Cost-Aware Optimisation of Cache Allocation for Information-Centric Networking

Haozhe Wang, Jia Hu*, Geyong Min and Wang Miao
College of Engineering, Mathematics and Physical Sciences
University of Exeter,
Exeter, UK
Email: {h.wang3, j.hu, g.min, wm255}@exeter.ac.uk

Nektarios Georgalas
Research and Innovation
British Telecom,
UK
Email: nektarios.georgalas@bt.com

***Abstract*—Information-centric networking (ICN) is an emerging paradigm that decouples content from the host to achieve fast and cost-efficient communication and content distribution in the future Internet. A key feature of ICN is the deployment of ubiquitous in-network caching to speed up service delivery and improve network resource utilisation. ICN caching has been widely studied in terms of caching strategies and caching performance. However, the economic aspect of ICN has received marginal consideration so far, although it is vital to understand the potential cost-efficiency of ICN before its wide deployment in service provider network. To address this issue, we propose a cost-aware caching scheme to study the Quality-of-Service (QoS) and cost of ICN and investigate the inner association between them. Two new models are designed to characterise the cost and QoS of ICN with arbitrary topology under heterogeneous bursty content requests. A multi-objective evolution algorithm is adopted to find the optimal cache resource allocation. Numerical results show the effectiveness of the proposed scheme in achieving cost-efficiency and QoS guarantee in ICN caching.**

I. Introduction

The current Internet architecture is built on the host-to-host model, aiming at resource sharing and end-to-end connection. However, with the emerging content-rich applications such as Internet-of-Things, mobile cloud computing and multimedia services [1], the focus of the Internet has shifted from host centric to content centric. Information-centric networking (ICN) emerges as a promising architecture for the future Internet to meet the increasing demand for scalable and efficient content distribution. ICN treats content as the first class entity in the network architecture and decouples content from host to achieve fast and cost-efficient communication and content dissemination in the future Internet.

In order to alleviate the pressure of high demands for network bandwidth and service quality posed by various services, a key feature of ICN is the deployment of transparent, ubiquitous in-network caching to improve network resource utilisation and reduce service latency [2]. In-network caching in ICN differs from the present caching of Internet, because ICN caching is transparent to applications and content object to be cached is finer-grained. This poses new challenges for in-network caching: 1) Various services differ considerably in their request rate, content size and popularity, this heterogeneity requires ICN to efficiently share and allocate cache resources among different services. 2) Requests in traditional file-based caches are typically assumed to follow the independent reference model (IRM), which is not valid in ICN caching where correlations are common among requests in an arbitrary cache network.

ICN caching has been widely studied in terms of caching strategies and caching performance [3]–[7]. Besides, some research [8]–[10] focused on the optimisation of cache allocation to answer the questions that where the cache resources should be placed and how much cache should be allocated wherein. Despite the many efforts on ICN caching, the economic aspect of ICN has received marginal consideration so far [11]. But it is vital to understand the potential cost-efficiency of ICN before its wide deployment in Internet Service Providers' (ISPs) network. To address this problem, Pham et at. [12] proposed a game-based pricing model that provides economic incentives for caching and sharing content in ICN. Kocak et al. [13] also utilised game theory to study a price-convex demand-response pricing model. Both works focused on the interaction between different players in the network without considering the caching strategy design and performance. Two models were proposed in [14] to investigate the impact of content retrieval cost on the caching design, but the paper made strong assumptions for the models such as unrealistic content requests and simplified topology. Therefore, a new cache allocation scheme that considers more realistic network scenarios such as arbitrary network topology and heterogeneous bursty content requests and can dynamically optimise the cache resource distribution is in demand.

To foster the practical deployment of ICN in the future Internet for service providers, it becomes crucial to understand the performance and cost bounds and trade-off of ICN caching. Therefore, this paper proposes a cost-aware QoS optimisation scheme for cache resource allocation in ICN. The main contribution of this paper is threefold:

(i) Two new models are developed to investigate the relation between economic cost and network performance in ICN. The cost model considers deployment and operation cost of ICN caching under a realistic ISP network. The QoS model considers multiple key networking and caching parameters and formulates the service delay.
(ii) Arbitrary network topology, heterogeneous bursty content requests and different content popularity distributions are considered to provide a more practical ICN environment.
(iii) We propose a cost-aware caching scheme that jointly optimises the cost and QoS of ICN. A multi-objective evolution algorithm is adopted to find the optimal caching resource allocation.

The remainder of the paper is organised as follows. Section II is devoted to the design of the cost model and

TABLE I
SUMMARY OF NOTATIONS

Parameter	Meaning
N	Number of ICN nodes in the network
E	Set of links between nodes
K	Number of types of services
p_e^{cap}	Device cost for an ICN node
$p_{c,n}^{cap}$	Caching unit cost for ICN node n
$p_{l,n}^{cap}$	Bandwidth unit cost for ICN node n
C_n	Cache size of node n
BW_n	Amount of bandwidth supported by node n
$p_{c,n}^{op}$	Cache unit cost for operation at node n
$p_{s,n}^{op}$	Unit cost for retrieving a content at node n
$p_{t,n}^{op}$	Unit cost for forwarding a content request at node n
r_n, $\overline{r_n}$	Number of requests satisfied/forwarded at node n
λ_n^k	Arrival rate of request for contents of service k at node n
$\lambda_{tot,n}^k$	Combined content request rate for service k at node n
$h_{v_n}^k$	Cache hit ratio of requests for service k at node n
S^k	Average delay of service k
$s_{m,o}^k$	Delay of requests from node m for content o of service k
Ω	Set of contents in the network
$\boldsymbol{\Omega_k}$	Total number of different contents of service k
$\boldsymbol{P^k}$	Set of cache hit rate for service k at nodes
$\boldsymbol{L}$	Set of link delay
σ_n	Fraction of requests at node n to the total network traffic
$\boldsymbol{p_{m,n}^k}$	Set of hit probability for service k at each node along the path from m to n
$\boldsymbol{l_{m,n}}$	Link delay of each segment along the path from m to n
α_n	Zipf exponent characterizing the skewness of popularity
τ_n^k	Time interval for generating a request missing process
ρ_n^k	Fraction of requests for service k at node n

QoS model followed by the analysis of cache performance. Section III proposes the cost-aware caching scheme and gives solution for the associated multi-objective optimisation problem. The effectiveness of the proposed scheme is evaluated by the numerical simulations in Section IV. Finally, Section V concludes the paper.

II. SYSTEM MODEL

This section develops the cost model and QoS model followed by the caching performance analysis which is used to quantify the models. Table I summarises the notations.

A. Cost Model

ICN aims for increasing the efficiency of content distribution, but whether it can be cost-effective for service providers is a key enabler for its realistic implementation. Bearing in mind the cost-efficiency issue, we develop a new model to evaluate the cost for ICN nodes, which includes capital expenditure (CAPEX) and operational expenditure (OPEX).

The CAPEX represents the installation expenditure of ICN devices. Assuming a fixed device cost (p_e^{cap}), the rest of cost is determined by two other specifications, i.e, the cache size and the bandwidth. We denote C as the cache size of an ICN node, and BW as the bandwidth supported by that node. The unit cost for cache and bandwidth are denoted as p_c^{cap} and p_l^{cap} that are various depending on locations. Therefore, the cost of CAPEX for one node, $Cost_n^{\text{CAPEX}}$ is expressed as

$$Cost_n^{\text{CAPEX}} = p_e^{cap} + p_{c,n}^{cap} \cdot C_n + p_{l,n}^{cap} \cdot BW_n \tag{1}$$

The OPEX includes the caching cost and traffic cost. The caching cost, $Cost_c^{op}$, is composed by two factors, the cost of storing that is proportional to cache size, and the cost of retrieving content from the cache when incoming requests are hit in the cache. The traffic cost, $Cost_t^{op}$, is proportional to the traffic forwarded to the neighbour nodes when requests are not satisfied by the cache. Because both costs depend on the unit cost related to the location, such as electricity tariff and rental fee, we denote the unit price for caching cost, retrieving cost and forwarding cost as p_c^{op}, p_s^{op} and p_t^{op}, respectively. Then the OPEX for node n is given by

$$\begin{aligned} Cost_n^{\text{OPEX}} &= Cost_{c,n}^{op} + Cost_{t,n}^{op} \\ &= p_{c,n}^{op} \cdot C_n + p_{s,n}^{op} \cdot r_n + p_{t,n}^{op} \cdot \overline{r_n} \end{aligned} \tag{2}$$

where r_n denotes the amount of content requests that are satisfied by the cache at node n, and $\overline{r_n}$ denotes the requests that are missed at the current node and forwarded to the adjacent nodes. To calculate the amount of requests that are satisfied by the cache, we denote the arrival rate of requests for service k as λ^k, and the hit ratio for such service as $h_{v_n}^k$. Since the arrival rate and cache hit ratio are different for each service at each node in a realistic scenario, we firstly consider the cost for serving one kind of service. Let λ_n^k denote the number of requests for service k arrived at node n, and $h_{v_n}^k$ as the proportion of those requests that are hit in the cache, then we have

$$r_n = \sum_{k \in K} \lambda_n^k \cdot h_{v_n}^k \quad \overline{r_n} = \sum_{k \in K} \lambda_n^k \cdot (1 - h_{v_n}^k) \tag{3}$$

Eq. (3) is only valid under the independent reference model (IRM), which means that the incoming requests for the same content at a node have the same probability and does not depend on any other sources. However in a network of caches, because the miss requests at one node are forwarding and becoming part of the incoming demand of its neighbours, the effects of miss requests from neighbouring nodes should be considered, which can be given by

$$r_n^k = \lambda_{tot,n}^k \cdot h_{v_n}^k \tag{4}$$

$$\lambda_{tot,n}^k = \lambda_n^k + \sum_{n' \in E_{<n,n'>}} \overline{r_{n'}^k} \tag{5}$$

$$\overline{r_n^k} = \lambda_{tot,n}^k (1 - h_{v_n}^k) \tag{6}$$

Accordingly, Eq. (3) should take into account the additional content requests and be updated as

$$\begin{aligned} r_n &= \sum_{k \in K} (\lambda_n^k + \sum_{n' \in E_{<n,n'>}} \overline{r_{n'}^k}) h_{v_n}^k \\ \overline{r_n} &= \sum_{k \in K} (\lambda_n^k + \sum_{n' \in E_{<n,n'>}} \overline{r_{n'}^k})(1 - h_{v_n}^k) \\ &= \sum_{k \in K} (\lambda_n^k + \sum_{n' \in E_{<n,n'>}} \overline{r_{n'}^k} - r_n^k) \end{aligned} \tag{7}$$

From Eq. (7) we can see that r_n^k and $\overline{r_n^k}$ are functions of each other, and both of them relied on the exogenous arriving rate λ_n^k and the cache hit ratio. The calculation of cache hit ratio of the node is detailed in Sec. II-C, then we can use Eq. (4)-(6) to iteratively solve the equations.

B. QoS Model

Delay of service is one of the most important factor to measure QoS. For delay-sensitive services, such as video and voice, the QoS is largely determined by service delay, which has a similar role as round-trip time in the TCP. So we use service delay to quantify the service quality. Let S^k denote the delay for service k, then it can be written as a function of

$$S^k = \mathcal{F}(\boldsymbol{\Omega_k}, \boldsymbol{P^k}, \boldsymbol{L}) \tag{8}$$

where $\boldsymbol{\Omega_k}$ is a set of content objects contained in service k, i.e., $\boldsymbol{\Omega_k} = \{o_i^k | i \in \Omega, k \in K\}$; $\boldsymbol{P^k}$ is a vector with the elements denoting the probabilities that the requested service is satisfied by the node on the path to content providers' servers, hence $\boldsymbol{P^k} = [h^k_{v_1}, h^k_{v_2}, \ldots, h^k_{v_N}]$; L is a vector that denotes the link delay along the path, which can be written as $\boldsymbol{L} = [l_{1,2}, \ldots, l_{N-1,N}]$, where $l_{i,j}$ is the link delay between node v_i and node v_j. Similarly, we define the link delay of each segment of the path from v_m to v_n as $\boldsymbol{l_{m,n}} = [l_{m,m+1}, \ldots, l_{n-1,n}]$.

Furthermore, we define vector $\boldsymbol{p^k_{m,n}}$, which denotes the probability that requests for service k arriving at node m is satisfied by node n, and can be easily generated from $\boldsymbol{P^k}$ as $\boldsymbol{p^k_{m,n}} = [1 - h^k_{v_m}, 1 - h^k_{v_{m+1}}, \ldots, h^k_{v_n}]$ where v_{m+1} means the next hop towards node n from node m.

Finally, Eq. (8) can be expressed as

$$S^k = \sum_{m \in N} \sigma_m \cdot S^k_m = \sum_{m \in N} \sigma_m \cdot |\boldsymbol{\Omega_k}| \cdot s^k_{m,o} \tag{9}$$

where σ_m is the ratio of the incoming requests at node m to that at the whole network, $|\boldsymbol{\Omega_k}|$ is the number of requested objects, and $s^k_{m,o}$ denotes the average delay for retrieving a single object o in service k at node m. It can be computed as

$$\begin{aligned} s^k_{m,o} &= h^k_{v_{m,o}} + (1 - h^k_{v_{m,o}}) h^k_{v_{m+1,o}} l_{m,m+1} + \ldots \\ &\quad + (1 - h^k_{v_{m,o}})(1 - h^k_{v_{m+1,o}}) \ldots (1 - h^k_{v_{m+d-1,o}}) \\ &\quad h^k_{v_{m+d^k_m,o}} \sum\nolimits_{i=1}^{m+d^k_m} \boldsymbol{l}_{m,m+d^k_m}(i) \\ &= \sum_{n=m}^{m+d^k_m} \prod_{q=1}^{n-m} \boldsymbol{p}^k_{m,n}(q) \sum_{i=1}^{n-m} \boldsymbol{l}_{m,n}(i) \end{aligned} \tag{10}$$

where d^k_m denotes the maximum distance (i.e., the number of hops) travelled by a request arriving at node m for service k. Note that the contents in the same service are requested under the same probability, for the sake of simplicity, we assume that the cache hit ratio is independent of the specific object, thus, $h^k_{v_m}$ is equivalent to $h^k_{v_{m,o}}$. At last, we can obtain the average delay for service k as

$$S^k = |\boldsymbol{\Omega_k}| \sum_{m \in N} \sigma_m \sum_{n=m}^{m+d^k_m} \prod_{q=1}^{n-m} \boldsymbol{p}^k_{m,n}(q) \sum_{i=1}^{n-m} \boldsymbol{l}_{m,n}(i) \tag{11}$$

C. Calculation of cache Performance

Cache performance, such as cache hit ratio is crucial for quantifying the cost and QoS in ICN. Cache hit ratio is determined by multiple parameters, such as cache size, traffic pattern, cache strategy, network topology, and content catalog and popularity distribution. For calculating the cache hit ratio, the following parameters and assumptions are used in this paper.

1) Caching strategy: Caching strategy includes cache decision policy and cache replacement policy. A simple standard Leave Copy Everywhere (LCE) policy for ICN [15] is considered in the paper, because caching operation is required to run at line-rate and complex policies are not suitable due to high computation and communication overhead.Least Recently Used (LRU) replacement policy is used for each node, since LRU has low complexity and satisfies the line speed operation requirement [2].

2) Traffic pattern: Markov-modulated Poisson process (MMPP) is leveraged to model the heterogeneous and bursty nature of ICN traffic. MMPP is capable of modelling the bursty content requests because it captures the time-varying arrival rate of various services via two matrices [16]. The arrival matrix Λ_n represents the intensity of service requests under different states at a node, while the state transition matrix Q_n represents the hidden state transition process.

3) Topology: We consider an ICN network with arbitrary topology, which is more realistic compared to specific hierarchical or flat topology.

4) Content distribution: Cache hit ratio depends crucially on the popularity of different contents. The popularity of contents in the Internet has been observed following a Zipf distribution, which has been used in [4], [6] to characterise the content distribution of ICN. Heterogeneous popularities are considered with various Zipf exponent α, which characterises the skewness of distribution.

Next, we can evaluate the performance of caching in ICN under arbitrary network and bursty traffic. We leverage the model that we developed in [17], so the cache hit ratio for service k at node n can be written as

$$h^k_{v_n} = 1 - \mathbb{P}(Req_{v_n}(\tau^k_n) \geqslant C_n) \tag{12}$$

where $Req_{v_n}(\tau^k_n)$ denotes that the number of different content requests arriving at node v_n between two subsequent requests of the same content in service type k during an inter-arrival time τ^k_n. τ^k_n is a key parameter to the generation of a cache missing, and largely depends on the traffic pattern, cache size and popularity distribution. τ^k_n can be expressed as a function of multiple variables

$$\tau^k_n = \mathcal{F}(C_n, \Lambda_n, Q_n, \alpha_n), \quad \forall k \in K \tag{13}$$

The bursty arrival rate at node n for service k, λ^k_n can be given by

$$\lambda^k_n = \rho^k_n \cdot \lambda_n = \rho^k_n \cdot \pi_n \cdot \Lambda_n \tag{14}$$

$$\rho^k_n = \frac{1/k^{\alpha_n}}{\sum_{i=1}^{K} 1/i^{\alpha_n}}, k \in K \tag{15}$$

where ρ^k_n is the fraction of requests for service k at node n, following the Zipf distribution, and π_n is the steady-state vector that subjects to $\pi_n Q_n = 0, \pi_n \boldsymbol{e} = 1$. In our previous work [17], the inter-arrival time, τ^k_n had been derived as

$$\begin{aligned} \tau^k_n &= C^{\alpha_n}_n / g^k_n \\ g^k_n &= \frac{1}{2} \cdot \Gamma(1 - \frac{1}{\alpha_n})^{\alpha_n} \frac{\sum_{k \in K} \lambda^k_{tot,n}}{\sum_{k \in K} 1/k^{\alpha}_n} \cdot |\Omega_k|^{\alpha_n} \end{aligned} \tag{16}$$

At last, the cache hit ratio for service k at node n, $h^k_{v_n}$ can be expressed in the form of

$$h^k_{v_n} = 1 - \beta^k_n e^{-u^k_n \tau^k_n} - (1 - \beta^k_n) e^{-v^k_n \tau^k_n} \tag{17}$$

where β_n^k, u_n^k and v_n^k are the parameters derived from MMPP [17]. We omit the details here for the reason of limited space.

III. Cost-Aware Optimal Cache Allocation

In this section, we first provide the rational reason behind our method, showing that the cost efficiency and QoS guarantee are opposite goals. Next, we propose a cost-aware optimal caching scheme that integrates the cost model and QoS model as a multi-objective optimal problem, which is solved through an evolutionary algorithm. At last, we show that the proposed scheme can be scale-up for large scale networks.

A. Multi-objective Optimisation Goals

Most recent caching allocation schemes aim at optimising the cache hit ratio, however, according to the two models presented in Section II, a large cache will increase both CAPEX and OPEX. The improvement of QoS and reduction of the cost are conflicting. In the mean time, because cost depends on locations and quality demands are related to services, there is a trade-off between the cost and QoS. To investigate this trade-off, we design two goals to explicitly describe two conflicting caching strategies.

$$Obj\_Cost = min\{\sum_{n\in N} Cost_n\} \tag{18}$$

$$Obj\_QoS = min\{\mathbb{E}(S^k), \forall k \in K\} \tag{19}$$

Obj_Cost aims to find the cache placement strategy that leads to the minimum cost of the whole network, while Obj_QoS is optimal for minimising the service delay for various services, which means maximising the cache hit ratio of nodes.

The solution for either goal must yields to the following design constrains: 1) *Cache budget*: the total size of cache to be deployed in the network has an upper bound, C_{tot}; 2) *Service Level Agreement (SLA)*: the requirements of QoS for various services differ from each other. For each type of service k, the global maximum delay should be less than $\widehat{S^k}$; 3) *Bandwidth capability*: the volume of traffic being transmitted at one node should be less than the bandwidth capability of that node, BW_n; 4) *Cost budget*: for service providers, the cost of network should be kept low to make profits and remain competitive. Therefore, a cost budget $Cost_{max}$ is set to control the total cost.

To guarantee the SLA, the main objective is to maximise the cache hit rate. In practice, it is desirable to place the cache resources at the nodes that receive large amounts of requests. Nevertheless, it is difficult to balance the network cost and QoS performance. Therefore, a new optimal caching scheme is designed to balance the cost and QoS goals. The multi-objective optimisation problem can be described as follows:

$$\begin{aligned} & min[Obj\_Cost, Obj\_QoS] \\ \text{s.t.}, & \forall n \in N : \sum_{n\in N} C_n \leqslant C_{tot},\ C_{min} \leqslant C_n \leqslant C_{max} \\ & \forall k \in K : \mathbb{E}(S^k) \leqslant \widehat{S^k} \\ & \forall n \in N : \sum_{k\in K} \lambda_{tot,n}^k \leqslant BW_n \\ & \forall n \in N : \sum_{n\in N} Cost_n = Cost_{tot} \leqslant Cost_{max} \end{aligned} \tag{20}$$

B. Design of a Cost-aware Caching scheme

We propose a cost-aware caching scheme in order to achieve significant saving in cost and also guarantee the SLA. To achieve this, we leverage the two models developed in Sec. II to investigate the inner connection between the two goals.

The goal function of network cost can be written as

$$\begin{aligned} \sum_{n\in N} Cost_n &= \sum_{n\in N} Cost_n^{\text{CAPEX}} + Cost_n^{\text{OPEX}} \\ &= \sum_{n\in N} p_e^{cap} + (p_{c,n}^{cap} + p_{c,n}^{op})C_n + p_{l,n}^{cap} \cdot BW_n \\ &\quad + p_{s,n}^{op} \cdot r_n + p_{t,n}^{op} \cdot \overline{r_n} \end{aligned} \tag{21}$$

The cost goal is determined by the cache size, location of node, intensity of requests and cache hit ratio. While the goal of QoS

$$\begin{aligned} \mathbb{E}(S^k) = \sum_{k\in K} \rho^k \cdot S^k &= \sum_{k\in K} |\Omega_k| \sum_{m\in N} \rho_m^k \cdot \sigma_m \\ &\quad \sum_{n=m}^{m+d_m^k} \prod_{q=1}^{n-m} \boldsymbol{p}_{m,n}^k(q) \sum_{i=1}^{n-m} \boldsymbol{l}_{m,n}(i) \end{aligned} \tag{22}$$

is determined by the topology, content distribution, traffic pattern and cache hit ratio.

Topology and traffic pattern have impact on both goals. However, the evaluation of all possible cache placements to find the Pareto optimal results with respect to predefined constrains for a particular network topology with different costs, bursty content requests, and various popularity distributions is very complex, time-consuming, and beyond the computation resources of machines. Evolutionary algorithms are popular for solving multi-objective optimisation. To accelerate the convergence speed and maintain the diversity, the non-dominated sorting genetic algorithm II (NSGA-II) is used in the paper. The input for the scheme consists of two parts. One is the population vector of cache resources allocated for each node. In the case of no prior knowledge, initial population vector is generated according to the uniform random distribution and subjected to the constrain. The other part including the topology information, request rate of nodes, popularity distribution and location-related cost is fed into the scheme to evaluate the objectives.

The complexity of the scheme largely depends on NSGA-II and increases linearly with the size of network and variety of service. The computational complexity of one iteration is governed by the non-dominated sorting of the two goals that requires $O(2\times S_p^2)$, where S_p is the population size of the first non-dominated front. Combining with calculation of cache hit ratio, the overall complexity is $O(N \times K \times S_p^2)$, where N and k are the number of nodes and services. Therefore, the proposed scheme can be scale-up for large network and multiple services with the linear increasing of complexity.

IV. Numerical Evaluation

In this section, we present numerical simulations to demonstrate the effectiveness of the proposed cost-aware optimal caching scheme. The goal is to find an optimal caching placement strategy for saving the network cost while maintaining the service quality.

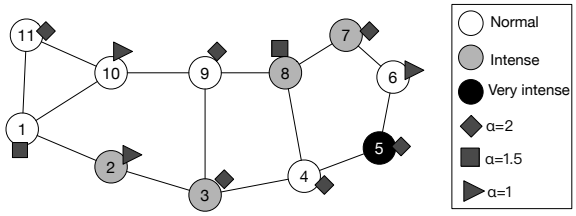

Fig. 1. Network topology of the simulation with heterogeneous content requesting rates and content popularity distributions. Each node represents the aggregation of multiple ICN edge routers that receive content requests from end-users.

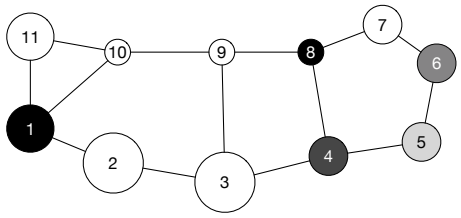

Fig. 2. The diversity of node cost based on location shown on the topology. Higher CAPEX nodes are darker in colour and higher OPEX nodes are larger in size.

A. Experiment Setup

Network topology. We consider a real-world network as the representative network topology, which is shown in Fig. 1. The network contains 11 nodes and each node represents the aggregation of multiple ICN edge routers that receive content requests from end-users. We choose the aggregation to simplify the topology and model representation, since the cost is location-based, there is no need for the calculation on each single node level within the same location.

Content request. All the nodes in the network receive bursty content requests that are modelled by MMPP. In order to represent the heterogeneous scenario, the nodes receive various rates of content requests with different content popularity. As shown in Fig. 1. Three types of arriving rate and content popularity are considered.

Diversity cost. Nodes at different locations have different CAPEX and OPEX. 5 types of prices for CAPEX and 4 types of prices for OPEX are considered, illustrated in Fig. 2. Without loss of generality, the unit prices are randomly assign to every node.

The setting for variables are summarised in Table II. Furthermore, we consider 10 types of services and a content catalog consisting of 10^6 contents. The contents are participated into 10 different services following the Zipf distribution. The total cache budge, $C_{tot} = \{1.0e6, 1.2e6, 1.5e6\}$, is in the unit of content. To avoid the extreme case where the cache size is too small, the minimum cache size is set to $C_{min} = 5000$ contents, and $C_{max} = C_{tot} - (N-1)C_{min}$. The link delay is set to $2ms$ for all links between ICN nodes. The maximum delay $E(S^k)$ is set to $350ms$. One repository that contains all the contents is connected to node 3. The bandwidth supported by each node and the corresponding cost, BW_n and $p^{cap}_{l,n}$, are set to same values for all nodes as 10^6 and 0.2, respectively.

TABLE II
SETTINGS OF VARIABLES FOR EVALUATION

Node	Λ_n[i]	α_n	$p^{cap}_{c,n}$[ii]	$p^{op}_{c,n}$	$p^{op}_{s,n}$	$p^{op}_{t,n}$
1	10	1.5	5	1.2	1.6	1.6
2	60	1.1	1	1.5	1.8	1.8
3	60	2	1	1.5	1.8	1.8
4	10	2	4	1.1	1.4	1.4
5	100	2	2	1.1	1.4	1.4
6	10	1.1	3	1.1	1.4	1.4
7	60	2	1	1.1	1.4	1.4
8	60	1.5	5	1	1.2	1.2
9	10	2	1	1	1.2	1.2
10	10	1.1	1	1	1.2	1.2
11	10	2	1	1.2	1.6	1.6

[i] The arrival rate is contents/second.
[ii] All the cost is expressed as the ratio to the basic price.

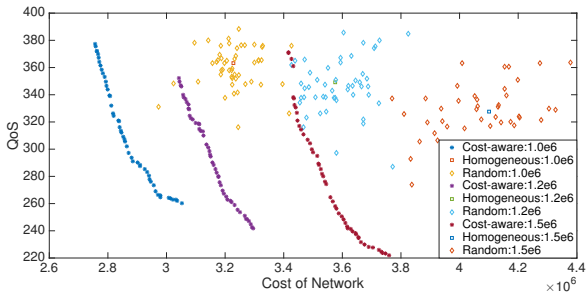

Fig. 3. Pareto optimal set under the goals (cost of network vs service delay) comparing with homogeneous and random cache allocation methods with cache budget $C_{tot} = \{1.0e6, 1.2e6, 1.5e6\}$.

B. Performance Evaluation

This subsection demonstrates the performance of the proposed optimal caching scheme. The achieved optimal results are compared with the equal caching allocation which is the default method for ICN and the random caching allocation to verify the effectiveness of the proposed scheme. The population size is set as 50 and the maximum generation is set as 500 for the evolutionary algorithm. The crossover probability for NSGA-II is set to 0.9, and the mutation probability is $1/N$.

Fig. 3 depicts the Pareto optimal set of the Obj_Cost versus Obj_QoS trade-off under three different cache budgets. The optimal set can be leveraged by network managers to select the most appropriate solutions for different purposes. The multi-objective optimisation results show that the cost of network and service delay are conflicting goals. The figure illustrates that with some service quality losses, a significant reduction in the cost of network can be achieved. For example, by increasing the total cost for 4%, a significant improvement up to 25% can be achieved in service quality.

Figs. 4 illustrates the results of different caching allocation methods under the same experimental environment. The results of the cost-aware caching scheme in the two figures are the mean values of the Pareto frontier set. The results of random method are the mean values of 50 allocation cases. The figures show that the proposed cost-aware caching scheme outperforms the homogeneous and random cache placement methods in terms of both network cost and service delay. As shown in the figures, the cost of network is up to 15.1% less than the homogeneous method and 14.2% less than the random method. As for the service delay, the cost-

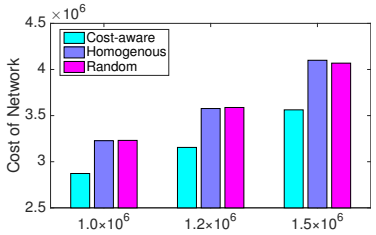

Fig. 4. Comparing of network cost among three cache allocation methods under caching budget $C_{tot} = \{1.0e6, 1.2e6, 1.5e6\}$.

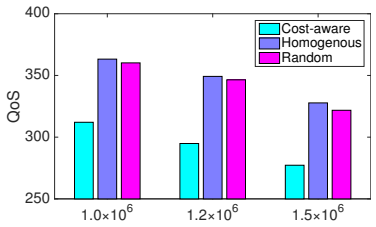

Fig. 5. Comparing of service quality among three cache allocation methods under caching budget $C_{tot} = \{1.0e6, 1.2e6, 1.5e6\}$.

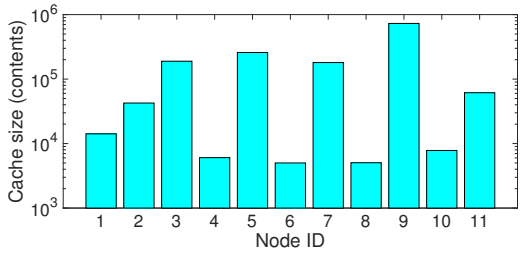

Fig. 6. Cache resource allocates for each node. The result is the median of Pareto optimal set under cache budget $C_{tot} = 1.5e6$.

aware caching placement reduces the delay up to 18.4% and 17.5% comparing to homogeneous and random methods.

Fig. 6 shows the cache allocations result to the nodes. There are different factors that could impact the allocation. One key metric that has a significant impact on cache is the popularity distribution. When Zipf parameter is larger, e.g., $\alpha = 2$ in the experiment, the allocated cache sizes tend to be larger as well, such as Nodes $3, 5, 7, 9, 11$, although the costs are different on these nodes. Node 9 is given the largest cache due to the low cost and its location. Node 9 locates at the core network and on the shortest path of multiple nodes to the repository. A higher cache hit ratio on the node will reduce the delay for many missing requests. Furthermore, cost also has a moderate influence on the cache allocation. For nodes with a higher cost, such as Nodes $1, 4, 8$, even though they are located on the paths of many nodes to the repository, the size of their caches are small. The cache sizes of Nodes $2, 6, 10$ are the synthetic impacts of the factors. The cost cannot be the dominating factor and needs to be balanced with the QoS, since the SLA has to be assured during the optimisation.

V. Conclusions

In this paper, a cost-aware optimal caching scheme has been proposed to find the trade-off between cost of network and QoS for ICN. A cost model has been developed to capture the main parameters that contribute to the cost of ISP in ICN, and a QoS model has been derived to measure the service quality by quantifying the service delay under heterogeneous bursty content requests and arbitrary network topology. This trade-off has been solved as a multi-objective optimisation problem, which aims to optimize both the QoS and network cost. Numerical simulations have been conducted to evaluate the effectiveness of the proposed scheme. The results have shown that the proposed scheme can achieve better QoS and lower cost comparing to the homogeneous and random cache allocation methods. Some insights have been observed from the optimal cache allocation: content distribution has the most significant impact on the cache allocation; Node location, which determines the traffic intensity and routing path, also greatly affects the allocating decision; Unit price of CAPEX and OPEX has a moderate impact on the cache allocation.

References

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2015-2020," Tech. Rep., 2016.

[2] M. Zhang, H. Luo, and H. Zhang, "A Survey of Caching Mechanisms in Information-Centric Networking," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 3, pp. 1473–1499, 2015.

[3] A. Ioannou and S. Weber, "A Survey of Caching Policies and Forwarding Mechanisms in Information-Centric Networking," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 4, pp. 2847–2886, 2016.

[4] W. K. Chai, D. He, I. Psaras, and G. Pavlou, "Cache "less for more" in information-centric networks (extended version)," *Comput. Commun.*, vol. 36, no. 7, pp. 758–770, 2013.

[5] M. Garetto, E. Leonardi, and V. Martina, "A Unified Approach to the Performance Analysis of Caching Systems," *ACM Trans. Model. Perform. Eval. Comput. Syst.*, vol. 1, no. 3, pp. 1–28, may 2016.

[6] G. Rossini and D. Rossi, "A dive into the caching performance of Content Centric Networking," in *2012 IEEE 17th Int. Work. Comput. Aided Model. Des. Commun. Links Networks*. IEEE, sep 2012, pp. 105–109.

[7] M. Mangili, F. Martignon, and A. Capone, "A comparative study of content-centric and content-distribution networks: Performance and bounds," *GLOBECOM - IEEE Glob. Telecommun. Conf.*, pp. 1403–1409, 2013.

[8] Y. Wang, Z. Li, G. Tyson, S. Uhlig, and G. Xie, "Design and Evaluation of the Optimal Cache Allocation for Content-Centric Networking," *IEEE Trans. Comput.*, vol. 65, no. 1, pp. 95–107, 2016.

[9] D. Rossi and G. Rossini, "On sizing CCN content stores by exploiting topological information," in *2012 Proc. IEEE INFOCOM Work.* IEEE, mar 2012, pp. 280–285.

[10] I. Psaras, W. K. Chai, and G. Pavlou, "In-network cache management and resource allocation for information-centric networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2920–2931, 2014.

[11] P. Agyapong and M. Sirbu, "Economic incentives in information-centric networking: implications for protocol design and public policy," *IEEE Commun. Mag.*, vol. 50, no. 12, pp. 18–26, dec 2012.

[12] T.-m. Pham, S. Fdida, and P. Antoniadis, "Pricing in Information-Centric Network Interconnection," in *IFIP Netw. Conf.*, 2013, pp. 1–9.

[13] F. Kocak, G. Kesidis, T.-M. Pham, and S. Fdida, "The Effect of Caching on a Model of Content and Access Provider Revenues in Information-centric Networks," in *2013 Int. Conf. Soc. Comput.* IEEE, sep 2013, pp. 45–50.

[14] A. Araldo, M. Mangili, F. Martignon, and D. Rossi, "Cost-aware caching: Optimizing cache provisioning and object placement in ICN," *2014 IEEE Glob. Commun. Conf. GLOBECOM 2014*, pp. 1108–1113, 2014.

[15] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proc. 5th Int. Conf. Emerg. Netw. Exp. Technol. - Conex. '09*. New York, New York, USA: ACM Press, 2009, p. 1.

[16] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," *Perform. Eval.*, vol. 18, no. 2, pp. 149–171, 1992.

[17] H. Wang, G. Min, J. Hu, W. Miao, and N. Georgalas, "Performance Evaluation of Information-Centric Networking for Multimedia Services," in *2016 IEEE Symp. Serv. Syst. Eng.* IEEE, mar 2016, pp. 146–151.