Copyright ⓒ 2018 KSII

A Novel Service Migration Method Based on Content Caching and Network Condition Awareness in Ultra-Dense Networks

Chenjun Zhou[1], Xiaorong Zhu[1,*], Hongbo Zhu[1] and Su Zhao[1,2]
[1]Wireless Communication Key Lab of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing, China
[2]National Mobile Communications Research Laboratory, Southeast University, Nanjing, China
[*corresponding author:xrzhu@njupt.edu.cn]

Received July 18, 2017; revised October 11, 2017; accepted January 4, 2018; published June 30, 2018

Abstract

The collaborative content caching system is an effective solution developed in recent years to reduce transmission delay and network traffic. In order to decrease the service end-to-end transmission delay for future 5G ultra-dense networks (UDN), this paper proposes a novel service migration method that can guarantee the continuity of service and simultaneously reduce the traffic flow in the network. In this paper, we propose a service migration optimization model that minimizes the cumulative transmission delay within the constraints of quality of service (QoS) guarantee and network condition. Subsequently, we propose an improved firefly algorithm to solve this optimization problem. Simulation results show that compared to traditional collaborative content caching schemes, the proposed algorithm can significantly decrease transmission delay and network traffic flow.

Keywords: Content caching, Service migration, Ultra-dense networks, Firefly algorithm

1. Introduction

In recent years, with the ubiquitous adoption of mobile Internet across the globe and the rapid rise in the number of mobile users, mobile data traffic has witnessed a remarkable growth. Mary Meeker from Kleiner Perkins Caufield & Byers (KPCB) predicts that by the end of 2017, the figure is likely to reach 50% or more [1]. In future communication systems, data services will primarily be distributed in hotspots and indoors. In order to meet the needs of a large volume of data communication, small base stations (BSs) should be extensively deployed as hotspots [2]. Current research on 5G communication systems demonstrates that network architecture growth is now flattening. Additionally, the core network is sinking where service is closer to the users.

To solve the problem of capacity bottlenecks and users' service experience in the mobile communication network, some hot content is distributed in small BSs of the ultra-dense network, which ensures cooperative optimization of communication and calculation using storage resources to reduce the pressure of the backhaul and core networks. It also reduces end-to-end latency and network traffic, improving communication network performance. In [3], in order to solve the cooperative video caching problem in a Long Term Evolution (LTE) core network, the authors establish an optimization model to maximize the data volume of video hits and propose a cooperative traffic routing algorithm with a cooperative video rate. [4] studies the development trend of 5G mobile communication and points out that the content distribution network is one of the key technologies. In [5], the authors propose an optimal content delivery method that uses a multicast algorithm when hot content is cached in routers. The authors in [6] propose a content caching network composed of mobile terminal equipment and develop an optimization solution to determine the probability of minimal average caching failure rate when each terminal caches the content. [7] presents a content caching design method and establishes an optimal content delivery framework in the mobile device. Focusing on the resource optimization mechanism under the distributed mobility management framework, [8] proposes a content deployment method for various mobile internet network requirements and scenarios in the future. [9] proposes a combination of cache hit and historical cache hit rate content selection strategy to coordinate the determination of the popularity of content. Ultra dense cloud small cell network (UDCSNet), which combines cloud computing and massive deployment of small cells, is a promising technology for 5G LTE-U mobile communications because it can accommodate the anticipated explosive growth of mobile users' data traffic. [10] presents an overview of the challenges and requirements of the fronthaul technology in 5G LTE-U UDCSNets and surveys the advantages and challenges for various candidate fronthaul technologies. In [11], the authors propose a user association and power allocation algorithm based on the combination of Millimeter wave (mmWave) and ultra dense network (UDN).

In this paper, on the basis of our knowledge of content caching distribution and the existing network conditions, we propose a novel service migration method (SMM), which can minimize the cumulative transmission delay of the users' access to the content. We propose to use the improved firefly algorithm to find the solutions to minimize delay. Simulation results show that compared to traditional collaborative content caching schemes, the proposed algorithm can significantly decrease transmission delay and network traffic flow.

2. System Model

In this paper, we assume that the hotspot is covered by an ultra-dense network, which consists of a control base station (BS), many small BSs, and a local network controller. The control BS with a large coverage is mainly used for transmitting control signals and the small BSs with small coverage are mainly used for transmitting data. The local network controller performs mobility management and manages the control base stations of the hot area. In addition, we assume there are some small BSs with caching capability can cache hot content and some that cannot. The control BS may be a macro BS or a small BS with larger transmission power, and the small BSs will periodically report information to the control BS.

When the user moves or the network condition changes, the local network controller will migrate the user's current service from the original service base station to the appropriate service base station according to the network condition, the content distribution, and the terminal mobility characteristic. This process is referred to as service migration, as shown in **Fig. 1**. The migration process always maintains service continuity while minimizing the transmission delay, thereby improving the user experience and reducing network traffic.

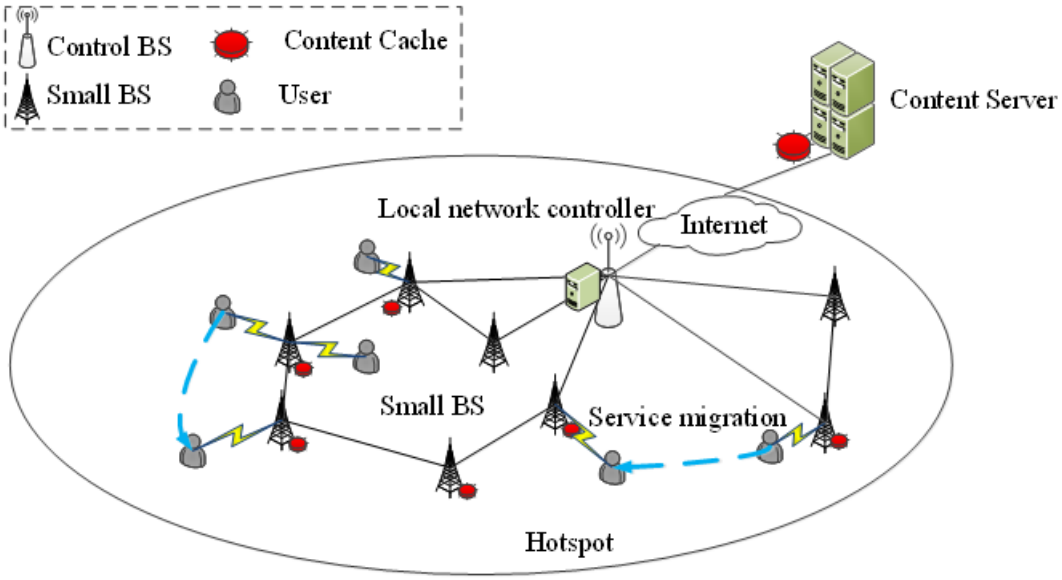


Fig. 1. Service migration system model diagram

2.1 The Hotspot Content Caching

In this paper, the hot contents are assumed to be cached in the distributed BSs based on the topics of interest of the users and the traffic statistics. We define the symbol f_j^i as the access frequency of BS j to content i. In [12], the Window-LFU algorithm was used to effectively predict the popularity of the content based on historical data, that is, f_j^i can also represent the content i on BS j in the updated cycle of popularity.

In this model, the set of BSs with caching capability is denoted as **N** with a size |**N**|. The content files of the entire system cache, i.e., the set of content files (|**S**|) that the users request for, are represented as **S**. The capacity of the serving BS j for the content caching of size is

denoted as Q_j ($j \in \mathbf{N}$) and the size of the content file i is s_i, $s_i \ll Q_j$, $\forall i \in \mathbf{S}$, $\forall j \in \mathbf{N}$.

We assume that when a user has a service request, the request hit can be classified as:

1) Local hit: if a copy of the requested content file is cached on the current serving BS.
2) Server hit: if the request is not a local hit, in which case the request needs to be sent to the content server over the Internet.

Based on the user's service request, the two hits result in different levels of delay: d_{enb}^{ij} and d_{ser}^{ij}. d_{enb}^{ij} represents the transmission delay of the content i from the small BS j and d_{ser}^{ij} represents the transmission delay of the content i from the content server. Normally, the content request for the server hit needs to be connected to the Internet Service Provider (ISP) network through the Packet Data Network (PDN) gateway, which then obtains the specified content file from the original server or Content Distribution Network (CDN) on the Internet. Unlike local hits, server hit content requests are also affected by the network conditions and it is assumed that $d_{enb}^{ij} \ll d_{ser}^{ij}$. Therefore, the selection of the system content library is aimed at maximizing the volume of locally hit data.

Though the content with low popularity occupies an equal amount of cache space as the content with high popularity, it contributes much less to local hit data. Therefore, in order to reduce the transmission delay for the user, the BS in the hotspot should cache the content with high popularity as much as possible.

Let us that the content set **S** is the caching of BS in the hotspot and the updating period T refers to the update time duration of content cached in BS. In the updating period, the small BS records the content access rate. Queue $\vec{f}$ represents the access frequency sorted by non-increment. We also assume that the content access rate follows the Zipf distribution $f(i)$, and satisfies $f(i) \geq \varphi$, where φ is the access frequency factor, $\varphi \in (0,1)$.

$$i \leq (\varphi \sum_{c=1}^{S} c^{-u})_{T}^{-\frac{1}{u}}, u>0 \tag{1}$$

The definition of $I \stackrel{\text{def}}{=} \lceil i \rceil + 1$, $I \leq |\mathbf{S}|$ and I is the serial number of hot content, the set C_{hot} caches the hot content and the set C_{nor} caches the normal content. During the updating cycle, the hot content set is distributed in the hotspot small BS. The content server caches a copy of all the cached content within the system, as shown in **Fig. 2**.

We define β_i as the content type of the user request. β_i =1 means that the user is asking for hot content and β_i =0 means that the user is asking for ordinary content.

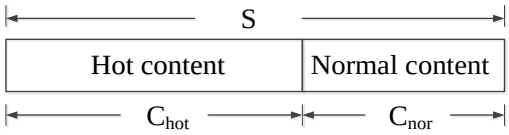


Fig. 2. System hotspot content distribution

2.2 Transmission Delay

The definition indicators y_j^i and χ_j^i indicate if the content i is cached at the BS j and if the content i belongs to the content server respectively. From the perspective of saving inter-network traffic costs, this study assumes that the content request will prioritize the delivery of the BS's service source from the hotspot. When a content request can get the

specified file within the hotspot, it will not request the content server outside the core network using the Internet network.

$$\chi_j^i = \begin{cases} 0 , & y_j^i > 0 \\ 1 , & \text{otherwise} \end{cases} \tag{2}$$

We let θ_j=1 represent BS j with a caching capacity Q_j and θ_j=0 represent BS j without caching capacity. The size of the content i cached in the BS is $s_i, i \in C_{hot}$ and the spatial size of all the contents of the BS caching must not exceed the BS caching size. The space constraint of BS j is expressed as:

$$\sum_{i \in C_{hot}} y_j^i s_i \le Q_j \ , \forall j \in \mathbf{N} \tag{3}$$

Accordingly, the local hit rate is expressed as $P_{ji}^B = \theta_j \cdot y_j^i$. When the user requests for content i that is in the BS j hit, P_{ji}^B =1. When the user requests the content i that is not in the BS j hit, P_{ji}^B =0. Contrary to local hits, the server hit is expressed as $P_{ji}^S = \chi_j^i$.When the user requests the content i that is in the server hit, P_{ji}^S =1. When the user requests the content i that is not in the server hit, P_{ji}^S =0. We assume that the entire system of cached content can always be delivered to the users and satisfies the formula expressed as:

$$P_{ji}^B + P_{ji}^S = 1 \tag{4}$$

For the users' service requests, in the case of a server hit, the request needs to be sent to the content server through the core network, which will bring the corresponding inter-network traffic, assuming the network generated traffic threshold of B_e. Since a content caused by the inter-network traffic is $\sum f_j^i s_i$, the total traffic flow across all contents in the system can be expressed as:

$$\sum_{j \in \mathbf{N}} \sum_{i \in \mathbf{S}} f_j^i \chi_j^i s_i \le B_e \tag{5}$$

The transmission delay generated by the user request is divided into two processes: the delay from the BS to the user and the delay from the content server to the BS. The definition delay D_j^i indicates that the BS j transmits the total transmission delay of the content i for the user. The definition delay d_{enb}^{ij} indicates the transmission delay when the BS j serves the content i for the user. The definition delay d_{ser}^{ij} yindicates that the server transmits the transmission delay of the content i to the BS j.

In this paper, we classified the delay into the following cases:

1)When the content requested by the user is not cached at the BS, the BS is required to initiate the request to the content server. In such cases, the delay for the content request of user A is expressed as:

$$D_j^i = (d_{enb}^{ij} + d_{ser}^{ij}) \tag{6}$$

2) When a user requests an ordinary content and the BS does not have a cache of the content, the BS also needs to request for content from the content server. In such cases, delay for the content request of user B is expressed as:

$$D_j^i = (d_{enb}^{ij} + d_{ser}^{ij}) \tag{7}$$

3) When the user requests for hot content and the BS can meet the needs of users. In such cases, delay for the content request of user C is expressed as:

$$D_j^i = d_{enb}^{ij} \tag{8}$$

From the three different delay types we can see, the formulas (6) and (7) have the same

expression, which can be termed as a delay expression. Thus, the total transmission delay generated for the requested content of the user can be expressed as:

$$D_j^i = P_{ji}^B d_{enb}^{ij} + P_{ji}^S (d_{enb}^{ij} + d_{ser}^{ij}) \tag{9}$$

The transmission rate of the BS is expressed as $R(p)=W\log_2(1+{}^{hp}\!/_{\delta^2})$, where W represents the channel bandwidth of the BS, h represents the channel gain of the BS, and p represents the transmit power of the BS. Consequently, we can rewrite the delay expression as:

$$d_{enb}^{ij} = {}^{s_i}\!/_{R_j(p)} \tag{10}$$

$$d_{ser}^{ij} = s_i \cdot D_{ser} \tag{11}$$

where D_{ser} is the transmission delay of the unit data quantity from the BS to the content server. It is reasonably assumed that it is a random variable with uniform distribution.

3. The Formation of Service Migration Model

In this paper, we propose an optimization model to minimize the cumulative transmission delay, with the optimization target and QoS guarantee defined as the constraints. The service migration method can be divided into the following three steps: 1) The local network controller in the hotspots manages the resource of the small BS by obtaining the cache content information and transmitting the current base network condition in the area. 2) The transmission latency of the current serving BS is compared to that of candidate objective BSs serving the user in the hotspot.3) According to the service migration conditions, determine if the current user service needs to carry out the service migration.

3.1 Service Migration Condition - Migration Factor σ_{jk}

We define the network graph model of the current service BS as $G_j=(\mathbf{N_0}, \mathbf{E}, D)$, where $\mathbf{N_0}$ represents the migration target candidate set, which indicates that the current service BS is located in the hot spot area with other caching content of the small BS. The symbol **E** represents the path set of the current service BS to the BS of the candidate target set. The symbol D represents the time delay on account of selecting different paths of different BSs.

The mathematical expression for single user transmission delay for the system is expressed as:

$$D = (1-\sigma_{jk}) D_j^i + \sigma_{jk} D_k^i \ (k \in N_0(G_j)) \tag{12}$$

The definition indicator σ_{jk} indicates whether the user needs to migrate. $\sigma_{jk}=1$, indicates that the user requests for content from the BS j to migrate to the BS k. On the other hand, $\sigma_{jk}=0$ indicates that no service migration action is performed. D_j^i represents the user's initial access to the BS transmission delay. D_k^i represents the service that has been transferred to the BS k transmission delay.

$$D_j^i = \left[\frac{\theta_j y_j^i}{R_j(p)} + \chi_j^i \cdot \left(\frac{1}{R_j(p)} + D_{ser}\right)\right] \cdot s_i f_j^i \tag{13}$$

$$D_k^i = \left[\frac{\beta_i}{R_k(p)} + (1-\beta_i)\left(\frac{1}{R_k(p)} + D_{ser}\right)\right] \cdot s_i f_k^i \tag{14}$$

In this paper, from the perspective of the system's content caching and the status of the BS network, we propose two conditions to determine whether the user has a service migration.

The conditions are as follows:

1) The migration target candidate set exists in the BS, which satisfies the transmission rate rather than the current service BS transmission rate.

$$\sigma_{jk}=\begin{cases}1\ , & R_j(p)\le R_k(p)\le \mathrm{R_m}\\ 0\ , & \text{else}\end{cases} \tag{15}$$

2) There is a BS in the migration target candidate set, and the content that satisfies the cache is not available to the current service BS.

$$\sigma_{jk}=\begin{cases}1\ , & R_j(p)\le R_k(p)\le \mathrm{R_m} \text{ and } y_j^i=0, y_k^i=1\\ 0\ , & \text{else}\end{cases} \tag{16}$$

According to (12) (13) (14), the transmission delay expression is expressed as:

$$D=\left[\sigma_{jk}f_k^i\left(\frac{\beta_i}{R_k(p)}+(1-\beta_i)\left(\frac{1}{R_k(p)}+D_{ser}\right)\right)+(1-\sigma_{jk})f_j^i\left(\frac{\theta_j y_j^i}{R_j(p)}+\chi_j^i\cdot\left(\frac{1}{R_j(p)}+D_{ser}\right)\right)\right]\cdot s_i \tag{17}$$

As the service migration condition $\sigma_{jk}\in\{0,1\}$ is a discrete variable, which makes this problem is difficult to solve, this paper proposes a solution that needs to remodify the variable.

3.2 Problem Formulation

3.2.1 Caching Capability and Hit Rate

We define the BS side hit rate as expressed $P_{ji}^B=\theta_j\cdot y_j^i$ and the content server hit rate as $P_{ji}^S=\chi_j^i$. Both to satisfy the formula $P_{ji}^B+P_{ji}^S=1$.

According to [13] we know that the BS average hit rate PB is given by

$$P^B=\frac{1}{\mathbf{N}}\sum_{j=1}^{\mathbf{N}}\sum_{i=1}^{\mathrm{C_{hot}}}P_{ji}^B=\frac{\sum_{j=1}^{\mathbf{N}}Q_j\theta_j}{\sum_{j=1}^{\mathbf{N}}Q_j\theta_j+Q_s}\sum_{i=1}^{\mathrm{C_{hot}}}n_i f_j^i \tag{18}$$

where $n_i, i=1,\cdots,\mathbf{N}$ is the total number of the content i cached in the BS. The symbol Q_s represents the contents of the server caching and the BS caching rate in the area is expressed as

$$\alpha=\frac{1}{\mathbf{N}}\sum_{j=1}^{\mathbf{N}}\theta_j\ ,\ j=1,\cdots,\mathbf{N} \tag{19}$$

Assuming that the BS caching space with caching capability is Q_0 and the caching content is the same hotspot content. We can use the number of BSs with caching content to express n_i.

Then the BS average hit rate is re-expressed as

$$P^B=\frac{(Q_0\alpha)^2}{NQ_0\alpha+Q_s}\sum_{i=1}^{\mathrm{C_{hot}}}f_i \tag{20}$$

3.2.2 Content Caching of BS

We define the set of BSs with content caching as $Y_j\in\{y_1,y_2,\cdots,y_{\mathrm{N}}\}$, where $y_j=1$ indicates that the BS has content caching, $y_j=0$ indicates that the BS has no content caching. The content caching matrix Y_{jk} indicates whether the comparison BS j and the BS k have content caching,

$$Y_{jk}=\begin{cases}Y_j\ge Y_k\ ,0\\ Y_j<Y_k\ ,1\end{cases}\ j,k\in\{1,2,\cdots,\mathrm{N}\}(k\ne j) \tag{21}$$

where $\mathrm{Y}_{jk}=0$ indicates that the BS j has content caching and the BS j has not content caching, or the BSs j and k have the same caching capacity. $\mathrm{Y}_{jk}=1$ indicates that the BS j has no content caching and the BS k has content caching.

3.2.3 Transmit Power of BS

We assume that the BS transmit power p is a random variable that satisfies the

function $\Lambda(t)=[\kappa O(\tau t)+\omega]^{\varepsilon}$. The coefficient κ is the peak factor of the function $\Lambda(t)$, the constant ω is the adjustment factor of the function $\Lambda(t)$, the exponent ε is the quality factor of the function $\Lambda(t)$, generally $\varepsilon \geq 1$, and $O(\tau t)$ is a random function with a constant value.

At the same time t, we define the set $\Psi(t)=\{p_1^t, p_2^t, \cdots, p_N^t\}$ to represent the transmit power of the BS. The power matrix Ψ_{jk}^t represents the comparison of the transmission power of the BS j and the BS k,

$$\Psi_{jk}^t=\begin{cases} p_j^t \geq p_k^t, 0 \\ p_j^t < p_k^t, 1 \end{cases} \quad j,k \in \{1,2,\cdots,N\}(k \neq j) \tag{22}$$

where $\Psi_{jk}^t=0$ indicates that the transmission power of the BS j (p_j^t) is greater than or equal to the transmission power of the BS k (p_k^t) at time t. $\Psi_{jk}^t=1$ indicates that the transmission power of the BS j is smaller than the transmission power of the BS k at time t.

3.2.4 Migration Rate Factor σ_l

From 3.2.2 and 3.2.3, we can see that the BSs transmission power is a random variable. At the same time t, the migration factor is defined as

$$\sigma_{jk}^t=\Psi_{jk}^t+(1-\Psi_{jk}^t)Y_{jk} \tag{23}$$

The number of candidate BS sets that all BSs need to migrate is $l=\sum_{k=1,k\neq j}^{N}\sigma_{jk}^t$, where symbol $\mu_{(l)}$represents the candidate BS migration probability and $\mu_{(l)}$is inversely proportional to the delay of target BS. We define the mobility factor $\sigma_l=\max\{\mu_{(l)}\}$. Since the migration factor σ_{jk} is a discrete variable, this problem is an NP-hard problem. Therefore, we use the mobility factor σ_l instead of σ_{jk}.

In this paper, we propose to minimize the cumulative transmission delay of user request content, and also to ensure the feasibility of the caching method, which is to satisfy the total amount of buffer space in the BS. The system satisfies the user request content hit rate constraint and keeps the data request data traffic between the network under the specified budget. This problem can be formulated as

$$\min_{p} \sum_{i\in S}\sum_{j\in N} D \tag{24}$$

$$\text{s.t.} \quad \sum_{i\in S} y_j^i s_j \leq Q_j, \forall j \in N \tag{25}$$

$$\sum_{i\in S}(1-P^B)\, s_i \leq B_e \tag{26}$$

$$R_j(p)=\log_2\left(1+\frac{h_j p_j}{\delta_j^2}\right) \leq R_m \tag{27}$$

$$R_k(p)=\log_2\left(1+\frac{h_k p_k}{\delta_k^2}\right) \leq R_m \tag{28}$$

$$\sigma_l, P^B \in [0,1], \forall l \in \{0,1,\ldots,N-1\} \tag{29}$$

According to the objective function of the optimization equation (24), the optimization equation (24) is not a convex optimization problem because the objective function is a non-convex function on the denominator of the power variable. In this paper, we propose a minimum optimization goal for the system transmission delay and the objective function $D(p)$ cannot be modified to convert it into a convex function.

[Prove] $D(p_k, p_j)$ is a continuous function in the convex region Ω, $(p_0,q_0)\in\Omega$ at any point.

$$D(p_0,q_0)=s_i\sigma_l\left(\frac{P^B}{R_k(p_0)}+(1-P^B)\left(\frac{1}{R_k(p_0)}+D_{ser}\right)\right)$$

$$+s_i(1-\sigma_l)).\left(\frac{P^B}{R_j(q_0)}+(1-P^B)(\frac{1}{R_j(q_0)}+D_{ser})\right) \tag{30}$$

In order to facilitate the calculation later, we will make the formula $\gamma_k=h_k/\delta_k^2$ and $\gamma_j=h_j/\delta_j^2$.

$$R_k(p_k)=\log_2\left(1+\frac{h_k p_k}{\delta_k^2}\right)=\log_2(1+\gamma_k p_k) \tag{31}$$

$$R_j(q_j)=\log_2\left(1+\frac{h_j q_j}{\delta_j^2}\right)=\log_2(1+\gamma_j q_j) \tag{32}$$

The expression for the first-order partial derivative of two-variate function $D(p_k, p_j)$ as

$$\frac{\partial D(p_k,q_j)}{\partial p_k}=-\frac{s_i.\sigma_l.\gamma_k}{\ln 2\cdot(1+\gamma_k p_k)\log_2^2(1+\gamma_k p_k)} \tag{33}$$

$$\frac{\partial D(p_k,p_j)}{\partial p_j}=-\frac{s_i.(\sigma_l-1).\gamma_j}{\ln 2\cdot(1+\gamma_j p_j)\log_2^2(1+\gamma_j p_j)} \tag{34}$$

Because $\gamma_k p_k\neq 0$ and $\gamma_j p_j\neq 0$, then the first-order partial derivative of $D(p_k, p_j)$ function is a continuous function in the domain of $\mathbf{\Omega}$, $D(p_k, p_j)$ function is differentiable, and the Hessian Matrix for $D(p_k, p_j)$ function is

$$G(p)=\nabla^2 D(p)=\begin{pmatrix}\frac{\partial^2 D}{\partial p_k} & \frac{\partial^2 D}{\partial p_k \partial p_j}\\ \frac{\partial^2 D}{\partial p_j \partial p_k} & \frac{\partial^2 D}{\partial p_j}\end{pmatrix}$$
$$=\begin{pmatrix}\frac{s_i\sigma\gamma_k^2}{(\ln 2)^2}\cdot\frac{\ln 2\log_2(1+\gamma_k p_k)+2}{(1+\gamma_k p_k)^2\cdot\log_2^3(1+\gamma_k p_k)} & 0\\ 0 & \frac{s_i(\sigma_l-1)\gamma_j^2}{(\ln 2)^2}\cdot\frac{\ln 2\log_2(1+\gamma_j p_j)+2}{(1+\gamma_j p_j)^2\cdot\log_2^3(1+\gamma_j p_j)}\end{pmatrix} \tag{35}$$

According to [14], the necessary and sufficient condition for the n-variable function $f(x_1, x_2, x_3, \cdots, x_n)$ to be a convex function in the convex region $\mathbf{\Omega}$ is that the Hessian matrix at any point within $\mathbf{\Omega}$ is positive definite. However, the Eigen values of the Hessian Matrix $G(p)$ contain negative numbers, and the Hessian Matrix $G(p)$ is a non-definite matrix. Hence, it is proved that the objective function $D(p)$ is a non-convex function.

4. Service Migration Method based on Firefly Algorithm

In this section, we use the firefly algorithm (FA) [15] to get the solutions of our proposed optimization problem. The FA is a meta-heuristic algorithm inspired by the flashing behavior of fireflies. The FA has good performance in terms of numerical optimization [16] [17] and combinatorial optimization [18], and outperforms other meta-heuristics. The FA is based on the following idealized behavior of the flashing characteristics of fireflies:

• All fireflies are unisex and so one firefly is attracted to other fireflies regardless of their sex;

• Their attractiveness is proportional to their brightness and thus for any two flashing fireflies, the less bright one will move towards the brighter one. Since the attractiveness is proportional to the brightness, both decrease as the distance increases. If there is no firefly

brighter than any other firefly, they move randomly;

• The brightness or light intensity of a firefly is affected or determined by the landscape of the objective function to be optimized.

In this algorithm, the firefly's attractiveness is proportional to the light intensity of the nearby firefly. We can define the diagonal coordinates of the firefly *i* and firefly *j* to represent the attractiveness.

The degree to which the fireflies are attracted to each other depends on the intensity of light they emit, and the higher the light intensity, the better the position is, and the better the target value.

Define the light intensity expression shown as

$$I(r)=\mathrm{I}_0 e^{-\varphi r^2} \tag{36}$$

where I_0 represents the brightness of the brightest firefly, φ is the fluorescein absorption rate, and r is the Cartesian distance between the fireflies. Assuming that the coordinates of firefly i and firefly j are x_i, x_j, the distance between two fireflies is $r_{ij}=\left\|x_i-x_j\right\|$.

The more attractive the fireflies are, the higher the attraction factor mentioned by the FA, the more the number of fireflies to attract, the more attractive factors that attract the perceived range of small fireflies to move in this direction. If the attraction factors are the same, the fireflies move randomly. The light intensity and attractiveness are inversely proportional to the distance between the fireflies and are reduced as the distance increases.

The definition of the degree of attraction shown as

$$\vartheta(r)=\vartheta_0 e^{-\varphi r^2} \tag{37}$$

where ϑ_0 indicates the attractiveness of the brightest firefly.

In every iteration, each firefly will move toward to the firefly with larger light intensity. The movement of this firefly is determined by

$$x_j^{T+1}=x_j^T+\vartheta\left(r_{ij}\right)\left[x_i^T-x_j^T\right]+\lambda\varepsilon_j \tag{38}$$

where T is the number of iterations of the algorithm; x_i and x_j are the location of the firefly; λ is the random step; and the general range is [0,1]. The vector ε_j is usually a random function vector generated by a Gaussian distribution, Uniform distribution, or other distribution.

The firefly algorithm can be summarized as follows: initializing the parameters such as fluorescein, updating the fluorescein, determining the next direction of movement, updating the decision radius and returning to the re-cycle.

In the question posed in this paper, when the distance between the firefly *i* and the firefly *j* exceeds the perceived radius of the fireflies, we determine that the two fireflies are invisible to each other, that is, the attraction of the two fireflies to each other is zero.

The attraction of the fireflies is determined by (37), and the direction of movement of the fireflies is determined by (38). The initial number of users per cell is regarded as the initial firefly group, and the hot content of the small BS is regarded as the light intensity of the initial area.This implementation procedure can be implemented using the following algorithm.

The iteration of the algorithm calculates the transmission rate of each BS in the hotspot and the transmission delay of its BS to the service. In each iteration, an increase in transmission rate of the BS will indicate more services and lesser transmission delay in identification of the transport destination BS.

Algorithm: Service migration method based on firefly algorithm

Objective function of optimization problem $D(p)$, $x=(x_1,x_2,x_3,\cdots,x_m)^T$

For each BS i, generate initial population of F fireflies x_{if} (f = 1, 2, ..., F) using generation of numbers with uniform distribution

Light intensity T_{if} at x_{if} is determined by $D(x_{if})$

Define light absorption coefficient

Initial generation, k = 0

While（k < *Max Generations*)

Update new members, k = k + 1

for f = 1 to F (all F fireflies)

for l = 1 to f (all F fireflies)

if (T_{il} < T_{if}) in case of a minimization problem

Move firefly f towards l in d-dimension;

end if

Attractiveness varies with distance r via –αr2

Evaluate new solutions and update light intensity

end for l

end for f

Rank the fireflies and find the current best

end while

5. Performance Evaluation

In this section, we evaluate the performances of the proposed algorithm by MATLAB and assume that there are 15 BSs in the hotspot with the deployment distribution shown in **Fig. 3**. Furthermore, we assume that there are 12 BSs with caching capability, which are BS numbers 1,2,3,4,7,8,9,10,12,13,14, and 15. Thus, in the region, the BS caching rate is α= 0.8. The simulation parameters are listed in shown in **Table 1**.

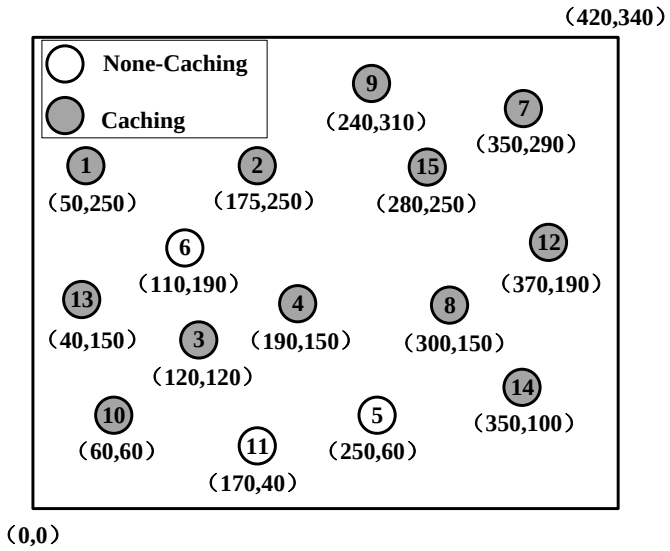


Fig. 3. Distribution of hotspot BSs

Table 1. Simulation parameters

Parameters	Values	Parameters	Values
BS caching rate	0.4、0.66、0.8、1	Number of caching contents	40
BS caching capacity	1 Tbit	Number of hot contents	30
System caching capacity	20 Tbit	Content size	[5,50] Gbit
BS transmit power	[18,22] mW	Zipf parameter	0.7
Unit data volume server hit delay	10^{-5} s/bit	Inter-network traffic threshold	300 Gbit
Number of firefly population	50	Step size	0.03
The firefly light update rate	0.6	Perceived radius	100 m

In **Fig. 4** and **Fig. 5** we improve the firefly algorithm to simulate the total transmission delay and the inter-network traffic with the change of content requests in the hotspot. The transmit power of each BS is in the range of 18 mW to 22 mW with an average transmit power of 20 mW, a BS bandwidth of 40 MHz, a caching capacity of 1 Tbit, the content caching rate α=0.8, the caching content obeys the Zipf distribution with the skew parameter of 0.7, the number of content files in the system is 40, the content file size range is 5 Gbit to 50 Gbit, the content server storage space is 20 Tbit, and the unit data volume Server hit delay D_{ser} is 10^{-5} s/bit.

By comparing the unused SMM and based on the SMM curve in **Fig. 4**, it is not difficult to find the cumulative transmission delay of the unused SMM and the SMM. When the number of requests increases to 730, the cumulative transmission delay of the system under the SMM

is about 2 seconds lower than the cumulative transmission delay under the unused SMM. This shows that the use of SMM can significantly reduce the system's cumulative transmission delay.

Fig. 5 shows the relationship between the inter-network traffic and the number of content requests. It can be seen from the figure that the inter-network traffic between the unused SMM and the SMM is increasing, and the difference between the two curves is increasing, but the inter-network traffic cannot exceed the inter-network traffic threshold B_e. As a result, the inter-network traffic curve becomes more stable as the number of content requests increases. If the request is not hot contents, the area BSs do not cache the contents, so BSs need to request the content server. It will inevitably lead to inter network traffic, which is why the use of SMM will also generate a large number of network traffic reasons. When the number of content requests reaches 100 to 150, access to content files is mostly non-hot content, which leads to a sudden increase in network traffic. When the number of requests reaches 500, the volume of inter-network traffic using the SMM is reduced by about 30 Gbit compared to the unused traffic migration method. This shows that the use of SMM can significantly reduce the network traffic within the system.

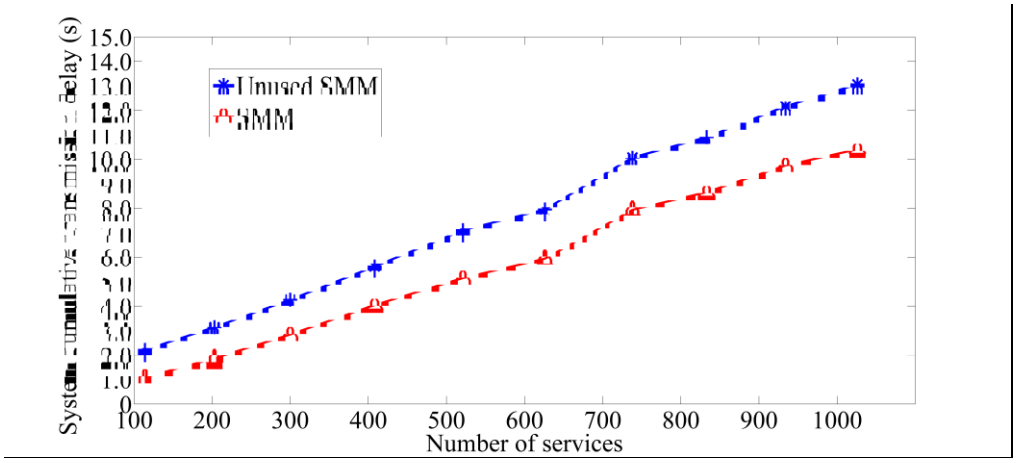


Fig. 4. System cumulative transmission delay and content access graph

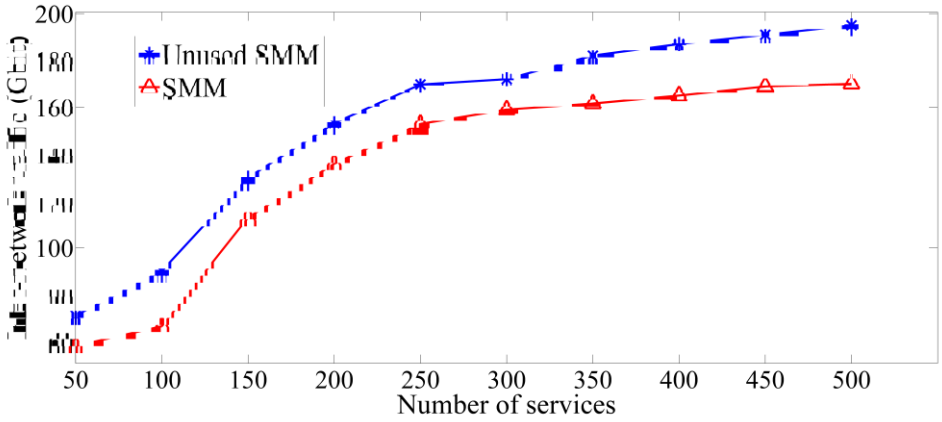


Fig. 5. System network traffic and content traffic graph

Fig. 6 shows a change in the number of requests before and after the BS service migration. It can be seen from the figure that the number of requests for content changes before and after the use of the SMM. Since the BSs on the 5th, 6th and 11th stations do not have the caching capability, it is clear that the number of accesses before the use is much lower than after using the SMM. Since the BSs on the 7th, 9th and 12th stations have caching contents and the network is in good condition, the number of requests by the BS after using the SMM is greatly increased. The central idea of this paper is that the current service BS caching status and poor network conditions can be improved through the SMM to the current service BS service migration to other caching conditions and network conditions of good BS services.

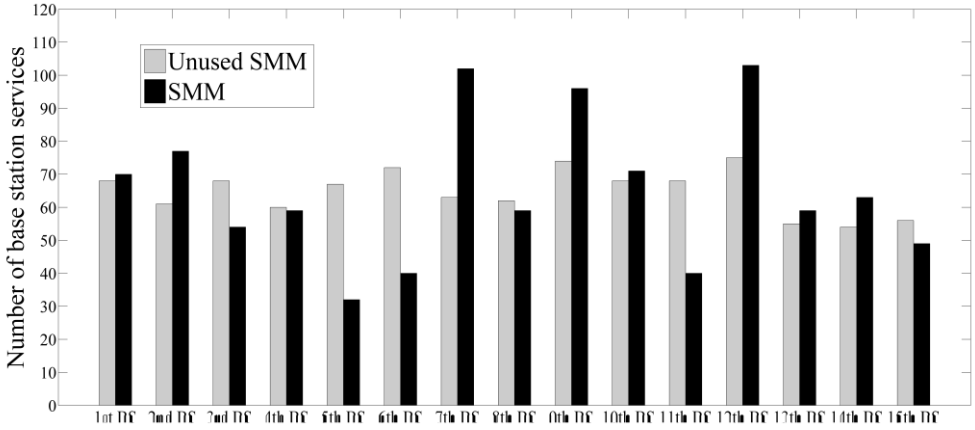
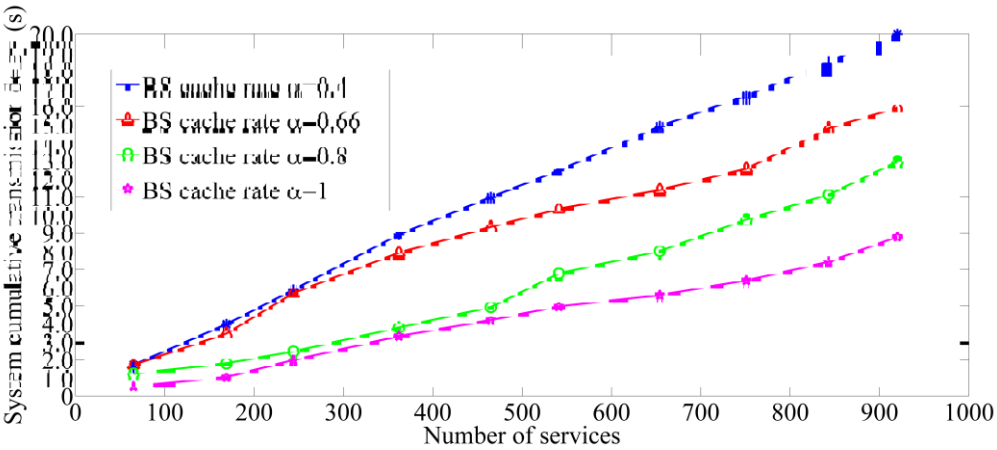
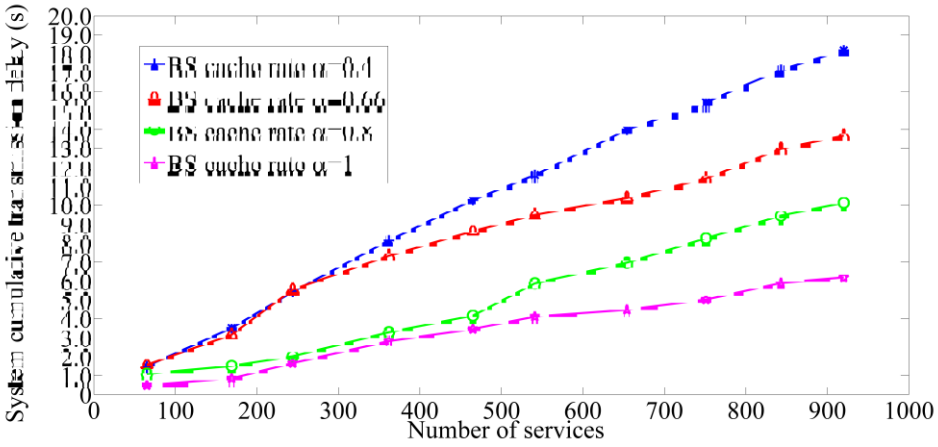


Fig. 6. Content access number of the BS service before and after the migration

Fig. 7(a) shows the cumulative transmission delay and the BS caching rate when the SMM is not used. **Fig. 7(b)** shows the cumulative transmission delay and the BS caching rate when the SMM is used. It can be seen from the figure that the change of the BS caching rate in the cell is also one of the factors that affect the cumulative transmission delay in the system. When α=0.4, the cumulative transmission delay of the system is the largest when the number of requests is the same, and the cumulative transmission delay gradually decreases as α increases.



(a) Unused SMM



(b) Used SMM

Fig. 7. System cumulative transmission delay and BS caching rate

6. Conclusion

In this paper, we propose a SMM that considers content caching and network state conditions in ultra-dense network. Through the service migration method proposed in this paper, we choose the target service BS with low transmission delay and good network condition. In this paper, the minimum cumulative transmission delay in the system is used to optimize the target model, and the solution of the minimum transmission delay BS is found in the iterative method based on the improved firefly algorithm. Simulation results show that compared with traditional collaborative content cashing scheme, the proposed SMM can significantly decrease transmission delay and network traffic flow.

Acknowledgements

This work was supported by Natural Science Foundation of China (61372125), National Science and Technology Major Project of Ministry of Science and Technology of China (2017ZX03001008), Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (16KJA510005), General university graduate degree innovation project of Jiangsu Province (SJLX16_0319), and the open research fund of National Mobile Communications Research Laboratory, Southeast University (2015D10).

References

[1] Kleiner P C, Byers. "Internet Trends 2016 - Code Conference [J]," June 2016. Article (CrossRef Link).

[2] I. Hwang, B. Song and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20-27, June 2013. Article (CrossRef Link).

[3] J He, H Zhang, B Zhao, et al. "A Collaborative Framework for In-network Video Caching in Mobile Networks," *Eprint Arxiv*, 2014. Article (CrossRef Link).

[4] C. X. Wang et al., "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122-130, February 2014. Article (CrossRef Link).
[5] J. Y. Kim, G. M. Lee and J. K. Choi, "Efficient Multicast Schemes Using In-Network Caching for Optimal Content Delivery," *IEEE Communications Letters*, vol. 17, no. 5, pp. 1048-1051, May 2013. Article (CrossRef Link).
[6] H. J. Kang and C. G. Kang, "Mobile device-to-device (D2D) content delivery networking: A design and optimization framework," in *Journal of Communications and Networks*, vol. 16, no. 5, pp. 568-577, Oct. 2014. Article (CrossRef Link).
[7] H. J. Kang, K. Y. Park, K. Cho and C. G. Kang, "Mobile caching policies for device-to-device (D2D) content delivery networking," in *Proc. of 2014 IEEE Conference on Computer Communications Workshops* (*INFOCOM WKSHPS*), Toronto, ON, 2014, pp. 299-304. Article (CrossRef Link).
[8] S. Yan, Q. Zhao, X. Huang and Y. Ma, "A migrating optimization method for CDN based on distributed mobility management," in *Proc. of 2013 5th IEEE International Conference on Broadband Network & Multimedia Technology*, Guilin, 2013, pp. 155-159. Article (CrossRef Link).
[9] Lei Yang, Chaowei Tang, Heng Wang and Hui Tang, "Multi-path Routing Policy for Content Distribution in Content Network," *KSII Transactions on Internet and Information Systems*, vol. 11, no. 5, pp. 2379-2397, 2017. DOI: 10.3837/tiis.2017.05.004. Article (CrossRef Link).
[10] Zhang H, Dong Y, Cheng J, et al. "Fronthauling for 5G LTE-U Ultra Dense Cloud Small Cell Networks," *IEEE Wireless Communications*, vol. 23, no. 6, pp. 48-53, 2016. Article (CrossRef Link).
[11] Zhang H, Huang S, Jiang C, et al. "Energy Efficient User Association and Power Allocation in Millimeter Wave Based Ultra Dense Networks with Energy Harvesting Base Stations," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1936 - 1947, 2017. Article (CrossRef Link).
[12] W C Hou, S Wang."Size-Adjusted Sliding Window LFU - A New Web Caching Scheme[M]," *Database and Expert Systems Applications. Springer Berlin Heidelberg*, pp. 567-576, 2001. Article (CrossRef Link).
[13] C. Yang, Z. Chen, Y. Yao, B. Xia and H. Liu, "Energy efficiency in wireless cooperative caching networks," in *Proc. of 2014 IEEE International Conference on Communications (ICC)*, Sydney, NSW, pp. 4975-4980, 2014. Article (CrossRef Link).
[14] S Boyd, L Vandenberghe, "Convex Optimization," *Cambridge University Press*, New York, NY, USA 2004. Article (CrossRef Link).
[15] Yang X S, "Firefly Algorithm, Stochastic Test Functions and Design Optimisation[J]," in *International Journal of Bio-Inspired Computation*, vol. 2, no. 2, 2010. Article (CrossRef Link).
[16] S. Lukasik, S. Zak, "Firefly algorithm for continuous constrained optimization tasks," in *Proc. of International Conference on Computational Collective Intelligence (ICCCI 2009)*, Lecture Notes in Artificial Intelligence, vol. 5796, pp. 97-100, 2009. Article (CrossRef Link).
[17] M. K. Sayadi, R. Ramezanian, N. Ghaffarinasab, "A discrete firefly meta-heuristic with local search for makespan minimization in permutation flow shop scheduling problems," *International Journal of Industrial Engineering Computations*, pp. 1-10, 2010. Article (CrossRef Link).
[18] X.S. Yang, S. S. S. Hosseini, A. H. Gandomi, "Firefly Algorithm for solving non-convex economic dispatch problems with valve loading effect," *Applied Soft Computing*, vol. 12, no. 3, pp. 1180-1186, 2012. Article (CrossRef Link).

Chenjun Zhou, is currently towards the master degree of Information and Communication Engineering at Nanjing University of Posts and Telecommunications. His research interest focuses on service migration in Ultra-Dense Networks.

Xiaorong Zhu, is a professor at the wireless communication key lab of Jiangsu Province, Nanjing University of Posts and Telecommunications. She received her PhD degree in wireless communications in 2008 from Southeast University, Nanjing, China. She was a postdoctoral in The Chinese University of Hong Kong in 2008 and 2009. Her research interests include heterogeneous wireless networks, wireless access technology, as well as The Internet of Things.

Hongbo Zhu, is a professor in the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications. His current research interests are in wireless communications and electromagnetic compatibility, sensor network technology and broad-band wireless communications.