the bursty content requests, given the requests for contents in type k with intensity $\lambda_{k_{tot}}(n)$ and transition matrix Q_k, $\mathbb{P}(Req_{v_n}(\tau_{k,n}) \geq C_{v_n})$ is a Bernoulli sequence and can be written as [21]

$$\mathbb{P}(Req_{v_n}(\tau_{k,n}) \geq C_{v_n}) = \beta_k e^{-u_k \tau_{k,n}} + (1-\beta_k) e^{-v_k \tau_{k,n}} \tag{13}$$

where the parameters u_k and v_k are the two eigenvalues of $(\Lambda_k - Q_k)$, and can be written as

$$\begin{aligned} u_k &= \frac{\frac{\lambda_{k_{tot}}(n)}{m} + \sigma_{k1} + \sigma_{k2} - d_k}{2} \\ v_k &= \frac{\frac{\lambda_{k_{tot}}(n)}{m} + \sigma_{k1} + \sigma_{k2} + d_k}{2} \end{aligned} \tag{14}$$

with

$$d_k = \sqrt{\left(\frac{\lambda_{k_{tot}}(n)}{m} + \sigma_{k1} - \sigma_{k2}\right)^2 + 4\sigma_{k1}\sigma_{k2}}$$

β_k is the transformation parameter from the *IPP* to the hyperexponential distribution, and is given by [20]

$$\beta_k = \frac{\frac{\lambda_{k_{tot}}(n)}{m} - v_k}{u_k - v_k} \tag{15}$$

Then Eq. (6) can be written as

$$h_{k,n} = 1 - \beta_k e^{-u_k \tau_{k,n}} - (1-\beta_k) e^{-v_k \tau_{k,n}} \tag{16}$$

Through examining the above equations, several interdependencies between the different variables of the model can be found. For example, Eq. (16) show that the cache hit ratio, $h_{k,n}$ is a function of $\tau_{k,n}$, while Eq. (11) and Eq. (12) reveal that $\tau_{k,n}$ is a function of total miss streams of neighbour nodes, which in turn requires the calculation of $h_{k,n}$. To derive the closed-form solutions to such interdependencies are very difficult, the equations of the model are computed through the iterative techniques.

Firstly, based on our previous work, Eqs. (8) (9) (13) are applied to every single node for all types k, with only the exogenous MMPP content requests. Then, the results are used to calculate Eq. (12) and update Eq. (11). In each iteration, Eq. (6) is finally solved and the result is fed into the next iteration.

IV. Model Validation and Performance Analysis

The accuracy of the developed analytical model is validated via a discrete-event simulator [17], developed under the OMNeT++ framework. This open-source simulator implements the Content Store (CS), Pending Information Table (PIT) and Forwarding Information Base (FIB) data structures, and content retrieve operations of ICN.

The model developed for caching network with arbitrary topology under bursty content requests is validated via a two-dimensional 5×5 torus network, illustrated in Fig. 1.

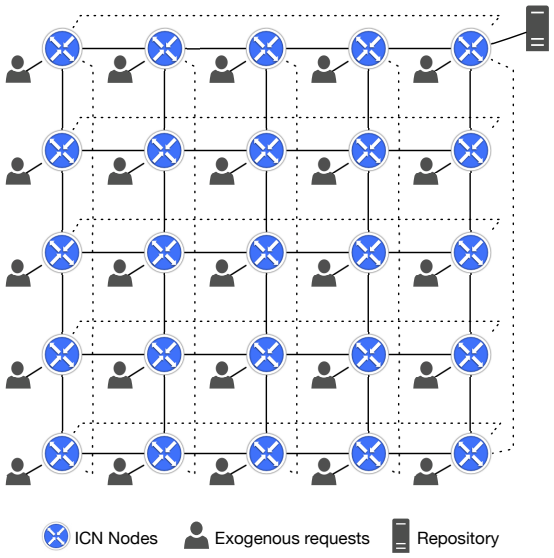

Fig. 1. Network topology: two-dimensional torus network

Each ICN node receives exogenous content requests from users. One repository which contains all the contents is placed randomly in the network on each simulation experiment. Torus topology has been widely used [13], [17], [24] to investigate the cache performance. Furthermore, by applying different routing methods, a torus topology can be formed into a cascade topology (deterministic routing) or a tree topology (shortest path routing). To achieve a more general topology, the random forwarding method is used in the validation, so in the event of a miss at a node, each neighbour node has 25 percent of chance to receive the forwarded missing request. In such situation, Eq. (10) can be written as

$$\lambda_{k_{tot}}(n) = \lambda_k(n) + \frac{1}{4}\sum_{n'=1}^{4} miss_k(n') \tag{17}$$

The parameters set for the validation are presented in Tab. II. A total of $M = 500$ different contents are considered and allocated into $K = 10$ sets with decreasing content popularity as the set number increases. The popularity of each set of contents follows the Zipf distribution with the exponent parameter $\alpha = 2$. The Zipf exponent $\alpha = 2$ is derived from the analysis of YouTube for a realistic Internet catalog size [19]. Each type of service owns $m = 50$ contents which are split into chunks of $10KB$ size, and the content size is geometrically distributed with average 10^3 chunks ($10MB$). The size of chunk of $10KB$ and average size of content of 10^3 chunks are widely used in the literature. In the torus network, each node receives the exogenous content requests generated by end-users accessing multimedia services. The exogenous requests are modelled by MMPP with a mean arrival intensity 10 $contents/s$, and the chunk transmission window size is set to $W = 1$ [3]. The standard Leave Copy Everywhere (LCE) decision policy [3] and Least Recently Used (LRU)

TABLE II
PARAMETERS SET FOR THE VALIDATION

Parameter	Values
N	25
M	500
m	50
K	10
chunk	10KB
C_{v_n}	1GB, 1.2GB, 1.5GB
F	10MB
α	2

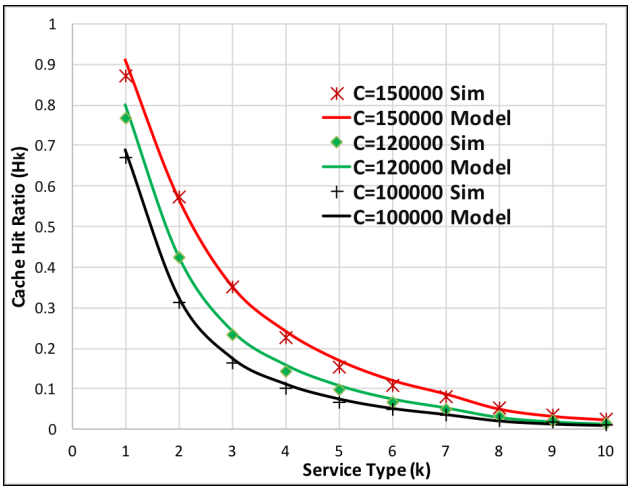

Fig. 2. Mean cache hit ratio H_k predicted by the model against those obtained through simulation under bursty traffic vs. content of different service types with different cache sizes C_{v_n}.

replacement policy [25] are implemented on each cache node with equal size C_{v_n}.

In the topology shown in Fig. 1, nodes are equivalent due to the utilization of random forwarding routing. In order to investigate the caching network performance, the mean cache hit ratio of contents in type k, H_k, is defined. H_k can be derived from the superposition of the cache hit ratio for all types of services at each node, and is given by

$$H_k = \sum_{v_n \in V} \omega_n h_{k,v_n} \tag{18}$$

where ω_n is the weight factor of the node and depends on the traffic load through it, with $\sum \omega_n = 1$. Fig. 2 depicts the mean cache hit ratio, H_k, as a function of the cache size for different cache sizes C, with 100000 chunks ($1GB$), 120000 chunks ($1.2GB$) and 150000 chunks ($1.5GB$) respectively. The figure reveals that the analytical performance results match well with those obtained from the simulation, which validates the accuracy of the developed analytical model. We can also see that the increase of cache size C causes the growth of cache hit ratio as expected.

Next, the developed model is used to investigate the impact of key network and content parameters on the caching performance. According to Eqs. (11) and (16), cache hit ratio h_{k,v_n} is a function of these parameters: cache size C_{v_n}, average

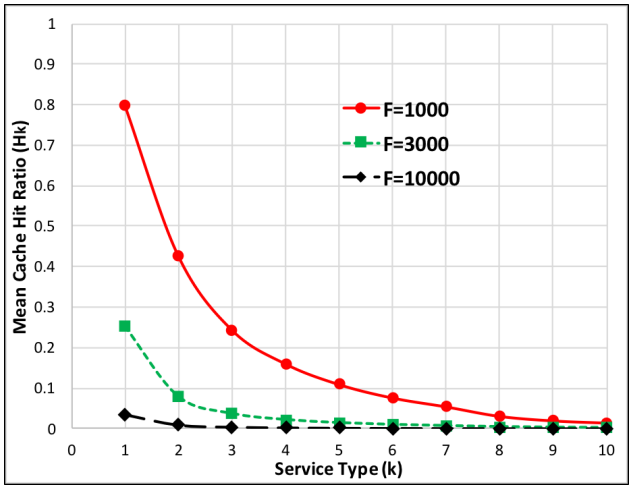

Fig. 3. Cache hit ratio predicted by the model for different content sizes with cache size $Cv_n = 120000$.

content size F, total number of contents O and Zipf exponent α. Unless otherwise stated, the values of the parameters in the performance evaluation is the same as those in Tab. II.

Content size F plays an important role in caching performance, as shown in Fig. 3, where the mean cache hit ratio, H_k, is clearly affected by the content size, F. As F increases, the cache hit rate decreases accordingly. As illustrated in Fig. 3, when the content size becomes very large ($F = 0.1GB$), the cache hit ratio for all kinds of contents become very low, indicating that larger content size will cause low cache hit ratio.

To investigate the impact of Zipf exponent α on the cache performance, the global cache hit ratio, H, is considered. H can be derived from the superposition of the hit sequences of different contents in all classes, and is given by

$$H = \sum_k q_k H_k \tag{19}$$

As depicted in Figs. 4 and 5, the global cache hit ratio, H, is a monotone increasing function of the Zipf exponent, α. This is because that the smaller α leads to a flatter popularity, which means that the popularity of each class is close to the others and the contents in each class are requested by a similar probability. In this case, contents in the cache are replaced more frequently than that with larger α, which hence pulls down the cache hit ratio.

The figures demonstrate that the developed analytical model can be used to predict the cache hit ratio of ICN nodes in the presence of arbitrary topology and bursty request process.

V. CONCLUSIONS

In this paper, a new analytical model has been developed to investigate caching performance of ICN with arbitrary topology and bursty content requests. The cache hit ratio at each node is derived as the key performance index. MMPP is adopted to capture the bursty nature of content requests from multimedia services. Simulation experiments have been

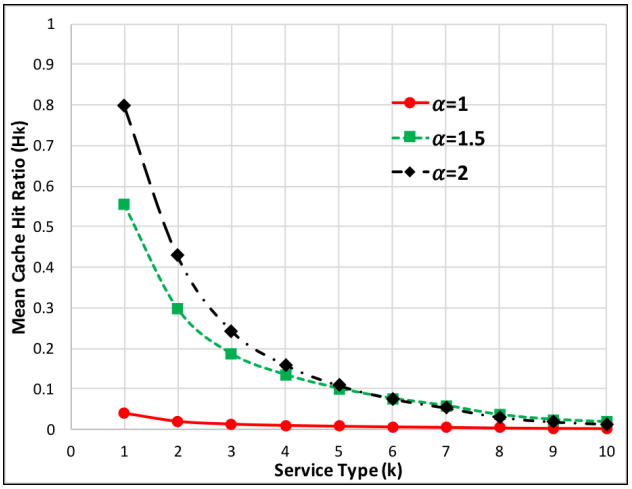

Fig. 4. Cache hit ratio predicted by the model for different Zipf exponent α vs. content of different classes with cache size $Cv_n = 120000$.

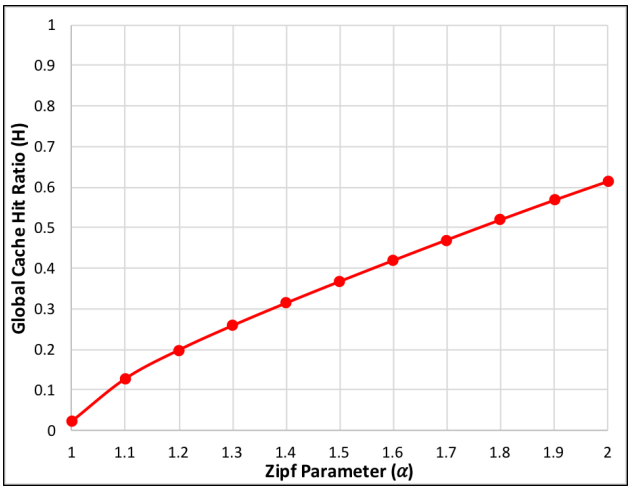

Fig. 5. Global cache hit ratio predicted by the model vs. different Zipf exponent α with cache size $C = 120000$.

performed to validate the effectiveness and accuracy of the analytical model, which has been used as a cost-effective tool to gain the insights of the impact of key network and content parameters on the cache performance in ICN.

REFERENCES

[1] Cisco, "Cisco visual networking index: Forecast and methodology, 2014–2019," *Cisco White paper*, 2015.

[2] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K. H. Kim, S. Shenker, and I. Stoica, "A data-oriented (and beyond) network architecture," in *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 4. ACM, 2007, pp. 181–192.

[3] 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, pp. 1–12.

[4] C. Dannewitz, D. Kutscher, B. Ohlman, S. Farrell, B. Ahlgren, and H. Karl, "Network of information (netinf) - an information-centric networking architecture," *Computer Communications*, vol. 36, no. 7, pp. 721–735, Apr. 2013.

[5] N. Fotiou, P. Nikander, D. Trossen, and G. Polyzos, "Developing information networking further: From psirp to pursuit," *Broadband Communications, Networks, and Systems*, vol. 66, pp. 1–13, 2012.

[6] Q. Wu, Z. Li, J. Zhou, H. Jiang, Z. Hu, Y. Liu, and G. Xie, "Sofia: toward service-oriented information centric networking," *IEEE Network*, vol. 28, no. 3, pp. 12–18, 2014.

[7] G. Xylomenos, C. Ververidis, V. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. Katsaros, and G. Polyzos, "A survey of information-centric networking research," *IEEE Communications Surveys Tutorials*, vol. 16, no. 2, pp. 1024–1049, Second 2014.

[8] G. Carofiglio, M. Gallo, L. Muscariello, and D. Perino, "Modeling data transfer in content-centric networking," in *Proc. International Teletraffic Congress*, 2011, pp. 111–118.

[9] I. Psaras, R. G. Clegg, R. Landa, W. K. Chai, and G. Pavlou, "Modelling and evaluation of ccn-caching trees," in *NETWORKING 2011*. Springer, 2011, pp. 78–91.

[10] L. Muscariello, G. Carofiglio, and M. Gallo, "Bandwidth and storage sharing performance in information centric networking," in *Proceedings of the ACM SIGCOMM workshop on Information-centric networking*. ACM, 2011, pp. 26–31.

[11] I. Psaras, W. K. Chai, and G. Pavlou, "Probabilistic in-network caching for information-centric networks," in *Proceedings of the Second Edition of the ICN Workshop on Information-centric Networking*, ser. ICN '12. New York, NY, USA: ACM, 2012, pp. 55–60.

[12] G. Zhang, Y. Li, and T. Lin, "Caching in information centric networking: A survey," *Computer Networks*, vol. 57, no. 16, pp. 3128 – 3141, 2013, information Centric Networking.

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

[14] S. Arianfar, P. Nikander, and J. Ott, "Packet-level caching for information-centric networking," in *Proc. ACM SIGCOMM, ReArch Workshop*, 2010.

[15] W. Chai, D. He, I. Psaras, and G. Pavlou, in *NETWORKING*. Springer, 2012, vol. 7289, pp. 27–40.

[16] K. Katsaros, G. Xylomenos, and G. C. Polyzos, "Multicache: An overlay architecture for information-centric networking," *Computer Networks*, vol. 55, no. 4, pp. 936–947, 2011.

[17] G. Rossini and D. Rossi, "A dive into the caching performance of content centric networking," in *Proc. IEEE Computer Aided Modeling and Design of Communication Links and Networks*, 2012, pp. 105–109.

[18] P. R. Jelenković and A. Radovanović, "Least-recently-used caching with dependent requests," *Theoretical computer science*, vol. 326, no. 1, pp. 293–327, 2004.

[19] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM on Internet measurement*, 2007, pp. 1–14.

[20] W. Fischer and K. Meier-Hellstern, "The markov-modulated poisson process (mmpp) cookbook," *Performance Evaluation*, vol. 18, no. 2, pp. 149–171, 1992.

[21] H. Wang, G. Min, J. Hu, H. Yin, and W. Miao, "Caching of content-centric networking under bursty content requests," in *Proc. IEEE WCNC*, 2014, pp. 2522–2527.

[22] A. Dan and D. Towsley, "An approximate analysis of the lru and fifo buffer replacement schemes," in *Proc. ACM SIGMETRICS*, 1990, pp. 143–152.

[23] H. Che, Z. Wang, and Y. Tung, "Analysis and design of hierarchical web caching systems," in *Proc. IEEE INFOCOM*, vol. 3, 2001, pp. 1416–1424.

[24] V. Sourlas, L. Gkatzikis, P. Flegkas, and L. Tassiulas, "Distributed cache management in information-centric networks," *Network and Service Management, IEEE Transactions on*, vol. 10, no. 3, pp. 286–299, September 2013.

[25] S. Podlipnig and L. Böszörmenyi, "A survey of web cache replacement strategies," *ACM Computing Surveys*, vol. 35, no. 4, pp. 374–398, 2003.