Distributed Caching Mechanism for Popular Services Distribution in Converged Overlay Networks

Wei Zhang, Jian Xiong, *Member, IEEE*, Lin Gui, *Member, IEEE*, Bo Liu, *Member, IEEE*, Meikang Qiu, *Senior Member, IEEE*, and Zhiping Shi

***Abstract*—With the proliferation of portable devices, the exponential growth of the global mobile traffic brings great challenges to the traditional communication networks and the traditional wireless communication technologies. In this context, converged networks and cache-based data offloading have drawn more and more attention based on the strong correlation of services. This paper proposes a novel popular services pushing and caching scheme by using converged overlay networks. The most popular services are pushed by terrestrial broadcasting networks. And they are cached in *n* router-nodes with limited cache sizes. Each router-node only interconnects with its neighbor nodes. Users are served through the router's WiFi link. If the services requested are cached in the routers, the user can be immediately responded; otherwise, the requests can be responded through the link from cellular stations to the router. In the proposed scheme, the cache size of the router, the maximum number of requests each router can serve, and the whole-time delay are limited. Three node-selecting and dynamic programming algorithms are adopted to maximize the equivalent throughput. Analytical and numerical results demonstrate that the proposed scheme is very effective.**

***Index Terms*—Converged overlay networks, distributed caching, offloading, popular services pushing and caching.**

I. Introduction

THE GLOBAL mobile traffic grew 63 percent in 2016 wih the proliferation of portable devices; and it will keep the growth rate of more than 60 percent annually [1]. The growth of traffic demand results in the heavy loads and the long-time crowed backhaul of the cellular networks. Hence, the bandwidth gain of the emerging novel wireless communication technologies is quickly offset; and the wireless communication bandwidth resource is still scarce. Improving the current spectrum resource is one of the main targets of wireless communication research. Usually, there are two ways to deliver these services. One way is to increase the bandwidth of users by using traditional physical technologies, such as the fifth generation of wireless communication technologies. Massive MIMO technology is one of the most efficient methods to improve the spectrum efficiency; and this technology can further improve the throughput rate [2]. Another way is to offload the most popular services, e.g., D2D interest sharing, WiFi offloading and converged broadcast and cellular networks [3]–[5]. They can offload the traffic by eliminating the correlation of the services.

In recent years, converged broadcast and cellular networks have drawn more and more attention in view of the strong correlation of services. There are several q̋olong tailq́ probability distributions to describe the phenomenon that the rich get richer (Rich-Get-Richer), e.g., Power-Law distribution, Zipf distribution and Pareto distribution [6]. Easley and Kleinberg [7] reveals that these models are equivalent.

Converged networks are researched in many literatures. In the converged networks, broadcast network can be used to preferentially transmit popular services; and the P2P cellular channels can respond individual requests. In industry, the third generation partnership project (3GPP) firstly defines the multimedia broadcast/multicast services (MBMS); and it further introduces Evolved MBMS (eMBMS) as the in-band convergence standard [8]. It is also one of the promising technology for forthcoming 5G technologies. Paper [9] researches the resource management in the dense heterogenous network (DenseNets). A hybrid unicast-multicast utility-based network selection algorithm (HUMANS) is proposed based on eMBMS scheme. The proposed scheme guarantees the outage percentage, average quality of transmission and more efficient resource utilization. Paper [10] analyzes the spectral coexistence between DVB-T2 and LTE networks. Its derivation and results provide further insight into the cooperation between the two networks. Paper [11] proposes an LTE hybrid unicast broadcast content delivery framework. However, the bandwidth efficiency is not improved since the broadcast

Manuscript received November 20, 2018; revised January 17, 2019; accepted February 15, 2019. Date of publication March 20, 2019; date of current version March 4, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61671295, Grant 61471236, and Grant 61420106008, in part by the Shanghai Key Laboratory of Digital Media Processing and Transmission, in part by the Shanghai Pujiang Program under Grant 16PJD029, in part by 111 Project under Grant B07022, and in part by the National Key Laboratory of Science and Technology on Communications under Grant KX172600030. *(Corresponding author: Jian Xiong.)*

W. Zhang, J. Xiong, and L. Gui are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xjarrow@sjtu.edu.cn).

B. Liu is with the Department of Engineering, La Trobe University, Bundoora, VIC 3086, Australia (e-mail: b.liu2@latrobe.edu.au).

M. Qiu is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: mq2203@columbia.edu).

Z. Shi is with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: szp@uestc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2019.2902818

content still occupies unicast channel in the proposed scheme. Furthermore, paper [12] proposes a cooperative structure of cellular network and broadcasting network by using cloud radio access network; and technical approaches for 3GPP and ATSC cooperation in physical layer are detailed and the cooperative frame structure of DVB-T2 and LTE are designed in this scheme. In order to deal with the unidirectional transmission of the broadcast system, the paper also designs dedicated return channel of broadcast network to enable a seamless interaction between broadcasters and users. A hit-rate based scheme and user fairness based scheme have been proposed respectively in [5] and [13]. Their goals are to maximize the systemąs equivalent throughput.

However, there are two main drawbacks of the traditional converged networks with centralized cache scheme if compared to distributed cache scheme: the limited storage capacity of users' devices and low efficiency of the centralized cache between users' devices.

On one hand, WiFi routers can be adopted overcome the problem of limited storage capacity of users' devices. Paper [14] gains the further insight of the WiFi offloading in heterogeneous networks without considering the broadcasting which takes the radio access technology as the starting point. Paper [15] considers the stochastic network with dynamic traffic; and independent Poisson point processes is used to model the spatial distribution of access points and users to reveal the impacts of network features on the network performance. Moreover, paper [16] proposes a novel pipeline network coding-based multipath transmission control protocol in heterogeneous wireless networks; and in this scheme, video services are delivered over a muti-path distribution network which consists of base stations and WiFi router nodes. Combining the converged broadcast and cellular networks with the WiFi offloading is a promising method to conquer the shortcoming of traditional converged networks. This converged overlay network consists of cellular network, broadcast network and router nodes which are overlapping to each other. There are places to cache the services under the converged overlay networks, for example, terminals, base stations, relays, and *et al.* With the help of WiFi router offloading, paper [17] researches the multi-layer converged networks in high-speed scenario. By adding a relay with WiFi router in the converged networks, the penetration loss and fading loss caused by high speed mobile Doppler effect can be overcome.

On the other hand, services can be distributively cached in the terminals. Distributed caching uses a dynamically scalable architecture, with multiple cache nodes sharing the load and accessing the cache content through addressing devices. Hence, it can improve the performance such as reliability, availability, scalability, and access efficiency [18]. In the content distribution network, the implementation mechanism mainly includes distributed caching based on D2D, distributed caching based on wireless mesh network and distributed caching based on network coding [3], [19], [20]. For wireless mesh networks based caching, paper [19] proposed a novel heterogeneous network architecture of hybrid LTE and wireless mesh routing, and proposed a corresponding routing protocol to realize the deep fusion between two different networks. For D2D based caching, both paper [3], [21] research the data offloading based on the incentive mechanism by D2D technologies; and reverse auction and the Stackelberg game theory are employed and researched. Moreover, paper [22] research the interference cancelation at receivers in cache-enabled wireless networks; and achieve lower packet loss rate in the D2D based distributed caching. For network coding based caching, paper [23] uses the maximum distance separable code (MDS) to enable any k segments of n segments to recover the whole content. Paper [20] uses a special network code stream to deliver individual information through the common channel in the proxy cache scenarios.

We combine the two-layer converged broadcast and cellular networks with WLAN in view of the aforementioned work. This combination can be traced from the traditional eMBMS heterogeneous network architecture (HetNets). Paper [24] proposed a distributed scheme of traffic offloading from multicast to WLAN in the current LTE and WLAN heterogeneous network. Moreover, the promising Software Defined Network (SDN) centralized approach is proposed to deal with the issues of resource management, implementation efficiency, and implementation flexibility to the network. In our proposed scheme, the services are pushed and distributively cached in the WiFi routers by broadcasting rather than broadband networks. As a result, the data traffic can be offloaded to improve the cache efficiency.

The early work and results of the research are published in paper [25]; and the paper is included in proceeding of broadband multimedia systems and broadcasting in 2018 (BMSB2018). A distributed cache scheme based on mesh WiFi router nodes is proposed to deal with the shortcomings of the centralized cache scheme in traditional converged networks; and we formulated the model based on the proposed scheme as an optimal math problem. This problem is based on the overlapped number of nodes' requests and is proved to be NPH problem by a brief deduction. Then, an alternation based node-selecting algorithm has been proposed to solve the problem.

The main contributions of this paper are summarized as follows by compared to the early work:

- Distributively caching model: To overcome the shortcomings of centralized cache in wireless converged network, such as low cache utilization rate, limited improvement of system throughput, a distributed caching mechanism based on wireless mesh network is introduced. Also, We define the load balancing coefficient to measure the load balance; and the overlapped number of nodes' requests is defined in a strict arithmetic expression. Then, we give a more generalized mathematic model; and it can be applied into a more universal scenario.
- Node-selecting algorithms: We formulate the ET objective of the distributed cache scheme as an optimal problem; and it is proved to be an NP-complete (NPC) problem. Two more node-selecting algorithms are proposed to choose the proper candidate router-nodes firstly; then, the dynamic programming algorithm

is adopted to cache the scheduling services to each router-node. Simulations show the efficiency of the proposed algorithms.

The remaining contents are organized as follows: Section II introduces the architecture of the converged overlay networks and the centralized caching scheme; the system model of the proposed scheme is given in Section III. In Section IV, the distributed cache problem is formulated. Numerical simulation results and discussions are shown in Section V. Finally, conclusions of this paper are drawn in Section VI.

II. Architecture of Converged Overlay Networks and Centralized Caching Scheme

A. Architecture of Converged Overlay Networks

There are mainly two types of architecture of the wireless converged networks:

Two-layer converged networks: the early studies mainly pay attention to the two-layer converged networks, in which cellular station and broadcast station directly connect to the user layer. In this scheme, popular services are pushed directly to mobile devices by broadcasting and less popular services are transmitted by cellular station. Paper [26] makes an extensive exploration of popular services pushing based on the mobile devices in converged cellular and broadcast networks. In the proposed scheme, the network capacity and response time are deduced from the queuing probability model which theoretically analyzes the characteristic of the networks.

Multi-layer converged networks: more attention are paid to multi-layer converged networks recently. An extra relay router is added into the networks and serves as the middle layer connecting station layer and user layer. In the multi-layer scheme, popular services are pushed to the relay router by broadcasting and the less popular services are delivered through the cellular unicast channel. Then, the user can request the services by the WiFi links of the relay router. Paper [17] researches on the multi-layer converged network in the high-speed-train scenario. Adding a cache-based relay router in the converged networks can overcome the penetration loss and big fast fading by caused high speed movement.

B. Centralized Caching Scheme and Solution

Current researches on cache-based wireless converged networks mainly focus on centralized scheme. Tt can be referred that both the above-mentioned architectures are based on the centralized cache scheme. For two-layer architecture, each mobile device has no connection with each other; and services cached on the mobile devices cannot be shared by the other users. For multi-layer architecture, the cache scheme has no difference from the two-layer architecture. In fact, the router caching the popular services can be regarded as a big terminal holding a group of users.

Since resources for communication are usually scheduled in the cloud computation center of the wireless converged networks, the cache scheme is a typical centralized cache scheme which can be modeled as 0-1 knapsack problem. Greedy algorithm and dynamic algorithm can be employed to solve the this problem. The classical and optimal way to solve this problem is well known as dynamic algorithm; and it can achieve the best result in the condition that the service size is an integer. Generally, the dynamic programming method has two implement ways: recursion and loop. Due to the limitation of the recursion in terms of efficiency, loop iteration is superior to it. The concrete realization of loop iteration is a two-layer circulation; and it reflects the idea from the bottom up. The computer complexity of loop iteration is $O(mC)$, wherein m is the size of services and C is the average storage size of a router or a base station.

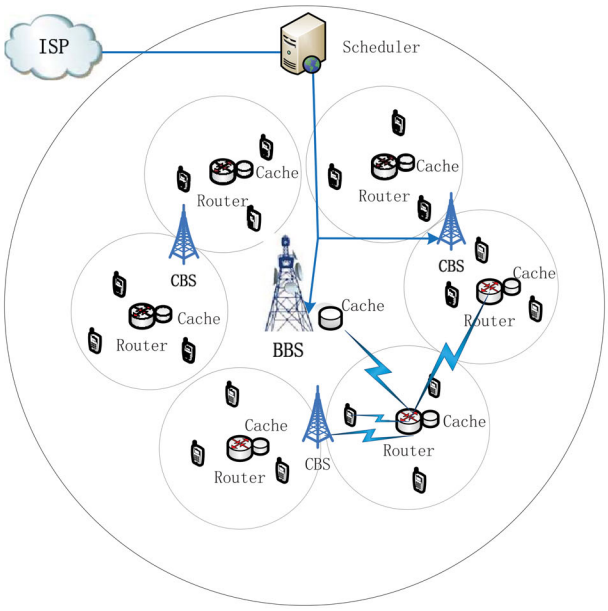


Fig. 1. The diagram of the proposed CON.

III. System Model and Performance Evaluation

The diagram and the system model of the proposed converged overlay networks (CON) are respectively illustrated in Fig. 1 and Fig. 2. There are a broadcast base station (BBS) with a content set cache, a scheduler, cellular base stations (CBS) and router-nodes with served users in the proposed networks.

A. System Model of the Proposed Scheme

The general work flow of the converged overlay networks is as followed: the scheduler detects popular services from the Internet service provider (ISP); and it delivers popular services to BBS in the form of transport stream over IP through the broadband backbone network. Meanwhile, the BBS pushes the services to the router-nodes; and the scheduler decides which services to be cached through the return channel of routers. Then, the routers provide the pre-cached services for the local users. If the requested services are not found in the cache, the requests are switched to CBS through routers.

B. Properties of Services

Researches show that Zipfian distribution characterizes statistically the popularity of services [5]. A fixed content set

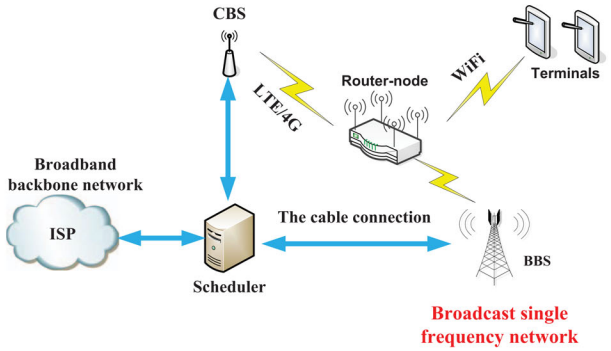


Fig. 2. The system model of CON.

relative to one scheduling period is considered in the model. The content set S is composed of M distinct services. The characters of services are defined as follows:

1) Service Popularity: The probability function of the Zipfian distribution is given by

$$f(i;z,M)=\frac{(1/i)^z}{\sum_{j=1}^{m}(1/j)^z} \tag{1}$$

where m is the number of the services and i represents the index and rank of the services. In that case, f_i is used to describe the popularity of the i-th service for simplicity.

2) Service Size: Based on the fact that only a small number of services have a large size, a long tail style distribution is adopted in [6] to model the size of the services:

$$P(X>\varepsilon)=\left(\frac{\varepsilon}{\varepsilon_{\min}}\right)^{-\beta} \tag{2}$$

where $\varepsilon_{min}>0$, $\beta>1$ and ε_i is the size of the i-th service. The average size of these services is $\bar{\varepsilon}$; and $\bar{\varepsilon}=\frac{\beta}{\beta-1}\varepsilon_{\min}$.

C. Performance Evaluation

1) *Equivalent Throughput (ET):* From the network operator perspective, the goal of the proposed scheme is to maximize offloading backhaul traffic. Then, the equivalent throughput of the proposed network in the limited scheduling time slot is

$$ET=ET_b+ET_c=\sum_{i\in S}\left|\bigcup_{j\in V}a_{ij}Q_j\right|.f_i.\varepsilon_i+B_{cs}\eta_{ce}N_c \tag{3}$$

where the total throughput equals to the sum of broadcast and cellular throughput; V is the set of the router-nodes; Q_j represents the set of requests of node j and its neighbor nodes; B_{cs} is the bandwidth of each cellular subchannel with the bandwidth efficiency η_{ce}; N_c is the subchannel number; and $\sum_{j\in V}a_{ij}$ is the router-node occupation number of the service i. In the formula, ET provided by the routing node is defined as the sum of the cache value of the routing node to the popular services, that is, the product of the size of the cached services and the number of requests for it.

2) *Repetition Coefficient (RC):* In order to get some insight of the advantage of distributed caches over centralized caches, we define the repetition coefficient (RC) to indicate cache utilization efficiency,

$$RC=\sum_{i\in S}\sum_{j\in V}a_{ij}.\frac{f_i\varepsilon_i}{\sum_{i\in S}f_i.\varepsilon_i(\sum_{j\in V}a_{ij}\geq 1)}\Bigg/n \tag{4}$$

where $\frac{f_i\varepsilon_i}{\sum_{i\in S}f_i.\varepsilon_i(\sum_{j\in V}a_{ij}\geq 1)}$ is the weight of service i.

3) *Delay Cost (DC):* As we know, the resource scheduling problem is just a trade-off between the system throughput and delay cost. Due to limited node capacity and user request tolerance time, this paper only considers that the user requests can only reach the neighbor nodes through the accessing node. Then, the DC of the proposed scheme is defined as:

$$\frac{\sum_{i\in S}\sum_{j\in V}a_{ij}|N_j|f_i.1+\sum_{i\in S}(\left|\bigcup_{j\in V}a_{ij}Q_j\right|-\sum_{i\in S}a_{ij}.|N_j|).f_i.2}{\sum_{i\in S}\left|\bigcup_{j\in V}a_{ij}Q_j\right|.f_i} \tag{5}$$

where the first part in the numerator is the delay cost of one-hop services and the second part is the delay cost of two-hop services. The denominator represents the total number of requests served by the cache services to normalize the DC. Then DC can measure the average hop number of user's one request for services, representing the average latency degree of user requests.

4) *Load Balancing Coefficient (LBC):* Load balancing first appears in the field of data offloading where traffic index is usually used to measure the load level of equipment. LBC is

TABLE I
LIST OF SYMBOLS

Parameter	Definition
T	The length of the time slot
B_b	Broadcasting bandwidth
η_{be}	Broadcasting bandwidth efficiency
B_{cs}	Cellar subchannel bandwidth
η_{ce}	Cellar subchannel bandwidth efficiency
B_{WiFi}	The maximum data rate of router-node
N_c	The number of the cellar subchannels
n	The number of the router-nodes
m	The number of the services
c_j	The cache size of router-node j
u_j	The number of users of router-node j
z	Zipf factor
β	Pareto parameter
d	The number of one user requests
f_i	The popularity of service i
ε_i	The size of service i
$\bar{\varepsilon}$	The average size of services
ε_{min}	Pareto independent variable's lower bound
q_j	The number of requests of node j and its neighbor nodes
p_{ij}	The cache potential of node j about service i
a_{ij}	The flag of whether node j caches service i
L_j	The number of requests of node j served by other nodes
C	The average storage size of nodes
k	The scale of nodes $n=k^2$
$\boldsymbol{V}$	The set of the router-nodes
$\boldsymbol{No_j}$	The set of node j and its neighbors
$\boldsymbol{Q_j}$	The set of requests of node j and its neighbor nodes
$\boldsymbol{N_j}$	The set of local requests of node j
$\boldsymbol{S}$	The set of the services
$\boldsymbol{Nb_j}$	The neighbor nodes set of the node j

defined as

$$\mathrm{D}\left(\left(\begin{array}{l}\sum\limits_{i\in S} a_{ij}\cdot\left(\left|Q_j\right|-\left|N_j\right|\right)\cdot f_i\cdot\varepsilon_i+\\ \sum\limits_{i\in S}\left(\sum\limits_{j\in No_j} a_{ij}>1\right)\cdot f_i\cdot\varepsilon_i\cdot\left|N_j\right|\end{array}\right)/B_{WiFi}T\right) \tag{6}$$

where the formula in parentheses represents the node traffic load; and the standard deviation of each node traffic load is denoted as $D(.)$ to represent LBC.

IV. Problem Formulation and Solutions

In this section, we discuss how to efficiently cache the pushed services to router-nodes. The proposed cache scheme is formulated as an optimization problem. We prove that this problem is an NP-complete problem. In that case, we propose three algorithms to get the approximate optimal solution: alternation based node-selecting and odd-even based dynamic programming algorithm (ABNS-ODP), alternation based node-selecting and dynamic programming algorithm (ABNS-DP), degree-based greedy-set-cover node-selecting and dynamic programming algorithm (DBGNS-DP). Both algorithm ABNS-ODP and algorithm ABNS-DP divide services into two sets, while algorithm DBGNS-DP divides services into multiple sets as much as possible. And DGNS-DP adds extra computational complexity to improve throughput performance and outperforms the other two algorithms.

A. Problem Formulation

Fig. 3 depicts the topology of these router-nodes. Note that there are $f_i\cdot|N_j|$ requests for the i-th service in the j-th router-node, where $|N_j|=u_j\cdot d$, and u_j means router-node j's user number. We assume each user has d requests in the scheduling time slot. Then, the total requests can be served by the j-th router-node for services is defined as $q_j=|N_j|+\sum\limits_{k\in Nb_j}|N_k|$, where Nb_j means the neighbor nodes set of the node j. The set of this q_j requests is denoted as Q_j. The set of the node j and its neighbors is denoted as $No_j=j\cup Nb_j$. In order to evaluate the cache potential of the j-th router-node about service i, we have

$$p_{ij}=a_{ij}\cdot\left|Q_j\right|\cdot f_i\cdot\varepsilon_i \tag{7}$$

where a_{ij} is the flag indicating whether the j-th router caches the i-th service or not; and S represents the set of the sevices. Due to the limited bandwidth of the router-node, the maximum capacity size of the nodes limited by the bandwidth of the node B_{WiFi} and the scheduling time slot T. Then, we have

$$\sum_{i\in S} a_{ij}\cdot\left(\left|Q_j\right|-\left|N_j\right|\right)\cdot f_i\cdot\varepsilon_i+\sum_{i\in S}\left(\sum_{j\in No_j} a_{ij}>1\right)\times f_i\cdot\varepsilon_i\cdot\left|N_j\right|\le B_{WiFi}T \tag{8}$$

where the left side represents the node's actual load, including the traffic load as request node or source node. Due to the wireless connection between router-nodes is intrinsic intermittent, there is a big delay for the services transmitted through three hops or more. Moreover, for the limitation of the nodes storage size and serving capacity, the number of users, and the number of neighbor nodes, one node can serve is limited. In that case, one-hop and two-hop scenario are only considered. We intend to maximize the ET in the scheduling time slot T:

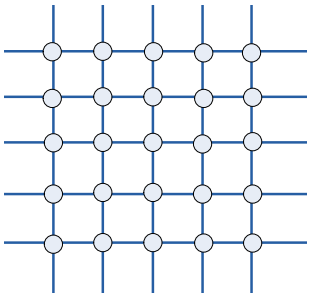


Fig. 3. The topology of router-nodes.

$$\max \quad ET=\sum_{i\in S}\left|\bigcup_{j\in V} a_{ij}Q_j\right|.f_i.\varepsilon_i+B_{cs}\eta_{ce}N_c \tag{9}$$

$$s.t. \quad \sum_{i\in S} a_{ij}\cdot\left|Q_j\right|\cdot f_i\cdot\varepsilon_i\le B_{WiFi}T \tag{10}$$

$$\sum_{i\in S} a_{ij}\varepsilon_i\le c_j \tag{11}$$

$$a_{ij}\in\{0,\,1\} \tag{12}$$

where c_j is the storage capacity of router-node j; and V stands for the set of router-nodes.

Equation (9) represents the total throughput of the CON where the cellular part can be regarded as a constant. Constraint (10) shows that the router-node delivery capacity is restricted by its bandwidth in the scheduling time slot. We can increase node density to make constraint (10) to be satisfied because the uniform distribution scenario is considered. Constraint (11) prevents the services cached to each selected router-nodes from exceeding the storage size of the nodes. Constraint (12) indicates this optimal problem is a 0-1 programming problem. To get further insight, it is obvious that the problem is similar to the set cover problem (SCP). We use the reduction method in [27] to prove that the proposed problem pertains to NPC.

Theorem 1: The cache allocation problem of CON is an NP-complete problem.

Proof: The formal language of decision problem corresponding to the proposed problem (TPP) is: $TPP=\{\{a_{ij}\},\{f_i.\varepsilon_i\},\{Q_j\},\langle k\rangle : a_{ij}\in\{0,1\}$, select an array of a_{ij} to make the value of ET at least $k\}$. Firstly, the proposed problem is proved to be NP. Given an instance of the TPP, ET can be calculated to judge whether the value exceeds k in polynomial time. As a result, TPP is an NP problem. ■

In order to prove that TPP pertains to NPH, we first prove $SCP\le_p TPP$, where $\le_p$ means a problem can be transformed into another problem in polynomial time. Assuming that $\{\{a_{1j}\},Q_j.f_1.\varepsilon_1\}$ is an instance of SCP, we construct an instance of TPP where the number of contents is two. In that case, if the instance of SCP selects a minimal number of sets to cover all elements of sets, the instance of TPP uses minimal nodes to cache content 1 and uses the remaining maximal

nodes to cache service 2 to get a maximal value k, and vice versa. In conclusion, SCP has a minimal set if and only if TPP has a maximal value of ET exceeding k. Finally, it can be proved that TPP is NP-complete by the definition of NPC in [27].

This optimal cache allocation problem can be solved by the exhaustive search algorithm. However, the complexity exponentially increases with the nodes scaling up. We propose two approximation algorithms and a heuristic algorithm; and we use node-selecting algorithm to choose candidates of router-nodes and use dynamic programming algorithm to schedule the services in these algorithms.

B. ABNS-ODP Algorithm

Definition: If a user can access the same services kept in two or more different router-nodes, we call the the requests of the user are overlapped; and the router-nodes corresponding to the overlapped requests are also overlapped. A simple case with 3×3 router-nodes is just shown in Fig. 4. For example, when we choose node 1 and 2 to cache service i, the overlapped requests are $N_1 + N_2$; when we choose node 1 and 3, the overlapped requests are N_2 ; when we choose node 1 and 6, the overlapped requests are zero.

Problem Transformation: Noted that the overlapped area of each node is relevant to the selecting matrix $\{a_{ij}\}$. The overlapped requests of node j is defined as L_j. So, we use L_j to represents the set union. Then, the proposed problem can be transformed into:

$$\max \quad ET = \sum_{j \in V} \sum_{i \in S} a_{ij} \cdot \left(q_j - L_j\right) \cdot \varepsilon_i \cdot f_i + B_{cs} \eta_{ce} N_c \quad (13)$$

$$s.t. \quad \sum_{i \in S} a_{ij} \varepsilon_i \leq c_j \quad (14)$$

$$a_{ij} \in \{0,\ 1\}. \quad (15)$$

Theorem 2: L_j is relevant to the selecting matrix $\{a_{ij}\}$.

Proof: Gotten from inclusion-exclusion principle, we have:

$$\left| \bigcup_{j=1}^{n} a_{ij} Q_j \right| = \left| \begin{array}{l} (a_{i1} Q_1 \cup \cdots \cup a_{i(j-1)} Q_{j-1} \cup \\ a_{i(j+1)} Q_{j+1} \cup \cdots \cup a_{in} Q_n) \cup a_{ij} Q_j \end{array} \right|$$

$$= \left| a_{ij} Q_j \right| + |(\ldots)| - \left| a_{ij} Q_j \cap (\ldots) \right| \quad (16)$$

It is obvious that L_j is relevant to the selecting matrix $\{a_{ij}\}$ in the above equations. Consequently, if the overlapped requests can be reduced, more utility can be obtained. Based on this fact, the nodes are selected based on alternation which not only guarantees smaller overlapped requests but also ensures the coverage of the whole requests. Therefore, alternation-based node-selecting and odd-even based dynamic programming algorithm is proposed, in which alternation based node-selecting determines the nodes to cache and odd-even based dynamic programming algorithm is responsible for allocating services to each selected nodes.

The ABNS-ODP algorithm is shown in Alg. 1. We assume that the services are sorted in descend order according to their popularity. The step 2 divides the content set into two parts: odd index part and even index part. In that case, the candidate services parts of odd indexes and the even indexes have the similar numbers and the similar popularity. Codes in line $3 \sim 15$ are used to select nodes. The odd part of services are cached to the corresponding nodes by step $5 \sim 9$, otherwise by step $10 \sim 15$. After the nodes selecting, we use dynamic programming by step $16 \sim 20$ to allocate the services to each node. The time complexity of the proposed algorithm is $O(mC + n)$.

Algorithm 1 Alternation-Based Node-Selecting and Odd-Even Based Dynamic Programming Algorithm

Input: Input parameters $\mathbf{V}$, f_i, ε_i, c_j, $\mathbf{S}$, N_j **Output:** $\mathbf{A} = \left(a_{ij}\right)_{m \times n}$

1: Initialization: $\mathbf{A} = 0$, $|\mathbf{V}| = n = k \cdot k$, $\mathbf{A}_1 = \left(a_{ixy}\right)_{m \times k \times k} = 0$
2: Divide $\mathbf{S}$ into $\mathbf{S}_{even}$ and $\mathbf{S}_{odd}$
3: **for** each node $j \in \mathbf{V}$ **do**
4: Transform j into coordinates on XY plane: $j = (x-1) \cdot k + y$
5: **if** $x \leq \lfloor k/2 \rfloor$, $x + y$ is even **then**
6: $a_{ixy} = 1$, $\forall i \in \mathbf{S}_{odd}$
7: **else**$x > \lfloor k/2 \rfloor$
8: $a_{ixy} = a_{i(k-x)y}$, $\forall i \in \mathbf{S}_{odd}$
9: **end if**
10: **if** $x \leq \lfloor k/2 \rfloor$, $x + y$ is odd **then**
11: $a_{ixy} = 1$, $\forall i \in \mathbf{S}_{even}$
12: **else**$x > \lfloor k/2 \rfloor$
13: $a_{ixy} = a_{i(k-x)y}$, $\forall i \in \mathbf{S}_{even}$
14: **end if**
15: **end for**
16: **for** $\mathbf{S} \in (\mathbf{S}_{even}, \mathbf{S}_{odd})$ **do**
17: Use dynamic programming algorithm of the knapsack problem to get the service selecting factor: $\mathbf{x}$
18: $\mathbf{X} = zeros(1, m)$
19: Calculate $\mathbf{X}$ according to $\mathbf{x}$
20: **end for**
21: **for** each node $j \in \mathbf{V}$ **do**
22: $a_{ixy} = a_{ixy} \circ \mathbf{X}$
23: **end for**
24: Return $\mathbf{A}$

C. ABNS-DP Algorithm

This algorithm is similar to ABNS-ODP algorithm just as Alg. 2 describes. However, algorithm ABNS-DP does not consider the problem of load balance and only considers how to maximize system throughput. Algorithm ABNS-DP divides the service sets according to the order of service popularity, while algorithm ABNS-ODP divides service sets according to the order of parity. The former can ensure that the high popularity rank services are cached where its ET performance is better, while the latter has better LDC performance. Dynamic programming is also used to allocate the services to each node. The time complexity of the proposed algorithm is $O(mC+n)$.

Fig. 4 is an example of Alg. 1 and Alg. 2 which have the same nodes division.

Algorithm 2 Alternation-Based Node-Selecting and Dynamic Programming Algorithm

Input: Input parameters $\mathbf{V}$, f_i, ε_i, c_j, $\mathbf{S}$, N_j **Output:** $\mathbf{A} = (a_{ij})_{m\times n}$

Initialization: $\mathbf{A} = 0$, $|\mathbf{V}| = n = k\cdot k$, $\mathbf{A}_1 = (a_{ixy})_{m\times k\times k} = 0$
Repeat Alg. 1 steps (3) − (15)
$\mathbf{B} = zeros(|\mathbf{S}|+1, C+1)$, $W_i = \varepsilon_i \cdot f_i$
for $i = |\mathbf{S}|$ **do**
  for $j = (1{:}C)$ **do**
    if $\varepsilon_i > j$ **then**
      $\mathbf{B}(i+1, j+1) = \mathbf{B}(i, j+1)$
    else
      $\mathbf{B}(i+1, j+1) =$ $\max\{\mathbf{B}(i, j+1), W_i + \mathbf{B}(i, j-\varepsilon_i+1)\}$
    end if
  end for
end for
$\mathbf{x} = zeros(1, m)$, $J = C+1$
for $i = m{:}-1{:}1$ **do**
  if $\mathbf{B}(i+1, J) > \mathbf{B}(i, J)$ **then**
    $\mathbf{x}_i = 1$
    $J = J - \varepsilon_i$
  else
    $\mathbf{x}_i = 0$
  end if
end for
$\mathbf{S} = \mathbf{S}-$ selected services: the remaining services set
Repeat step (16)-(34) with some parameters adaptive to the $\mathbf{S}$
Calculate the final $\mathbf{x}$
for each node $j \in \mathbf{V}$ **do**
  $a_{ixy} = a_{ixy} \circ \mathbf{x}$
end for
Return $\mathbf{A}$

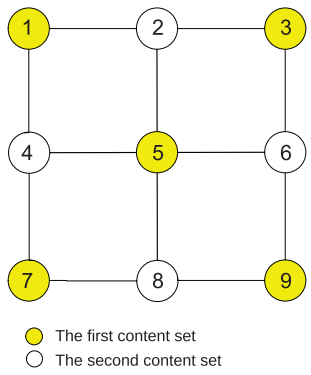


Fig. 4. An example of 3×3 node-selecting algorithm.

D. DBGNS-DP Algorithm

Problem Transformation: As mentioned above, the proposed problem is similar to SCP. To get further insight, Q_j contains the requests of node j and its neighbors that can be served. Note that $No = j + Nb_j$ and the local requests of each node is uniform; then, $|\sum_{j\in V} a_{ij}Q_j|$ can be substituted for $|\sum_{j\in V} a_{ij}No_j|.u_j.d$. Then, the proposed problem is transformed into:

$$\max \quad ET = \sum_{i\in S}\left|\bigcup_{j\in V} a_{ij}No_j\right|.u_j.d.f_i.\varepsilon_i + B_{cs}\eta_{ce}N_c \quad (17)$$

$$s.t. \quad \sum_{i\in S} a_{ij}\varepsilon_i \le c_j \quad (18)$$

$$a_{ij} \in \{0,\ 1\} \quad (19)$$

It can be observed that the fewer nodes are used to cache the most popular services; and the higher cache utilization can be obtained. By parity of reasoning, the fewer nodes are selected to cache the less popular services, the higher cache utilization can be achieved for the rest services and the remaining nodes.

The proposed algorithm can be divided into two parts: node selecting algorithm and content allocating algorithm. For the node selecting part, how to select fewer nodes to cover the whole area can be modeled as a set cover problem. Since the SCP is also NP-complete, a degree-based greedy-set-cover node-selecting algorithm is proposed. In this proposed node selecting algorithm, degree means the maximal number of the neighbor nodes that correspond to No_j. For the selected nodes, dynamic programming algorithm is responsible for allocating services to each node. In each iteration, a degree-based greedy-set-cover node-selecting algorithm firstly selects the fewest nodes from the current nodes; and then, dynamic programming algorithm is adopted to select proper services to be cached until the nodes or services are used up.

The details are illustrated in algorithm 3. Steps $4 \sim 9$ are used to select the minimal number of nodes by using the greedy-set-cover algorithm. Then, steps $10 \sim 29$ of the dynamic programming algorithm allocate the services to the selected nodes in advance. The steps $4 \sim 29$ is treated as one iteration. The loop body of the first 'While' is the iterations. Steps $30 \sim 33$ are the iteration control statements. The time complexity of the proposed algorithm is $O(n^3 + mC)$.

V. Simulation and Discussions

In this section, the numerical results of the proposed three algorithms are presented. The default parameter settings are shown in Table II which are referred from the industrial standards and convention.

A. Throughput Improvement of the Proposed Scheme

As shown in Fig. 5, the network throughput is improved with the increasing of Zipf factor. However, the growth rate becomes smaller with the Zipf factor increasing; and the network throughput converges on a constant. This is because the upper bound is mainly determined by the most popular services; and with increment of Zipf factor, the dominant factor is converged on the most popular services regardless of other services. Compared the three proposed distributed-caching algorithms, it is found that the distributed cache schemes have big throughput improvements when z is smaller than 1.5. Among these proposed algorithms, Alg.2 performs a

Algorithm 3 Degree-Based Greedy-Set-Cover Node-Selecting and Dynamic Programming Algorithm

Input: Input parameters $\mathbf{V}$, f_i, ε_i, c_j, $\mathbf{S}$, No_j **Output:** $\mathbf{A} = (a_{ij})_{m\times n}$

Initialization: $\mathbf{A} = \{a_{ij}\}_{m\times n} = 0$, $|\mathbf{V}| = n = k \cdot k$
$Se = []$, $No = No_j$, $U = U_1 = 1{:}n$, $Num = m$, $BI = 1{:}100$
while U$_1$! =$\emptyset$ or $Num! = 0$ **do**
  while U! =$\emptyset$ **do**
    Select an $No_j \in No$ that maximizes $\left|U \cap No_j\right|$
    $U = U - No_j$
    $Se = Se \cup \{No_j\}$
    Substitute the selected No_j in No for $\emptyset$
  end while
  $\mathbf{x} = zeros(1, Num)$, $J = C + 1$
  $\mathbf{A} = zeros(Num + 1, C + 1)$, $W_i = \varepsilon_i \cdot f_i$,
  for $i = |\mathbf{S}|$ **do**
    for $j = (1{:}C)$ **do**
      if $\varepsilon_{BI(i)} > j$ **then**
        $\mathbf{A}(i+1, j+1) = \mathbf{A}(i, j+1)$
      else
        $\mathbf{A}(i+1, j+1) = \max\{\mathbf{A}(i, j+1), W_{BI(i)} + \mathbf{A}(i, j - \varepsilon_{BI(i)} + 1)\}$
      end if
    end for
  end for
  for $i = Num{:} - 1{:}1$ **do**
    if $\mathbf{A}(i+1, J) > \mathbf{A}(i, J)$ **then**
      $\mathbf{x}_i = 1$
      $J = J - \varepsilon_{BI(i)}$
    else
      $\mathbf{x}_i = 0$
    end if
  end for
  $Num = Num - sum(\mathbf{x})$
  Calculate BI_1 through finding position of $\mathbf{x} == 0$
  $BI = BI(BI_1(:))$
  Calculate $U_1 = \bigcup_{j\in V} No_j$
end while

little better than Alg.1; and the ET curves of them are almost overlapped. Mover, Alg.3 is optimal among these proposed algorithms in terms of the ET.

Fig. 6 shows impact of router scales on the network throughput. It is obvious that with the scale of router nodes becoming bigger, equivalent throughput has a linear increment. This phenomenon can be explained that the object function has a linear correlation with the nodes number. In fact, the scale of router nodes can't be infinite since the number of nodes that a broadcast station can cover is finite.

B. The Delay Cost of the Proposed Scheme

The impact of Zipf factor on the normalized delay and the comparison of three proposed algorithm with DP in centralized cache are illustrated in Fig. 7. Obviously, Alg.1 and Alg.2 have little difference in the normalized delay. It proves

TABLE II
SIMULATION PARAMETERS

Parameter	Values
T	$1800s$
B_b	$3MHz$
η_{be}	$1bps/Hz$
B_{cs}	$1MHz$
η_{ce}	$1bps/Hz$
N_c	20
n	$(9*9, 10*10, 11*11)$
m	100
c_j	$1GB$
u_j	10
z	$(0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2)$
β	2
d	10
B_{WiFi}	$37.5MBps$

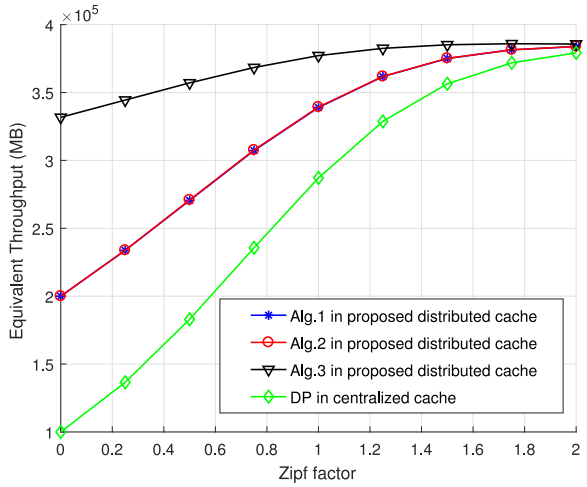


Fig. 5. The impact comparison of Zipf factor on the network throughput.

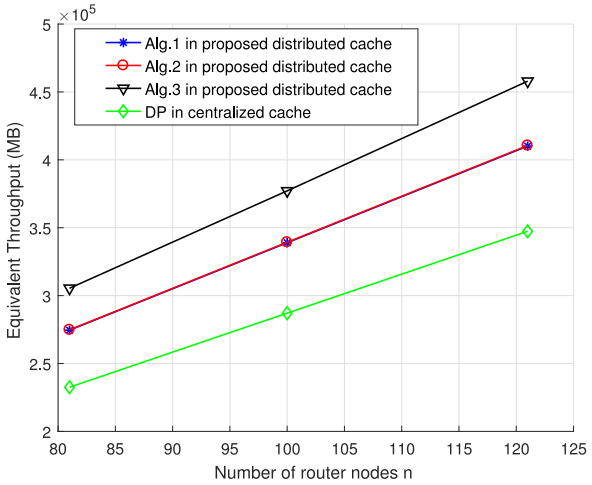


Fig. 6. The impact comparison of the router nodes scale on the network throughput.

that they have similar performances. When the Zipf factor is below 1, Alg.3 behaves much better than the other two algorithms. While the Zipf factor surpasses 1, the situation is reversed. The reason is that the equivalent throughput gain becomes smaller with the increasing of Zipf factor. Generally,

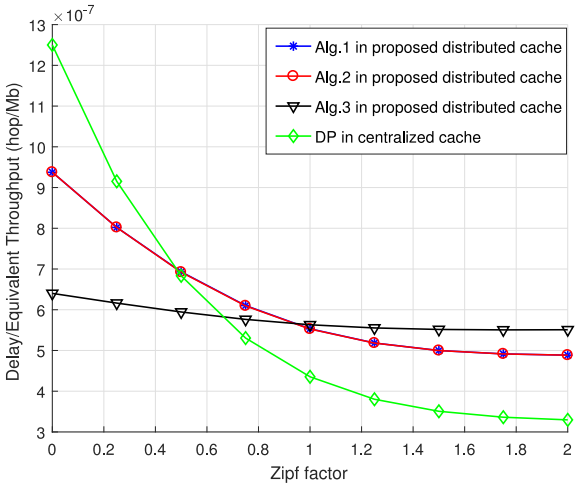


Fig. 7. The impact comparison of Zipf factor on the normalized delay.

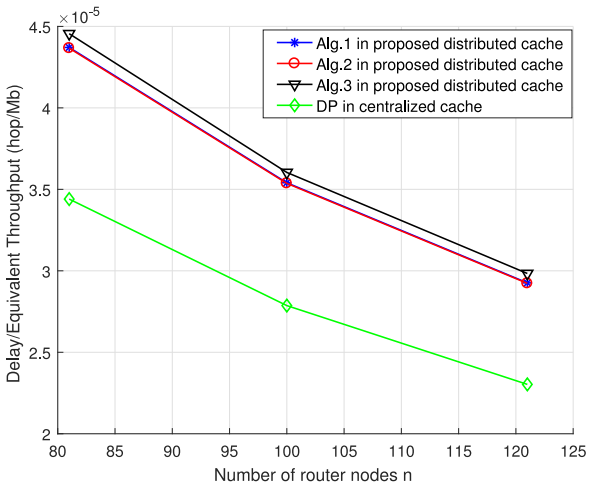


Fig. 8. The impact comparison of the scale of router nodes on the normalized delay.

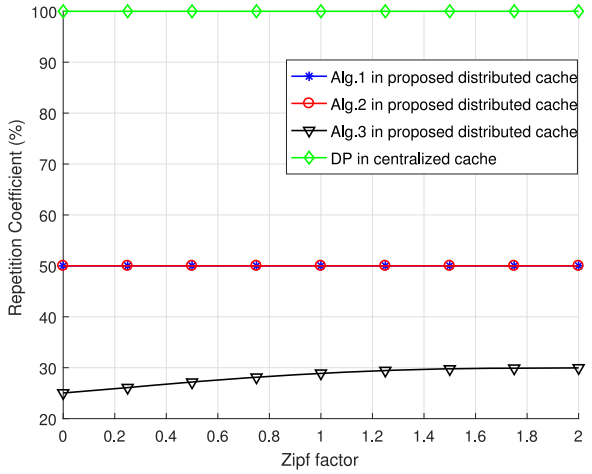


Fig. 9. The impact comparison of Zipf factor on the repetition coefficient.

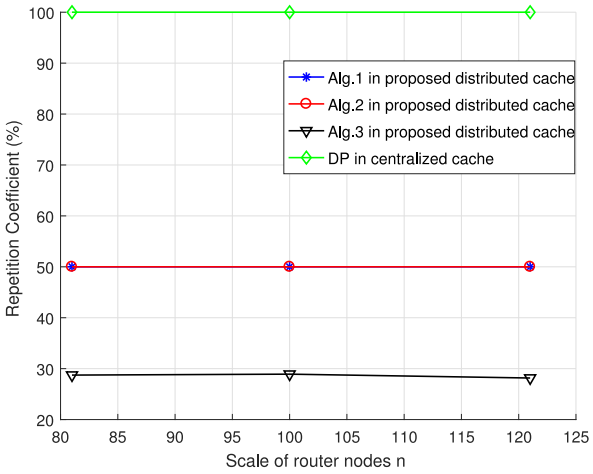


Fig. 10. The impact comparison of Zipf factor on the repetition coefficient.

TABLE III
SIMULATION PLATFORM

Parameter	Values
CPU	$i5-6400$
Frequency of CPU	$2.7GHz$
Coreness of CPU	4
Size of RAM	$8GB$
Version num. of simulation platform	*MATLAB R2016b 64-bit(win64)*

the proposed distributed cache schemes outperform the centralized cache schemes when Zipf factor is between 0 to 0.6. In other words, the more domain factor focuses on few most popular services, the little effect resource scheduling has. As shown in Fig. 8, the normalized delay decrease linearly with the scale increasing of the router nodes.

C. The Repetition Coefficient of the Proposed Scheme

We can find out that the RCs of distributed-cache schemes are smaller than that of the centralized-cache scheme from Fig. 9. This means the proposed distributed scheme have higher cache utility. For Alg.1 and Alg.2, the RCs are only 50 percents of that of the centralized cache scheme. Moreover, the RC is constant. The reason is that the nodes only cache half of the services by using the proposed algorithms. For Alg.3, the RC is reduced to about 30 percents. Furthermore, the the RC of Alg.3 has a little increment with the increasing of Zipf factor since the RC highly depends on a few most popular services.

In terms of the scale of router nodes, Fig. 10 demonstrates that the RC is nearly constant when the scale of router nodes becomes bigger. For Alg.1 and Alg.2, it can be derived that the RC is constant. That is because both Alg.1 and Alg.2 divide the whole nodes into two sets while Alg.3 divide the whole nodes into as many sets as possible. For Alg.3, the RC is about 30 percents; and it demonstrates that the proposed Alg.3 can achieve stable cache utilization and it is quite superior to others two schemes.

D. The Convergence Time of the Proposed Scheme

The complexities of the proposed three algorithms are also compared; and the convergence time is adopted to evaluate the performances. The more complexity an algorithm has, the more computing resources the computer has to allocate. The parameters of the simulation platform are shown in Table III.

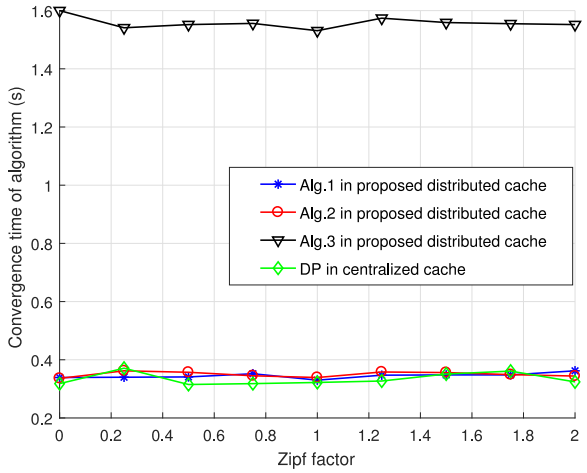


Fig. 11. The impact comparison of Zipf factor on the convergence time.

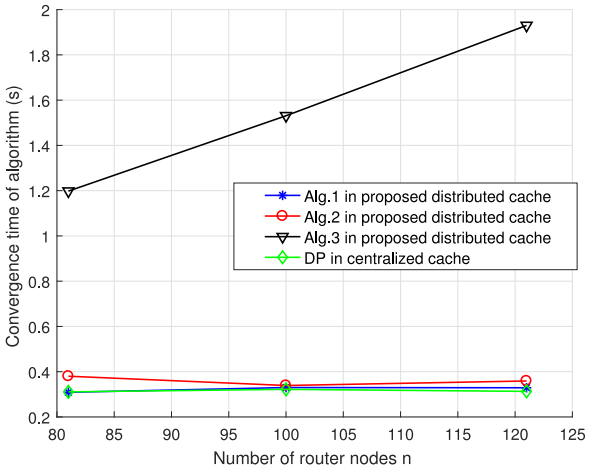


Fig. 12. The impact comparison of the scale of router nodes on the convergence time.

As shown in Fig. 11, the convergence time curves of the proposed algorithms are almost stable regardless of Zipfian factor's variation. Moreover, the convergence time of Alg.3 is much longer than those of the other two algorithms which contributes to much better performance than the other proposed algorithms. However, the complexity of Alg.3 is not so high since it is a greedy and dynamic programming algorithm. In terms of the scale of router nodes, for Alg.1 and Alg.2, the convergence time is almost stable in that both algorithms have the same complexity. Alg.3 has a linear increment of convergence time. The reasonable explanation is that the complexity of the Alg.3 is the polynomial of the three order square of n. While the complexity of the other algorithms is only the polynomial of the one order square of n. This reveals that the scale of router nodes affects Alg.3 much more than Alg.1 and Alg.2.

E. The Load Balancing Coefficient of the Proposed Scheme

As shown in Fig. 13, the LBCs increase along with the rising of the Zipf factor. The reason is that the popularity of services becomes increasingly uneven as the Zipf factor increases; and it concentrates in a few services which leads to the unbalanced

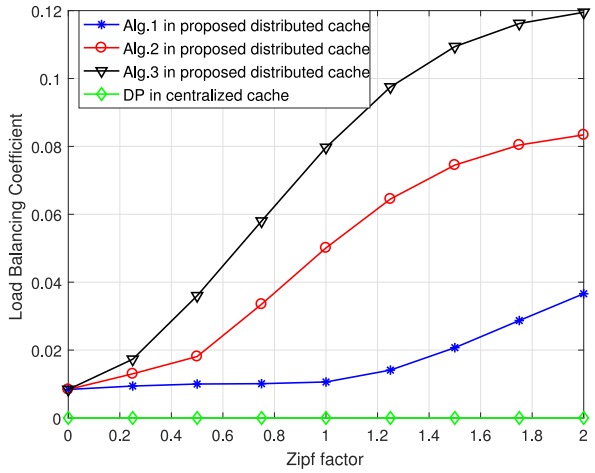


Fig. 13. The impact comparison of Zipf factor on the load balancing coefficient.

load between node groups. Although the throughout of ABNS-ODP is slightly lower than the algorithm ABNS-DP, its LBC performance is obviously better than that of ABNS-DP. The reason is that algorithm ABNS-ODP divide services by parity which can guarantee less nodes load difference than that of algorithm ABNS-DP. For the centralized cache algorithm DP, since the load of each node is exactly the same, the load balance coefficient of the system is 0. From above analysis, we can conclude that the performance improvement of algorithm DBGNS-DP is at the cost of the load balance. In the actual system, it is necessary to make a trade-off between ET, load balance, complexity, and so on.

VI. Conclusion

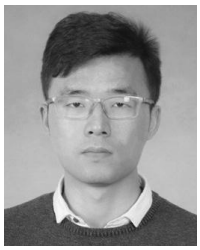
Due to the storage limitation of routers and user devices, this paper proposes a novel popular service pushing and caching scheme by using converged overlay networks. In the scheme, the deploying routers have the grid topology; and all router-nodes only connect with their neighbor nodes. The services scheduling problem is formulated as an optimal problem, which has been proved to be NP-complete. Then, the three algorithms: ABNS-ODP algorithm, ABNS-DP algorithm and DBGNS-DP algorithm are proposed to select routers and specifically cache services. Simulation results show that the proposed distributed cache scheme outperforms the traditional centralized cache scheme in terms of throughput at the cost of delay.

Further works will focus on the non-uniform case of user distribution. Moreover, the load balancing problem will also be discussed.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update 2016–2021," San Jose, CA, USA, Cisco, White Paper, Mar. 2017.
[2] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
[3] Q. Zhang, L. Gui, F. Tian, and F. Sun, "A caching-based incentive mechanism for cooperative data offloading," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2017, pp. 1376–1381.

[4] H. Ko, J. Lee, and S. Pack, "Performance optimization of delayed WiFi offloading in heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9436–9447, Oct. 2017.
[5] X. Li, J. Xiong, B. Liu, L. Gui, and M. Qiu, "A capacity improving and energy saving scheduling scheme in push-based converged wireless broadcasting and cellular networks," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2016, pp. 1–6.
[6] L. A. Adamic. (2002). *ZiPF, Power-Laws, and Pareto—A Ranking Tutorial*. [Online]. Available: http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html
[7] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[8] *Introduction of the Multimedia Broadcast Multicast Service (MBMS) in the Radio Access Network (RAN); Stage 2 (Release 7), V7.7.0*, document TS 25.346, 3GPP, Sophia Antipolis, France, Mar. 2008.
[9] G. Araniti, P. Scopelliti, G.-M. Muntean, and A. Iera, "A hybrid unicast-multicast network selection for video deliveries in dense heterogeneous network environments," *IEEE Trans. Broadcast.*, vol. 65, no. 1, pp. 83–93, Mar. 2019. doi: 10.1109/TBC.2018.2822873.
[10] H. Bawab, P. Mary, J.-F. Hélard, Y. Nasser, and O. Bazzi, "Spectral overlap optimization for DVB-T2 and LTE coexistence," *IEEE Trans. Broadcast.*, vol. 64, no. 1, pp. 70–84, Mar. 2018.
[11] L. Christodoulou, O. Abdul-Hameed, and A. M. Kondoz, "Toward an LTE hybrid unicast broadcast content delivery framework," *IEEE Trans. Broadcast.*, vol. 63, no. 4, pp. 656–672, Dec. 2017.
[12] Y. Wang *et al.*, "Media transmission by cooperation of cellular network and broadcasting network," *IEEE Trans. Broadcast.*, vol. 63, no. 3, pp. 571–576, Sep. 2017.
[13] P. Song, J. Xiong, L. Gui, M. Qiu, and Y. Zhang, "Resource scheduling for hybrid broadcasting and cellular networks," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, Jun. 2015, pp. 1–6.
[14] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.
[15] B. Xia *et al.*, "Opportunistic channel sharing in stochastic networks with dynamic traffic," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9587–9591, Oct. 2017.
[16] C. Xu, P. Wang, C. Xiong, X. Wei, and G. M. Muntean, "Pipeline network coding-based multipath data transfer in heterogeneous wireless networks," *IEEE Trans. Broadcast.*, vol. 63, no. 2, pp. 376–390, Jun. 2017.
[17] B. Li *et al.*, "Cache-based popular services pushing on high-speed train by using converged broadcasting and cellular networks," *IEEE Trans. Broadcast.*, to be published. doi: 10.1109/TBC.2018.2863102.
[18] Y. Yang, Y. X. Peng, G. Feng, and L. Y. Tian, "Application of distributed storage technology for financial management and control system in electric power system," in *Proc. 16th Int. Symp. Distrib. Comput. Appl. Bus. Eng. Sci. (DCABES)*, Oct. 2017, pp. 124–126.
[19] J. M. James and M. T. Themalil, "Design of heterogeneous wireless mesh network for LTE," in *Proc. Int. Conf. Circuit Power Comput. Technol. (ICCPCT)*, Apr. 2017, pp. 1–5.
[20] F. Rezaei and B. H. Khalaj, "Stability, rate, and delay analysis of single bottleneck caching networks," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 300–313, Jan. 2016.
[21] Z. Chen, Y. Liu, B. Zhou, and M. Tao, "Caching incentive design in wireless D2D networks: A Stackelberg game approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.
[22] C. Yang *et al.*, "Interference cancelation at receivers in cache-enabled wireless networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 842–846, Jan. 2018.
[23] S. Gao and M. Tao, "Joint multicast scheduling and user association for dash-based video streaming over heterogeneous cellular networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jul. 2016, pp. 1–6.
[24] J. Bukhari and W. Yoon, "Multicasting in next-generation software-defined heterogeneous wireless networks," *IEEE Trans. Broadcast.*, vol. 64, no. 4, pp. 915–921, Dec. 2018.
[25] W. Zhang *et al.*, "On popular services pushing and distributed caching in converged overlay networks," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2018, pp. 1–6.
[26] H. Feng, Z. Chen, and H. Liu, "Performance analysis of push-based converged networks with limited storage," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8154–8168, Dec. 2016.
[27] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. Cambridge, MA, USA: MIT Press, 2001.



Wei Zhang received the B.S. degree from Nanjing University, Nanjing, China, in 2016. He is currently pursuing the master's degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include converged networks resource scheduling and distributed caching.

Jian Xiong received the B.Sc. and M.Sc. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2002, respectively, and the Ph.D. degree from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2006. He was a Visiting Scholar with Columbia University in 2015. He is currently an Associate Professor with the Image Communication and Networking Engineering Institute, SJTU. He has published over 40 journal or conference papers and holds over 40 patents, including 25 awarded patents. His current research interests include wireless wideband transmission technologies, networking, and caching technologies of converged wideband and broadcast systems. He was a recipient of one Best Journal Paper Award (IEEE TB'14) and two Conference Best Paper Awards (IEEE/BMSB16, IEEE/BMSB18, and IEEE/SSC16). He is the TPC Co-Chair of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting from 2010 to 2018.

Lin Gui (M'08) received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2002. Since 2002, she has been with the Institute of Wireless Communication Technology, Shanghai Jiao Tong University, Shanghai, China, where she is currently a Professor. Her current research interests include HDTV and wireless communications.

Bo Liu received the B.Sc. degree from the Department of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004 and the M.Sc. and Ph.D. degrees from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2007 and 2010, respectively, where he has been an Assistant Professor with the Institute of Wireless Communication Technology since 2010. He was a Post-Doctoral Research Fellow with Deakin University, Australia, from 2014 to 2017. He is currently a Lecturer with the Department of Engineering, La Trobe University, Australia. His research interests include wireless communications and networking, and security and privacy issues in wireless networks.

Meikang Qiu (SM'07) received the B.E. and M.E. degrees from Shanghai Jiao Tong University and the Ph.D. degree in computer science from the University of Texas at Dallas. He is currently a Faculty Member with Shenzhen University and Columbia University. A lot of novel results have been produced and most of them have already been reported to research community through high-quality journal and conference papers. He has published 4 books, 400 peer-reviewed journal and conference papers, including over 200 journal articles, over 200 conference papers, and over 70 IEEE/ACM TRANSACTIONS papers. His research is supported by the U.S. Government, such as NSF, NSA, Air Force, Navy, and companies, such as GE, Nokia, TCL, and Cavium. His paper published in the IEEE TRANSACTIONS ON COMPUTERS about privacy protection for smart phones has been selected as a Highly Cited Paper in 2017. His paper about embedded system security published in the *Journal of Computer and System Science* (Elsevier) have been recognized as Highly Cited Papers in 2016 and 2017. His paper about data allocation for hybrid memory has been published in the IEEE TRANSACTIONS ON COMPUTERS has been selected as hot paper (1 in 1000 papers) in 2017. His research interests include cyber security, big data analysis, cloud computing, smarting computing, intelligent data, and embedded systems. He was a recipient of the IEEE SYSTEM JOURNAL 2018 Best Paper Award for his paper on tele-health system, the *ACM Transactions on Design Automation of Electrical Systems* 2011 Best Paper Award, the Navy Summer Faculty Award in 2012, the Air Force Summer Faculty Award in 2009, and over ten conference best paper awards in recent years. He is currently an Associate Editor of over ten international journals, including the IEEE TRANSACTIONS ON COMPUTERS and the IEEE TRANSACTIONS ON CLOUD COMPUTING. He has served as a Leading Guest Editor for the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, special issue on Social Network Security. He is the General Chair/Program Chair of a dozen of IEEE/ACM international conferences, such as IEEE TrustCom, IEEE BigDataSecurity, IEEE CSCloud, and IEEE HPCC. He is the Chair of IEEE Smart Computing Technical Committee. He is a Senior Member of ACM.

Zhiping Shi received the master's and Ph.D. degrees from Southwest Jiaotong University, Chengdu, China, in 1998 and 2005, respectively. She has two years of Post-Doctoral experience with the University of Electronic Science and Technology of China (UESTC) from 2005 to 2007. From 2009 to 2010, she was a Visiting Scholar with Lehigh University, Bethlehem, PA, USA. In 2007, she joined the School of Communication and Information, UESTC, where she is currently a Professor with the National Key Laboratory of Communications. Her research interests fall in the areas of wireless communication and coding theory.