Optimal Information Centric Caching in 5G Device-to-Device Communications

Optimal Information Centric Caching in 5G Device-to-Device Communications

Changqiao Xu, *Senior Member, IEEE,* Mu Wang, Xingyan Chen, Lujie Zhong, and Luigi Alfredo Grieco, *Senior Member, IEEE*

Abstract—Device-to-Device (D2D) communications are a prominent feature of 5G systems, introduced to provide a native support to distributed services in mobile environments. D2D technologies enable straight interactions between mobile terminals without a compulsory involvement of base stations. In this manuscript, we study and propose an optimized caching strategy to content distribution on top of D2D technology, based on Information Centric Networking (ICN) principles. The rationale is that ICN architectures can provide seamless support to mobile services and decouple contents from node identifiers, thus providing a promising match with D2D requirements. To this end, a novel fluid-based model in proposed hereby that catches the interplay between ICN functionalities, D2D requirements and 5G specifications. Then, based on this model, an optimal content replication problem is formulated, encompassing caching overhead and system load. Additionally, this problem is thoroughly analyzed to prove that it has an optimal solution with time threshold form. A practical algorithm ς^*-OCP is further proposed in order to implement the optimal caching control in realistic environments. Finally, a massive simulation campaign is carried out to test the proposed algorithm in comparison to state of the art solutions.

Index Terms—Information Centric Networks(ICNs), 5G Wireless Networks, D2D, In-network Caching, Optimal Control.

1 Introduction

The joint efforts of standardization groups such as 3GPP and IMT-2020PG are turning the vision of fifth generation (5G) [1] [2] networks into reality. Thanks to the integration of heterogeneous technologies (e.g., LTE, WiFi, LiFi, mmWave [3]) deployed across macro, pico and femto base stations, 5G networks can provide seamlessly access and 1000x increase of network capacity. The expected hyper-dense deployment of 5G systems and the rich and diversified eco-system of services they support the call for new data sharing models that push to the edge traffic and complexity to magnify the return of investment [4].

As one of the key technologies of 5G systems, device-to-device (D2D) communications [5] reuse the cellular spectrum and caching at mobile devices to enable users sharing content with each other directly, hence offloading traffic to the edge and shortening the packet latencies. In D2D scenarios, we expect mobile and static nodes that support information centric services [6] and Internet of Things [1] applications thanks to a capillary exchange of data without a compulsory mediation of base stations. Besides, 5G D2D as a heterogeneous network enables mobile devices concurrently using multiple communication technologies [7] including LTE-D2D, WiFi-Direct and mmWAVE, thus further improving the D2D data delivery performance. Unfortunately, such a heterogeneous environment, coupling with high dynamic of mobile nodes, does not immediately match IP objectives, thus requiring overlay structures that optimize content distribution to users [8], [9]. In particular, while the underlying IP protocol provides host to host services, users demand content oriented applications. In particular, while the underlying IP protocol provides host to host services, users look for content oriented applications. This mismatch, especially in mobile scenarios with in-network caching capabilities, could impair the performance of the 5G D2D technology [10], [11].

These limitations can be overcome thanks to Information Centric Networks (ICNs) [12], which represent a possible evolution towards future Internet. As a matter of fact, ICNs are grounded on name based networking primitives and can natively provide content oriented services also in mobile networks [13], [14]. As Fig. 1 shows, the key idea of ICN for 5G D2D scenarios is that: (i) a data consumer asks the network for a specific content name by issuing a request contains the name of requested content instead of address of destination; (ii) the network locates one or more providers (mobile devices that hold the copies of the asked content) and sets up a D2D route; (iii) the providers return the asked content along that route. As content requests are routed by names instead of host identifiers, ICN can easily support mobile scenarios because if a user switches its position in the network this will not affect the way it will keep asking and receiving contents (i.e., by names). This name-based routing design also enables the identical requests from different interfaces being aggregated at intermediate nodes and concurrently served by forwarding data back via corresponding incoming interfaces, hence, inherently accommodating multihoming of 5G and providing multicasting data delivery. Moreover, mobile nodes that act as information relayers can proactively cache contents to serve the same requests

- *C. Xu, M. Wang and X. Chen are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, P. R. China. E-mail: {cqxu, wangmu, chenxingyan}@bupt.edu.cn.*
- *L. Zhong is with Information Engineering College, Capital Normal University, Beijing 100048, P. R. China. E-mail: zljict@gmail.com.*
- *L. A. Grieco is with the Department of Electrical and Information Engineering, Politecnico di Bari, Bari, Italy. E-mail: a.grieco@poliba.it*

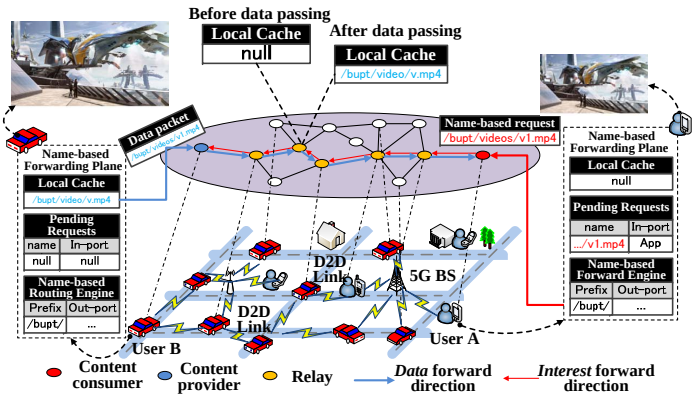


Fig. 1. An illustration of ICN 5G D2D networks

in the future. Thus, content access latency can be reduced and the load to the core alleviated. With the advantages of inherently supporting mobility, multihoming, multicast and in-network caching, ICN outperforms the conventional IP network at content distribution. Hence, by providing all the needs of D2D communications and a natural match for content delivery services (a killer application in 5G), the development of ICN architectures for 5G D2D systems has been considered as a very promising trend for future 5G in many recent papers [13], [15], [16].

As caching mechanism is the decisive factor for the content dissemination efficiency in information centric 5G D2D scenarios, it is necessary to optimize the way different contents are replicated and handled at mobile nodes [14], [17]. Traditional caching strategies already devised in wired ICN [18]–[21] cannot be directly adopted in 5G D2D scenarios because the way they pre-assign content copies to nodes does not take into account topology variations contributed by the mobility of nodes. Moreover, mobile users can be constrained in energy, memory and processing resources, so that each time they cache a content they loose device lifetime, thus requiring a multi-dimensional optimization approach to caching, able to embrace the interplay among ICN functionalities, D2D requirements and 5G specifications. Several studies consider the problem of caching coordination among mobile nodes, but they are tailored to different technologies than 5G [22]–[25].

To bridge this gap, the presented contribution affords the challenges related to ICN-based 5G D2D systems as follows:

(1) We consider ICN as a dynamical system and adopt a fluid-based model to characterize how mobile node parameters (i.e., number of content providers, consumers and forwarders) evolve with system parameters (arrival rate of content requests and caching policy). We also present several numerical results to validate the accuracy of our fluid-based model.
(2) Based on the proposed model, we focus on the tradeoff between caching redundancy and system load when optimizing the caching decision and give the comprehensive proof for the existences of the optimal caching control who has a time threshold form.
(3) We design a practical caching algorithm named ς^*-OCP based on the proposed optimal time-threshold control. We validate the proposed algorithm through a massive simulation campaign, showing that our algorithm outperforms state of the art solutions in terms of caching utilization and content access latency.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 introduces a novel fluid-based model to catch replicas dynamic in distributed ICN 5G D2D. Section 4 analyzes and derives the optimal control for caching policy. Sections 5 and 6 present a practical algorithm and conduct simulation tests to prove the validation of optimal control in realistic environments. Section 7 concludes the paper and discusses future works.

2 Related Work

So far, quite a few papers afforded the design of ICN-based caching mechanisms. Literatures [18]–[20] consider different objectives when formulating caching placement problem. Wang *et al.* [18] formulated the caching placement problem in order to maximize the caching hit ratio. Wu *et al.* in [19] intend to maximize the overall caching benefit, which is defined as the product of average request rate, popularity of content, and hop reduction. The objective function of the optimization problem in [20] aims to maximize the content provider's total caching revenue, including the traffic-proportional profit of serving user demand, the incentives cost for selecting APs to use their storage and access bandwidth, and the infrastructure costs triggered by cache misses at APs. Above studies formulate the problem of caching placement as integer programming problems, which have already been proved to be NP-hard. Consequently, the corresponding caching algorithms are heuristic and sub-optimal. Another caching scheme proposed by Kvaternik *et al.* [21] formulated the caching problem as a convex optimization problem by jointly considering the delivery/caching energy consumption, caching redundancy, and content completeness. As there exists a unique optimal solution for convex problem, a consensus-based caching algorithm is proposed which optimizes the caching configurations by solving the formulated convex optimization problem in a distributed way. However, solutions above focus on caching optimization in wired networks whose topology is static, which are unsuitable for high topology dynamic 5G D2D scenarios.

To introduce the information-centric design into 5G, Liang *et al.* proposed an information-centric virtualization architecture for 5G wireless systems [15]. They formulate the resource allocation and caching strategy as a joint optimization problem, which aims to minimize the inter/intra-ISP traffic and content access delay. The interior point method is employed to derive the solution of above joint optimization problem. Unfortunately, this solution still considers the caching at access points such as based stations, neglecting the D2D communications of 5G environment.

In fact, ICN in 5G D2D scenarios enable mobile users to donate their storage space and bandwidth resource to facilitate the data dissemination. Hence, it is necessary to take the dynamics of mobile users into consideration to design caching strategies for ICN 5G D2D. Recently, several caching mechanisms for ICN-based D2D have been proposed. For instance, Grassi *et al.* applied the ICN architecture into VANETs [22]. A caching everything every where (CEE) strategy is employed to enable vehicle nodes caching every content that received. However, unlike vehicles, which

Table 1
Comparison of existing works

Literatures	Analytical model	5G	D2D	User Mobility	Optimal control
CHPR [18], MBP [19], [20]	×	×	×	×	×
RCO-CCS [21]	×	×	×	×	√
5G ICN [15]	×	√	×	×	√
VNDN [22], DPC [23], GrIMS [24]	×	×	√	√	×
EcoMD [25]	√	×	√	×	×
ICN D2D [13]	×	√	√	×	√
Our solution	√	√	√	√	√

have enough storage and energy resources to execute CEE, most mobile devices such as smart phones and laptops are energy-hungry and resource constrained, which cannot sustain the high resource consumption of CEE. Random-probabilistic caching ($RND\,(p)$) in [10] reduces the caching redundancy by caching the forwarded content with a given probability p. For example, $p = 0.5$ implies that mobile nodes will have 50% percent probability to cache a receiving content, hence reducing the caching redundancy with respect to CEE. However, the randomness feature of $RND\,(p)$ results in the unguaranteed of caching performance.

Deng $et\ al.$ [23] proposed a distributed probabilistic caching strategy for vehicular environment. In this solution, each mobile node calculates the caching probability of given content by the weighted sum of users demand, node centrality and relative movement speed. However, the optimal caching configuration still cannot be achieved as the caching probability is determined by a heuristic method. In our early proposed information-centric architecture GrIMS [24] over VANETs environment, a Cloud-assist information centric architecture is built to monitor the balance between video supply and demand of system. If supply capacity of corresponding content is insufficient, a cooperative caching strategy will be employed to allocate the content replicas to selected nodes. We also proposed an information-centric multimedia content distribution framework over the vehicular networks named EcoMD [25]. A queuing model is built for each content to estimate the bandwidth requirement and average waiting delay. According to the estimation, all nodes on data delivery path will dynamically allocate caching space to storage passing content in order to minimize the overall average waiting delay. However, due to the lack of theoretical analysis on the usage and variation of caching, above solutions in D2D-based ICN still use heuristic methods that cannot provide theoretically proved performance bounds.

Another study that relates to our work introduces a virtualized wireless ICN architecture and considers the resource allocation under D2D communications [13]. Specifically, the optimization problem is formulated to maximize the utility of virtual network operators, and focuses on how to continuously optimize the caching resource allocation among mobile devices in order to improve the backhaul efficiency. Unlike [13], we build a fluid-based model to analyze how the population of consumers and content copies in 5G D2D scenarios evolve with the user demand and caching policy, hence providing theoretical guidelines for the optimal caching design. Moreover, instead of only considering the backhaul efficiency as in [13], we consider the caching optimal problem in 5G D2D scenarios as the tradeoff between system load and caching cost. The comparison of existing works with our work is shown in Table 1.

Table 2
Notations used in model

Symbol	Description
$A\,(t)$	population fraction of ordinary nodes at time t
$D\,(t)$	population fraction of activated consumers at time t
$B\,(t)$	population fraction of inactivated consumers at time t
$X\,(t)$	population fraction of satisfied consumers at time t
$D_f\,(t)$	population fraction of activated forwarders at time t
$B_f\,(t)$	population fraction of inactivated forwarders at time t
$Y\,(t)$	population fraction of satisfied forwarders at time t
β_k	average request rate of chunk k
$\|E\|$	average number of nodes in one-hop range
λ	information spreading rate
$\sigma\,(t)$	controllable cache probability at time t
υ_k	average cache eviction rate of chunk k

3 System Model

This section proposes a novel fluid based model to describe the dynamics of ICN caching in 5G D2D systems. The notations used in the model is summarized in Table 2

3.1 Assumptions

Before presenting our model, we make following assumptions:

First, without loss of generality, we consider a 5G D2D scenarios where Named Data Networking (NDN) [22] is employed [1], since NDN paradigm is a mature ICN architecture and continuously developed by researchers worldwide. In NDN, content consumers issue an $Interest$ packet for requesting a content, and routers, upon receiving that $Interest$, firstly check whether the requested content is in their content store (CS). If not, the name and incoming interface of $Interest$ packet will be recorded in pending $Interest$ table (PIT) and forwarded out to the next-hop according to the forwarding information base (FIB). The above process is repeated until at least one content provider if found, which is in possess of the asked content. Then, the content provider returns the $data$ packets encapsulating the requested content along the reverse direction of searching path and nodes on-path will proactively cache the content into their CS. We assume that every mobile device in NDN is equipped with 5G D2D interface. In this way, it is possible to set up a distributed system in which each node can act as content consumer, forwarder, and producer, thanks to the interplay between ICN, caching, and D2D capabilities of 5G.

Second, we assume the content is divided into several chunks and each chunk has equal size, nodes request/cache the content in the unit of chunks as done in [21].

Third, we assume the movement behavior of mobile nodes follows the Random Way Point (RWP) model, which is a general mobility model used in the studies of mobile networks, especially for D2D environment [26], [27]. In this model, mobile nodes move in a convex area A (i.e., square, unit disk, etc.) with randomly selected destinations and

1. The proposed model also holds for any other ICN architecture over 5G D2D based on in-network caching.

movement speed. For instance, according to RWP, node n will move from A_1 to A_2 (A_1 and A_2 as the start/destination points chosen randomly over A) along a straight line with velocity V_1 whose value is selected by distribution $f_V(v)$. Once reaching the destination, the node n will reselect a new destination A_3 from uniform distribution over A and move straightly to A_3 with newly selected velocity V_2.

3.2 Replica dissemination in mobile ICN

Let $\mathrm{K} = [1, ..., K]$ be the set of chunks belonging to all contents in the network. We now discuss how to build a fluid-based model for a given chunk k ($\forall k \in \mathrm{K}$) in ICN 5G D2D. First, we define four roles of mobile nodes: **Consumer**, the node which issues an *Interest* packet for k; **Relay**, an intermediate node which receives and forwards an *Interest* packet because the requested chunk is not in local cache; **Provider**, the node which holds the replica of k; **Ordinary**, the node which does not belong to any role above. To further describe the node states according to above roles, we introduce the following four bits:

- **Request bit ($\mathbb{R}$):** 1 if the node is a consumer for the chunk, 0 otherwise.
- **Forward bit ($\mathbb{F}$):** 1 if the node is a relay for the chunk, 0 otherwise.
- **Spread bit ($\mathbb{S}$):** 1 if the node is able to spread the request to neighbor node, 0 otherwise.
- **Have bit ($\mathbb{H}$):** 1 if node is a provider, 0 otherwise.

Then, we define 7 possible node states during the chunk dissemination by four bits above. Each state can be described as follows:

- **Ordinary state *A*** ($\mathbb{R}$=0, $\mathbb{F}$=0, $\mathbb{S}$=0, $\mathbb{H}$=0): a node in this state is an ordinary node, let $A(t)$ be the population fraction of nodes in this state at time t.
- **Activated consumer state *D*** ($\mathbb{R}$=1, $\mathbb{F}$=0, $\mathbb{S}$=1, $\mathbb{H}$=0): this state indicates that a consumer node is preparing to send a request, we denote $D(t)$ as the population fraction of nodes in state $\boldsymbol{D}$ at time t.
- **Inactivated consumer state *B*** ($\mathbb{R}$=1, $\mathbb{F}$=0, $\mathbb{S}$=0, $\mathbb{H}$=0): the node is consumer that already sent out request and are waiting for corresponding data back, we define $B(t)$ as the population fraction in state $\boldsymbol{B}$ at time t.
- **Consumer satisfied state *X*** ($\mathbb{R}$=1, $\mathbb{F}$=0, $\mathbb{S}$=0, $\mathbb{H}$=1): in this state, a consumer already obtained the asked chunk and can be considered as a provider in the system. We define $X(t)$ as the population fraction of nodes in $\boldsymbol{X}$ at time t.
- **Activated relay state $\boldsymbol{D_f}$** ($\mathbb{R}$=0, $\mathbb{F}$=1, $\mathbb{S}$=1, $\mathbb{H}$=0): a relay node enters in this state when it receives a request and prepares to forward it. We define $D_f(t)$ as the population fraction in $\boldsymbol{D_f}$ at time t.
- **Inactivated relay state $\boldsymbol{B_f}$** ($\mathbb{R}$=0, $\mathbb{F}$=1, $\mathbb{S}$=0, $\mathbb{H}$=0): the state of relay nodes already sent out the request for k and are waiting for data reply. We define $B_f(t)$ as the population fraction of B_f at time t.
- **Relay satisfied state *Y*** ($\mathbb{R}$=0, $\mathbb{F}$=1, $\mathbb{S}$=0, $\mathbb{H}$=1): a relay node in state Y that has received chunk k and decides to replicate a copy in its local caches, let $Y(t)$ be the population fraction of nodes in this state at time t.

Each node in networks is one of seven state, hence the sum of all nodes of each state is constantly equal to the total number of nodes in networks, namely:

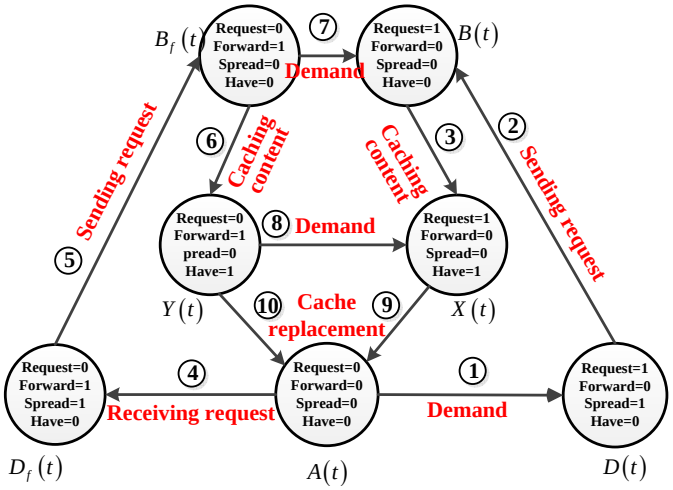


Fig. 2. Possible state transitions of nodes in ICN-based 5G D2D

$$A(t) + D(t) + B(t) + X(t) + D_f(t) + B_f(t) + Y(t) = 1$$

3.3 Fluid-based model

In this subsection, we build a fluid based model to describe the dynamics of $A(t)$, $D(t)$, $B(t)$, $X(t)$, $D_f(t)$, $B_f(t)$, $Y(t)$. As the dynamic of population fraction of each state is derived by the state transition rate, a key point is how to analyze the state transition among above states. According to the data dissemination process in ICN 5G D2D, the transition of above 7 states can be interpreted as Fig. 2, where ten possible types of state transitions (represented by the straight line with arrow) are existing among the states. Now we give the detail description of each state transition:

- **Transition 1**: If node in state $\boldsymbol{A}$ become interested in chunk k, it will convert to $\boldsymbol{D}$. We assume that each node becomes interested in k in a small time interval Δt with a given probability $\beta_k \Delta t$, β_k is the parameter of Poisson distribution, which is only related to the popularity of content k. As $\beta_k \Delta t \approx \beta_k dt$ when Δt is small enough, the conversion rate of a single ordinary node is β_k. Thus, the conversion rate of transition 1 can be denoted by $\beta_k A(t)$ when the population fraction of ordinary node is $A(t)$.
- **Transition 2**: After sending out the $Interest$ packet, an activated consumer in state $\boldsymbol{D}$ will transit to inactivated state $\boldsymbol{B}$. Because all consumers will send out the *Interest* packets to request the chunk k, the conversion rate is equal to $D(t)$ the population portion of nodes in D at current time t.
- **Transition 3**: After receiving requested chunk, inactivated consumers in state $\boldsymbol{B}$ will become satisfied consumers, namely convert from state $\boldsymbol{B}$ to $\boldsymbol{X}$. In our model, a consumer can obtain the content from a neighbor node who is in state $\boldsymbol{X}$ and $\boldsymbol{Y}$. In this case, the probability of converting from $B(t)$ to $X(t)$ can be approximated by the pairwise meeting probability between a node in state B and a provider of chunk k. Namely, the conversion rate of transition 3 is equal to the density of provider $X(t)$+$Y(t)$ times the density of $B(t)$. Let $P(t) := X(t)$+$Y(t)$, the conversion rate can be represented as $P(t)B(t)$.
- **Transition 4**: This transition indicates that an ordinary node becomes an activated relay. The $Interest$ forwarding in 5G D2D can be considered as an epidemic process

(EP) [29] since the activated relays are trying to transform their neighbors into activated relays. Hence, activated relays and consumers can be treated as infect individuals in EP, and ordinary nodes can be treated as suspected individuals in EP. Besides, nodes in NDN do not forward a request that already received [12], which indicates nodes in inactivated state can be treated as recovered individuals in EP that will not participate in the epidemic spreading process. In this case, the conversion rate of this transition can be represented by the following equation according to [29]:

$$\lambda|\bar{E}|A\left(t\right)\left(D\left(t\right)+D_{f}\left(t\right)\right) \tag{1}$$

where λ is the spreading rate, $|\bar{E}|$ denotes the average number of nodes that connect with an ordinary node. In our model, we consider two types of $Interest$ forwarding strategies: broadcast-based and unicast-based. For broadcast-based $Interest$ forwarding, all ordinary nodes that receive the *Interest* will certainly convert to activated relays within Δt, hence the λ in Eq. (1) can be set to 1. For unicast-based $Interest$ forwarding, the activated relay will forward the $Interest$ to only one next-hop and the spreading rate λ of unicast-based forwarding is $\frac{1}{|\bar{E}|}$. Now we discuss how to estimate $|\bar{E}|$ for RWP model. Recall that mobile nodes are moving in a convex area A, according to [26], the probability of node n locating at a position $\boldsymbol{r}$ ($\boldsymbol{r}$ is an two-dimension vector that indicates the coordinate over A) can be given as follows:

$$f\left(\boldsymbol{r}\right)=\frac{1}{\bar{l}s_{A}^{2}}\int_{0}^{\pi}a_{1}a_{2}\left(a_{1}+a_{2}\right)d\phi \tag{2}$$

where $\bar{l}$ is constant and set to 0.521 for RWP, s_A is the area of the A, a_1 and a_2 are simplify for the $a_1\left(\boldsymbol{r},\phi\right)$ and function $a_1\left(\boldsymbol{r},\phi\right)$, which are the length of line segments from $\boldsymbol{r}$ to broader of A whose direction are ϕ and $\pi-\phi$, respectively. We consider the communication range of node is a disk with radius R, for node n at position $\boldsymbol{r}$, the probability of a node in n's one-hop range is given by the following curve integral:

$$p_{\boldsymbol{r}}=\oint_{A_{\boldsymbol{r}}}f\left(\boldsymbol{r}\right)ds \tag{3}$$

where $A_{\boldsymbol{r}}$ is a circular function with centre $\boldsymbol{r}$ and radius R. Hence the average number of connected nodes of n at $\boldsymbol{r}$ is $Np_{\boldsymbol{r}}$, where N is the total number of nodes (as nodes in RWP is statistic equivalent). Therefore, based on Eq. (2) and (3), $|\bar{E}|$ can be derived by following:

$$|\bar{E}|=\oint_{f(A)}Np_{\boldsymbol{r}}f\left(\boldsymbol{r}\right)ds \tag{4}$$

where $f\left(A\right)$ denotes the curve function of area A.

- **Transition 5**: After forwarding the *Interest* packet, relays will convert from state $\boldsymbol{D_f}$ to $\boldsymbol{B_f}$. Similarly to transition 2, all activated relays in state $\boldsymbol{D_f}$ will send out the *Interest* packet for requested chunk and hence the conversion rate of this transition is $D_f\left(t\right)$.
- **Transition 6**: When a relay receives the requested chunk and decides to cache it in local, it will become a member in state $\boldsymbol{Y}$. Similarly, as transition 3, relayers can obtain the chunk from a neighbor who holds the replica. Accordingly, the conversion rate of transition 6 is $\sigma\left(t\right)P\left(t\right)B_f\left(t\right)$, where $\sigma\left(t\right)$ determines whether cache the receiving chunk or not at time t. Namely, $\sigma\left(t\right)$ can be considered as the caching policy. For instance, when employing CEE, the $\sigma\left(t\right)\equiv 1$.
- **Transitions 7** and **8**: Since relays are also mobile users and may become interested to k, they may covert from $\boldsymbol{B_f}$ to $\boldsymbol{B}$ and $\boldsymbol{Y}$ to $\boldsymbol{X}$. Similar as transition (1), the conversion rate of transitions 7 and 8 are equal to $\beta_k B_f\left(t\right)$ and $\beta_k Y\left(t\right)$, respectively.
- **Transitions 9** and **10**: These two transitions indicate mobile nodes evict chunk k from local cache. We denote the cache eviction probability of k in one node as υ_k, namely the rate of transition 9 and 10 are $\upsilon_k Y\left(t\right)$ and $\upsilon_k X\left(t\right)$, respectively. Now we discuss how to derive υ_k. As υ_k is the inverse of average cache lifetime $E\left(T_k\right)$, namely $\upsilon_k=E\left(T_k\right)^{-1}$, hence υ_k can be obtained by deriving $E\left(T_k\right)$. The cache lifetime T_k can be interpreted as the difference between cache miss interval t [28] and t_0, where t_0 is denoted as the time interval between cache eviction and cache miss. We consider the least recently used (LRU) as the cache replacement strategy which evicts the recent least used chunk from the cache, namely k will be replaced if the time interval between two consecutive requests is larger than given value τ_k. Once the requested chunk is not in cache, a cache miss will occur. Thus, according to [28], the cache miss interval t for LRU is composed of a sequence of independent random variable $\{t_1,t_2,...,t_m\}$, namely:

$$t=\sum_{i=1}^{n-1}t_i+t_m \tag{5}$$

where $t_i\left(i\leq m-1\right)$ denotes the epoch between two cache hit and t_m denotes the time interval between the last cache hit and cache miss. Apparently, for any $i\leq m-1$, $t_i\leq\tau_k$ and $t_m>\tau_k$. Since the request arrival probability follows a Poisson distribution with parameter β_k according to analysis in transition 1, the average value of t can be given as $E\left[t\right]=\beta_k^{-1}e^{\beta_k}e^{\tau_k}$ [28]. According to the definition of t_0, $t_0=t_m-\tau_k$, τ_k is a constant value for any given k. Because $t_m>\tau_k$, we thereby derive the expectation of t_m by the following conditional expectation:

$$\begin{aligned}E\left[t_m\right]&=E\left[t_m\mid t_m>\tau_k\right]\\&=\int_0^{\infty}t_m f\left(t_m\mid t_m\geq\tau_k\right)dt_m\\&=\frac{e^{-\beta_k\tau_k}\left(\tau_k+\frac{1}{\beta_k}\right)}{e^{\beta_k\tau_k}}\end{aligned} \tag{6}$$

Table 3
Update and conversion rate among states

Transition	State update	Conversion rate
1	(0,0,0,0)→(1,0,1,0)	$\beta_k A\left(t\right)$
2	(1,0,1,0)→(1,0,0,0)	$D\left(t\right)$
3	(1,0,0,0)→(1,0,0,1)	$P\left(t\right)B\left(t\right)$
4	(0,0,0,0)→(0,1,1,0)	$\lambda\|\bar{E}\|A\left(t\right)\left(D\left(t\right)+D_f\left(t\right)\right)$
5	(0,1,1,0)→(0,1,0,0)	$D_f\left(t\right)$
6	(0,1,0,0)→(0,0,0,1)	$\sigma\left(t\right)P\left(t\right)B_f\left(t\right)$
7	(0,1,0,0)→(1,0,0,0)	$\beta_k B_f\left(t\right)$
8	(0,0,0,1)→(1,0,0,1)	$\beta_k Y\left(t\right)$
9	(1,0,0,1)→(1,0,0,1)	$\upsilon_k X\left(t\right)$
10	(0,1,0,1)→(1,0,0,1)	$\upsilon_k Y\left(t\right)$

Table 4
Parameters setting for 5G D2D

Parameter	Value
Max BSs/UE Tx power for cellular	46/23 dBm
Max UE Tx power for D2D	10dBm
Noise figure	7,11 dB
Mac channel delay	250ms
Communication range	150m
Download Data Rate	300Mbps
Upload Data Rate	50Mbps
Operating frequency	3.5GHz
PropagationLossModel	FriisPropagationLossModel
EnergyDetectionThreshold	-71.9842

Therefore, we have

$$\begin{aligned} E\left(T_k\right) &= E\left[t\right] - E\left[t_0\right] \\ &= \beta_k^{-1} e^{\beta_k} e^{\tau_k} - \frac{e^{-\beta_k \tau_k}\left(\tau_k + \frac{1}{\beta_k}\right)}{e^{\beta_k \tau_k}} + \tau_k \end{aligned} \tag{7}$$

In Table 3, we summarize transitions rate of 10 types of transition we derived. According to Fig. 2 and Table 3, the dynamics of $\mathbf{U}\left(\mathbf{t}\right)$ can be expressed by following O.D.E functions with initial value U_{t_0} at time t_0:

$$\dot{A} = -\beta_k A\left(t\right) - \lambda|\bar{E}|A\left(t\right)\left(D\left(t\right) + D_f\left(t\right)\right) + \upsilon_k P\left(t\right) \tag{8}$$

$$\dot{D} = \beta_k A\left(t\right) - D\left(t\right) \tag{9}$$

$$\dot{B} = D\left(t\right) - P\left(t\right) B\left(t\right) + \beta_k B_f\left(t\right) \tag{10}$$

$$\dot{X} = P\left(t\right) B\left(t\right) + \beta_k Y\left(t\right) - \upsilon_k X\left(t\right) \tag{11}$$

$$\dot{D}_f = \lambda|\bar{E}|A\left(t\right)\left(D\left(t\right) + D_f\left(t\right)\right) - D_f\left(t\right) \tag{12}$$

$$\dot{B}_f = D_f\left(t\right) - \sigma\left(t\right) P\left(t\right) B_f\left(t\right) - \beta_k B_f\left(t\right) \tag{13}$$

$$\dot{Y} = \sigma\left(t\right) P\left(t\right) B_f\left(t\right) - \left(\beta_k + \upsilon_k\right) Y\left(t\right) \tag{14}$$

$$\boldsymbol{U}|_{t=t_0} = U_{t_0} \tag{15}$$

where initial value U_{t_0} $=(A\left(t_0\right),\ D\left(t_0\right),\ B\left(t_0\right),\ X\left(t_0\right),$ $D_f\left(t_0\right), B_f\left(t_0\right), Y\left(t_0\right))$.

Remark: Although the fluid-based model we built is for 5G-D2D environment, it can be also extended to other mobile scenarios with some modifications on model parameters. For instance, since different routing policies and mobility models are distinguished by the value of λ and $|\bar{E}|$, applying our model to mobile wireless networks [30] or wireless sensor networks [31] only need to re-set the λ and $|\bar{E}|$ by investigating the routing policy and mobility model of the considered scenarios. In addition, other caching replacement policies such as FIFO or LFU could also be considered in model by regulating the caching eviction rate υ_k .

3.4 Accuracy of the O.D.E Approximation

In order to evaluate the accuracy of our fluid-based model, we conduct a series of simulation tests by ndnSIM [32] based on NS-3 [33]. The simulation parameter settings in terms of network and users are given as follows:

We consider a 6000*6000 m^2 scenario and 2000 mobile nodes are moving in the scenario according to the RWP model, whose velocity ranges from $[10, 40]\, m/s$. In order to simulate the 5G-D2D scenarios in NS-3, we basically re-set the physical and MAC layer parameters and modulation schemes according to the requirement of 5G industrial standarization [34]. The detail parameter settings are given as Table 4. Our simulation considers videos with 2000kbps playback bit rate and $120s$ long. We further divide each video into 60 chunks, namely each chunk is $2s$ long with size of 500KB. The size of cache in each node is set to 20000 MTUs, where a MTU is equal to 1500B. In this case, mobile node can store at most 60 chunks. The cache operation of each node is in chunk-level as it will cache or switch out the whole chunk. The cache replacement policy is LRU.

Figs. 3 and 4 show the system evolving with different initial states based on broadcasting $Interest$ forwarding. Figs. 5 and 6 show the system evolving with different initial states based on random unicast $Interest$ forwarding. For each scenario, we repeat 20 runs with different random seeds and take the average of results. As shown in the figures, our fluid model converges well to the simulation tests in both scenarios with only small difference. We also observe that when request rate of content is given, the variation of $B\left(t\right)$ has a strong relationship with number of caching copies $Y\left(t_0\right)$. This because according to the O.D.E function (8)-(14), high value of $Y\left(t\right)$ can accelerate the transiting from $\boldsymbol{B}$ to $\boldsymbol{X}$. As a result, system load (number of nodes that wait for data returning, i.e., $\boldsymbol{B}\left(t\right)$) can be effectively alleviated. In contrast, lower $Y\left(t_0\right)$ will in turn lead to higher $\boldsymbol{B}\left(t\right)$. This observation is consistent with the fact that caching can speed up the data dissemination and reduce the waiting delay of users.

4 Caching Optimization

In this section, we will discuss how to optimize the caching policy in ICN 5G D2D based on our fluid-based model.

4.1 Problem Formulation

ICN 5G D2D offloads from networks by replicating content copies at mobile nodes. In principle, the more nodes enable caching operations the better are. On the other hand, since intermediate nodes in 5G D2D scenarios are also mobile users with constrained resources, caching operations should be carefully enabled to preserve device lifetime. Additionally, considering 5G D2D networks are highly dynamic in terms of topology and nodes state, it is necessary to enable mobile nodes continuously configure the cache strategies in order to adapt the dynamic environment. We conclude three cache optimization guidelines for ICN 5G D2D: 1) chunks that belong to more popular contents and having less replicas in the network should be cached with a higher priority; 2) resources devoted to caching should be properly tuned to strive a balance between network responsiveness and device lifetime; 3) the caching process duration $T\left(T \to \infty\right)$ in 5G-D2D should be divided into time slots as [21], [24], i.e., $T := \left(T_1, T_2, \ldots, T_n, \ldots\right)$, where each time slot $\Delta T_i = T_{i+1} - T_i, (i = 1, 2, 3, \ldots)$. Mobile nodes will reconfigure the caching strategies at the beginning of the time slot ΔT_i to accommodate the network dynamic.

According to the above guidelines, the objective function of caching optimization problem can be expressed by:

$$J_{\Delta T_i, \sigma} = \psi B_{T_i, \sigma}\left(T_\sigma\right) + \left(1 - \psi\right) Y_{T_i, \sigma}\left(T_\sigma\right) \tag{16}$$

where $B_{T_i,\sigma}\left(\centerdot\right)$ and $Y_{T_i,\sigma}\left(\centerdot\right)$ denote the value of $B\left(t\right)$ and $Y\left(t\right)$ with caching control $\sigma\left(t\right)$ and initial condition $\boldsymbol{U}|_{t=T_i}$. We define T_σ as the time of $B_{T_i,\sigma}\left(t\right)$ reaches the peak

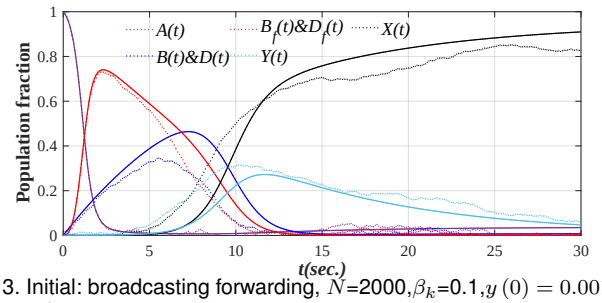


Fig. 3. Initial: broadcasting forwarding, N=2000,β_k=0.1,$y\left(0\right)=0.001$

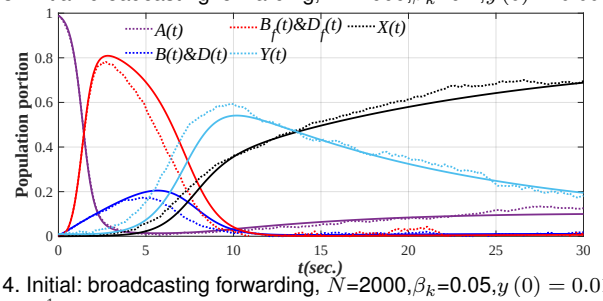


Fig. 4. Initial: broadcasting forwarding, N=2000,β_k=0.05,$y\left(0\right)=0.01$

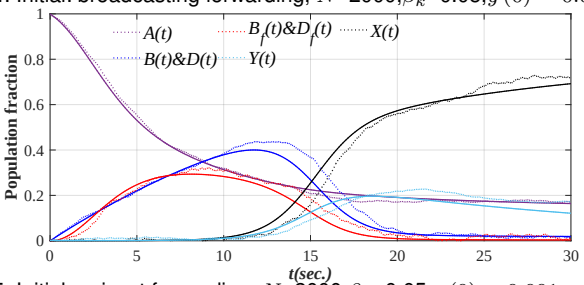


Fig. 5. Initial: unicast forwarding, N=2000,β_k=0.05,$y\left(0\right)=0.001$

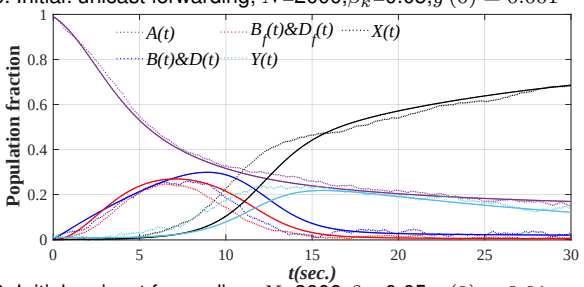


Fig. 6. Initial: unicast forwarding, N=2000,β_k=0.05,$y\left(0\right)=0.01$

value in time slot ΔT_i, namely $\left\{T_\sigma \left| B_{T_i,\sigma}\left(T_\sigma\right)=\max_{\Delta T_i} B_\sigma \right.\right\}$. Accordingly, $Y_{T_i,\sigma}\left(T_\sigma\right)$ indicates the population fraction of nodes in state $\boldsymbol{Y}$ when $B_{T_i,\sigma}\left(t\right)$ reaches the peak value. $\psi \in (0,1)$ and can be treated as weight parameters. Intuitively, the first term in Eq. (16) penalizes the system that suffers a higher peak load, which is consistent with the first guideline. The second term can be considered as a penalty for cache redundancy in networks, which is consistent with second guideline. In addition, since $Y_{T_i,\sigma}\left(T_\sigma\right)$ implies the caching cost which can reflect the total caching and energy consumption, the objective function we formulated inherently considers energy consumption when optimizing the caching configuration. The form of caching optimization problem in each time slot ΔT_i can be stated as follows:

$$\textbf{min} \quad J_{\Delta T_i,\sigma} \tag{17}$$

$$\boldsymbol{s.t} \quad 0 \le \sigma\left(t\right) \le 1, t \in \left[T_i, T_{i+1}\right]. \tag{18}$$

4.2 Optimal Control

In this subsection, we will discuss how to establish an optimal control $\sigma\left(t\right)$ for minimizing $J_{\Delta T_i,\sigma}$. Most current

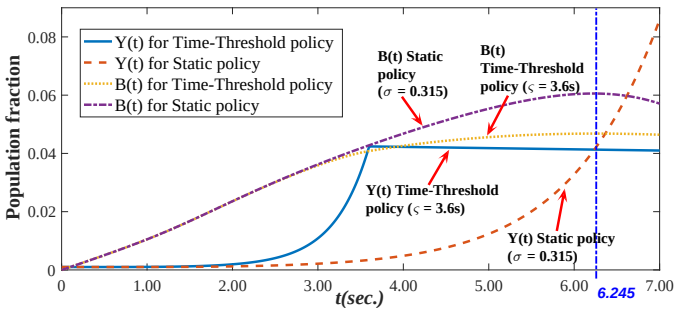


Fig. 7. Probabilistic caching policy vs. time threshold caching policy

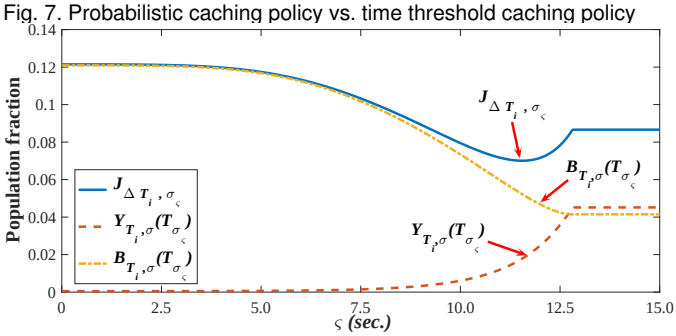


Fig. 8. Evaluation of optimal time threshold ς caching policy under the given initial condition of the O.D.E function

caching solutions in D2D environment [23], [24] are using random caching solutions which cache the content according to a given probability, namely the caching control parameter $\sigma\left(t\right)$ in these strategies is constant in each time slot ΔT (i.e., $\sigma\left(t\right) \equiv p$, $p \in (0,1)$). Instead of applying probabilistic caching strategy, we consider caching control with a time-threshold form $\sigma_{T_i,\varsigma}\left(t\right)$ in each ΔT_i:

$$\sigma_{T_i,\varsigma}\left(t\right) = \begin{cases} 1, T_i \le t < T_i + \varsigma \\ 0, T_i + \varsigma < t \le T_{i+1} \end{cases} \tag{19}$$

Eq. (19) indicates that chunk will be cached by intermediate nodes if receiving time is within ς. Otherwise, the received chunk will only be returned back to consumer without replicating a copy in local cache.

The main reason of optimizing caching with a time threshold form rather than probabilistic can be explained by numerical results of fluid model and theoretical analysis. Fig. 7 shows the numerical results of $\sigma_{T_i,\varsigma}\left(t\right)$ ($\varsigma = 3.5$) and probabilistic caching policy with $\sigma\left(t\right) \equiv 0.315$. We observed that when $T_{\sigma_\varsigma} = T_\sigma$ ($t = 6.245s$), $Y\left(T_{\sigma_\varsigma}\right) = Y\left(T_\sigma\right)$, $B\left(T_{\sigma_\varsigma}\right) \le B\left(T_\sigma\right)$. Namely, time threshold control $\sigma_{T_i,\varsigma}\left(t\right)$ has lower value of Eq . (16) than probabilistic control $\sigma\left(t\right)$. In addition, to justify the dominance of threshold-based caching control theoretically, we give the following theorem and prove it by the lemmas given in Appendix. A.

Theorem 1. *Given the cost function (16) according to the O.D.E system of (8)-(14), for any probabilistic cache policy with $\sigma\left(0 < \sigma \le 1\right)$ or other dynamic control $\sigma\left(t\right)$, there always exists a time-threshold based cache policy with following form:*

$$\sigma_{T_i,\varsigma}\left(t\right) = \begin{cases} 1, T_i \le t < T_i + \varsigma \\ 0, T_i + \varsigma < t \le T_{i+1} \end{cases} \tag{20}$$

with a lower value of $J_{\Delta T_i,\sigma}$.

Proof. See **Appendix. A.** □

Now we discuss the existence of a solution with time-threshold form for the optimization problem (17) (18). Let

$B_{T_i,\sigma}(T_{\sigma_\varsigma})$ denote the peak value of $B(t)$ with time threshold caching control $\sigma_{T_i,\varsigma}(t)$, $Y_{T_i,\sigma}(T_{\sigma_\varsigma})$ and $J_{\Delta T_i,\sigma_\varsigma}$ are the corresponding $Y_{T_i,\sigma}(T_\sigma)$ and $J_{\Delta T_i,\sigma}$ of $\sigma_{T_i,\varsigma}(t)$, respectively. The optimization problem (17) (18) for time threshold caching control can be rephrased as:

$$\textbf{min} \quad J_{\Delta T_i,\sigma_\varsigma} \tag{21}$$

$$\boldsymbol{s.t} \quad T_i \leq \varsigma \leq T_{i+1}. \tag{22}$$

Fig. 8 shows how $B_{T_i,\sigma}(T_{\sigma_\varsigma})$, $Y_{T_i,\sigma}(T_{\sigma_\varsigma})$ and $J_{\Delta T_i,\sigma_\varsigma}$ vary with ς, where ψ is set to 0.5. With the increasing of ς, we can see a monotonic increasing trend of $Y_{T_i,\sigma}(T_\sigma)$. It is because more intermediate nodes will decide to cache the passing chunk since the time threshold increasing. We also see a decrement trend for $B_{T_i,\sigma}(T_{\sigma_\varsigma})$, this is because the increase rate $\dot{B}(t)$ of state $\boldsymbol{B}$ descends with the increasing of population fraction $Y(t)$ according to the Eq. (10). For the value of cost function $J_{\Delta T_i,\sigma_\varsigma}$, we can see the corresponding curve firstly experience a decrease trend and then increase with the ς, namely exists a minimal value with the vary of ς. The following theorem ensures the existence of optimal solution with time threshold based form:

Theorem 2. *Given the optimization problem (21) and (22) according to the O.D.E system of (8)-(14) with initial condition $\boldsymbol{U}|_{t=T_i}$, there exists an optimal time threshold cache policy $\sigma^*_{T_i,\varsigma}(t)$ for optimization problem (21) (22), where caching time threshold is ς^*.*

Proof. See **Appendix. B.** □

5 Practical Algorithm

In order to validate the effectiveness of our time threshold caching control in realistic environment, we propose a practical caching policy named ς^*-opportunitistic caching policy (ς^*-OCP). We consider a mobile information-centric network where all mobile nodes use 5G-D2D interface to share content and equipped with GPS to record the geographical location and moving velocity. 5G base stations (BSs) in this scenarios act as coordinators to collect the network information.

To accommodate with the dynamic variation of network in terms of request rate and number of users in the area, ς^*-OCP is designed as an online algorithm that executes every time slot ΔT_i. To derive the optimal caching time threshold ς^* of all chunks in each time slot, we need to build the corresponding O.D.E equations (8)-(14) and calculate its numerical solutions. In ς^*-OCP, mobile nodes create a 4-bit map $(\mathbb{R},\mathbb{F},\mathbb{S},\mathbb{H})$ for each chunk k, where $\mathbb{R}$, $\mathbb{F}$, $\mathbb{S}$, $\mathbb{H}$ follow the same definitions of request, forward, spread, and have bit in section 3. A state list is maintained at mobile node to record the 4-bit map of each chunk. To further save storage space, chunks in ordinary state $A(\mathbb{R}=0,\mathbb{F}=0,\mathbb{S}=0,\mathbb{H}=0)$ will not be recorded in the state list. Each node submits this state list and current movement velocity to coordinators every ΔT_i. In order to alleviate the extra bandwidth consumption caused by submission process, we smuggle this information into MAC layer control frame. The population fraction of each state can be estimated by the state lists that submitted by all nodes, $|\bar{E}|$ can be estimated by equation (4) according to the area of scenarios and average moving speed. To calculate

Algorithm 1 ς^*-**Opportunitistic Caching Policy**

Coordinator side:
/* Algorithm proceed in coordinator side */
for each time slot ΔT_i
collect $(\mathbb{R},\mathbb{F},\mathbb{S},\mathbb{H})$ and velocity from mobile nodes;
 for all chunk $k \in \mathbb{K}$
  set network state in current time t_0 as initial condition;
  build the fluid-based model for k by Heun's method;
  calculate the optimal time threshold ς^* by searching the close interval;
  broadcast ς^* to all mobile nodes in communication range;
 end for
end for

User side:
/* Algorithm proceed at mobile user n */
upload $(C_n(t), R_n(t), S_n(t), V_n(t))$
wait for data packet of chunk k coming;
if n is a intermediate node
 if $ReceiveTime \leq \varsigma^*$
  create a replica of k in local CS;
 end if
 send k out according to the entry in PIT;
else if i is a consumer for k
 create a replica of k in local CS;
end if

the request producing rate β_k, let $\left(t^k_{n_i}\right)_{n_i=1}^{\infty}$ denote time sequence of node becoming $(\mathbb{R}=1,\mathbb{F}=0,\mathbb{S}=0,\mathbb{H}=0)$ for chunk k, we can approximate $\beta_k(t)$ in ΔT_i by :

$$\beta_k(t) = \frac{1}{N_A(t)\,\Delta T_i}\int_{\Delta T_i}\sum_{n_i=1}^{\infty}\delta\left(t-t^k_{n_i}\right)dt \tag{23}$$

where $\delta(.)$ is unit impulse function with a form of

$$\delta(t) = \begin{cases} 1, t=0 \\ 0, \text{otherwise} \end{cases}$$

$N_A(t)$ denotes the number of nodes in state A at time t.

Therefore, as the initial state and β_k, $|\bar{E}|$ and υ_k are given, the O.D.E equations of (8)-(14) for each chunk can be built. There are several numerical methods available for deriving the solution of O.D.E equation, such as Euler's, Runge-Kutta and Heun's method [35]. In our algorithm, we use Heun's method to derive the numerical solution of (8)-(14) by the following reasons: Euler's method is simple but has unstable performance in terms of accuracy. Runge-Kutta is more accurate than Euler's, yet requires more execution time and memory. Heun's method can be considered as a tradeoff between accuracy and execution overhead, which is preferred by resource limited mobile environment. To determine the value of weight parameter ψ, we analyze the sensitivity of ς^*-OCP to ψ, which is shown by the Fig. 9. $B^*_{T_i,\sigma}(T_\sigma)$ and $Y^*_{T_i,\sigma}(T_\sigma)$ denote the value of $B_{T_i,\sigma}(T_\sigma)$ and $Y_{T_i,\sigma}(T_\sigma)$ when corresponding $J_{\Delta T_i,\sigma}$ is minimum, respectively. As Figure shows, $B^*_{T_i,\sigma}(T_\sigma)$ ($Y^*_{T_i,\sigma}(T_\sigma)$, respectively) decreases (increases, respectively) with the rising of weight parameter ψ. Moreover, it is also observed that $B^*_{T_i,\sigma}(T_\sigma)$'s gradient declines gradually, and the growing rate of $Y^*_{T_i,\sigma}(T_\sigma)$ rises with ψ. This suggests that setting ψ too big (small) will result in the high cost of caching redundancy (system load), which hinders the performance of the algorithm. Thus, to balance the tradeoff between two optimizing objectives, we set the ψ in ς^*-OCP to 0.5. As the optimization problem (21) (22) has an optimal solution by **Theorem 2**, namely there exists a time threshold-based caching policy that can jointly optimize the system load and caching cost. To find the optimal time threshold-based caching policy practically, we employ the numerical

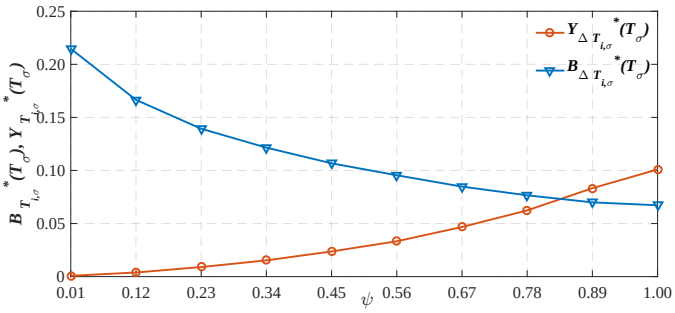


Fig. 9. The value of $B^*_{T_i,\sigma}(T_\sigma)$ and $Y^*_{T_i,\sigma}(T_\sigma)$ vary with ψ

sweeping method which obtains the optimal caching time threshold ς^* by traversing the close interval $[T_i, T_i + \Delta T]$. Consequently, ς^*-OCP derives the caching time threshold that optimizes the problem (21) (22), namely the corresponding caching policy ς^*-OCP is optimal.

After obtaining the optimal time threshold ς^*, the coordinator broadcasts ς^* of each chunk k to all nodes in the communication range via the MAC layer beacon frame. The receiving nodes of chunk k will decide whether to replicate a copy of this chunk in local according to the threshold time ς^* of k. Specifically, if the receiving time exceeds ς^*, the node only forwards the content to the upstream node. Otherwise, the node will replicate the content to local cache and send out data. The pseudo code of the above process is given as **Algorithm 1**. For complexity of this algorithm, we have following proposition.

Proposition 1. *The overall complexity of the algorithm proceed at coordinator side is limited by*

$$O\left(\mid \mathbb{K} \mid . \left(\max\{\frac{\Delta T}{\varepsilon}, H\}\right)\right)$$

where $\mid \mathbb{K} \mid$ *is the total number of chunks,* ε *is the accuracy of searching method and* H *is the complexity of Heun's method.*

Proof. In each time slot, the complexity of this algorithm at server side is determined by the number of chunks, the searching algorithm and Heun's method. As the size of chunk set is $\mid \mathbb{K} \mid$. Searching the interval with accuracy ε require at most $\frac{\Delta T}{\varepsilon}$ iterations. For Heun's method, the corresponding complexity is determined by the number of iterations H. As the searching algorithm and Heun's are proceed in parallel for each chunk. Hence, the proposition is proved. □

As the time complexity of Heun's method and searching is fixed when the accuracy requirement is given, the proposed algorithm only has a polynomial complexity as the number of chunk grows. The algorithm at user side requires them to submit state list and movement velocity to coordinator and receive the optimal caching time threshold every ΔT_i, which can be considered as $O(1)$.

6 Performance Evaluation

In this section, we conduct a series of simulation tests to compare the performance of ς^*-OCP with three state-of-art D2D-based caching strategies, GrIMS [24], DPC [23] and RND(0.5). The parameter settings of network and mobility model are the same as in Section 3.4. The simulation scenario is a 2000*2000 m^2 square with 200 mobile nodes, and simulation time is set to 1000s. Furthermore, to approximate the realistic environment, we adopt at most 40 different videos in our simulation tests. The length of each video is ranging from 120s to 240s, which means the number of chunks contained by a video is from 60 to 120. The distribution of users request for video-level content is described by Zipf distribution according to user behaviors analysis in [36], which means given a video set with n videos, the request probability of the r-th most popular video is

$$P(r) = \frac{\left(\sum_{k=1}^{N} \frac{1}{k^\rho}\right)^{-1}}{r^\rho} \tag{24}$$

where ρ is the Zipf parameter and set to 0.8. After determining which video to watch, users will request the chunks of video in sequence and choose another video after finishing the playback of current video.

We also deploy 25 base stations (BSs) uniformly in the network to collect state list and velocity of users and periodically broadcast the caching time-threshold every 30s according to **Algorithm 1**.

6.1 Simulation Test

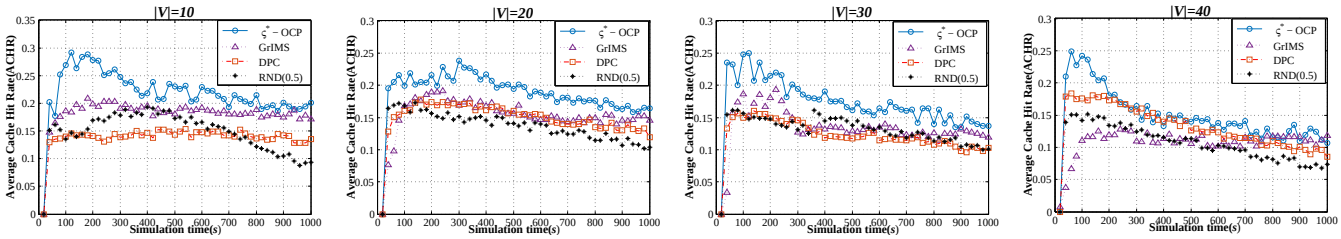
We test the performance of four solutions in terms of average cache hit ratio, caching cost, average downloading time and control overhead. The detail analysis is as following.

Average Cache Hit Ratio (ACHR): In ICN, if one node receives a request and the corresponding chunk is in its local cache, it will be considered as a cache hit event. Otherwise, it is a cache miss event. ACHR indicates average ratio between the number of cache hit events and the total number of received requests. We estimate the ACHR at time t by the following equation:

$$ACHR(t) = \frac{1}{|\mathbb{N}(t)|} \sum_{i \in \mathbb{N}(t)} \frac{H_i^h(t)}{H_i(t)}$$

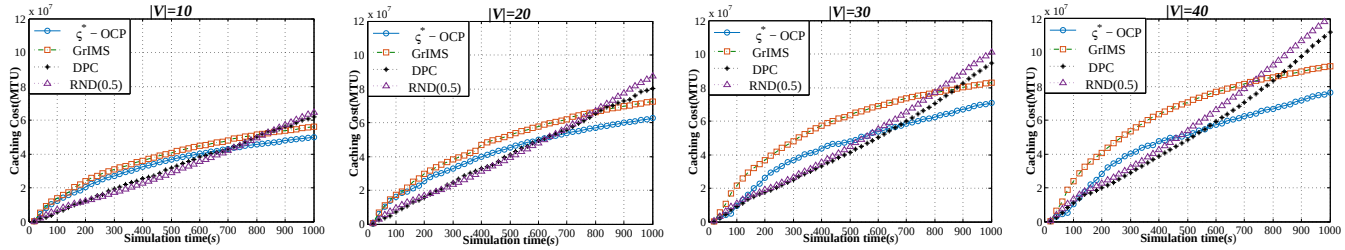
where $\mathbb{N}(t)$ denotes set of nodes received requests till t and $|\mathbb{N}(t)|$ is its cardinality, $H_i^h(t)$ and $H_i(t)$ denote the number of cache hit events and total number of received requests at node i till time t, respectively.

According to Fig. 10, the overall performance of ς^*-OCP is the best among four solutions, i.e., ς^*-OCP is about 10%/20%/16%/5% higher than the best of other three solutions when the size of video set $|V|$ is 10/20/30/40, respectively. GrIMS outperforms DPC in some cases. For example, GrIMS outperforms DPC after 500s in 10 (a) and 700s in 10 (b). RND(0.5) has the worst performance among all the solutions. The superiority of ς^*-OCP is because it has low system load with respect to probabilities solutions according to **Theorem 1** and **2**, which means the demand chunk can be quickly discovered, namely a higher cache hit ratio. Comparing with GrIMS that estimates the request rate from the whole system to configure the global cache resource, DPC determines the caching probability by calculating the local requests rate at each mobile node. Thus, the chunk demand estimation of DPC may be more inaccurate than that of GrIMS and hence DPC results in lower ACHR. The reason that RND(0.5) performs the worst is simply caching all passing content with constant probability not



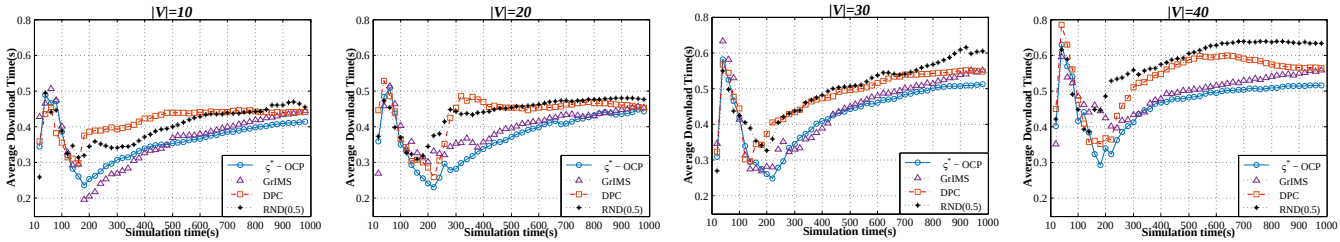
(a) (b) (c) (d)

Fig. 10. Average cache hit ratio vs. simulation time along 4 sizes of video sets: $(a)|V| = 10; (b)|V| = 20; (c)|V| = 30; (d)|V| = 40;$



(a) (b) (c) (d)

Fig. 11. Caching Cost vs. simulation time along 4 sizes of video sets: $(a)|V| = 10; (b)|V| = 20; (c)|V| = 30; (d)|V| = 40;$



(a) (b) (c) (d)

Fig. 12. Average download time vs. simulation time along 4 sizes of video sets: $(a)|V| = 10; (b)|V| = 20; (c)|V| = 30; (d)|V| = 40;$

only ignores the popularity diversity of different content chunks but also neglects the time varying characteristics of content demand.

***Caching Cost (CC)*:** The CC is defined as the total number of caching events and we calculate the CC at time t by following equation:

$$CC(t) = \sum_{s=0}^{t} \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{V}} C_{i,j}(s) \times M$$

where $C_{i,j}(s)$ is an impulse function, which indicates mobile node i cache chunk j at time s when $C_{i,j}(s)$=1 and 0 otherwise. $\mathbb{V}$ denotes the chunk set, M denotes the number of data packets contained by each chunk. $\mathbb{N}$ is the the set of nodes in simulation.

Fig. 11 (a)(b)(c)(d) show the CC comparison with different sizes of the video set, i.e., $|V|$=10, 20, 30 and 40. As shown in figures, ς^*-OCP achieves the best performance among four solutions during the second half of simulation in each figure. Especially, when video set grows, the difference between ς^*-OCP and other heuristic strategies is becoming more and more obvious, i.e., ς^*-OCP outperforms than GrIMS, DPC and RND(0.5) at 1000s by 5%, 22% and 25% in Fig. 11 (a). And this superiority is extended to 20%, 42% and 45% when the video set size reaches 40. The results are consistent with **Theorem 1** which declares that time threshold based solution overwhelms the probabilistic based caching control in terms of caching cost. GrIMS achieves lower caching cost than that of DPC and RND(0.5) when simulation time exceeds 800s. This is mainly because GrIMS allocates the caching resource by globally estimating the chunk demand and node residual caching space, which can alleviate part of unnecessary caching, namely reducing the caching cost. DPC and RND(0.5) have similar performance in terms of the CC since these two methods make caching decision locally, and the average caching probability of all chunks in DPC is also around 0.5.

***Average Download Time (ADT)*:** ADT denotes the average time interval between the time to issue the $Interest$ packet of chunk k and the time to receive the corresponding chunk. This metric reflects the access latency of content, which is an important metric for delay-sensitive application such as video streaming.

Fig. 12 (a), (b), (c) and (d) show that all curves firstly decrease sharply before 200s, then reveal an increasing trend during the rest of simulation. This is because all mobile nodes have enough caching space to ensure that content can be placed to users nearby at the beginning of simulation, which reduces the ADT. However, due to the growing cache hit distance caused by cache miss when local cache is full, the downloading latency as well increases. From Fig. 12, we can also see ς^*-OCP has lowest ADT

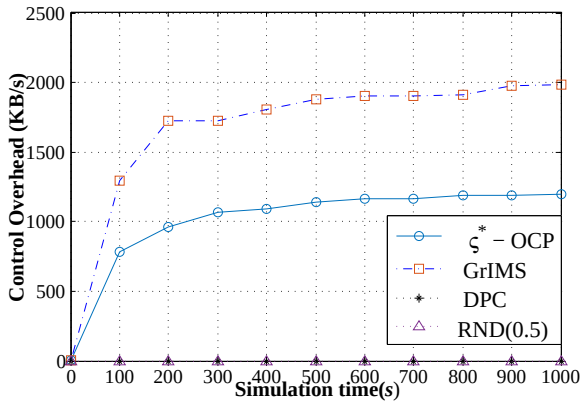


Fig. 13. Control overhead vs. simulation time

among four solutions. GrIMS performs better than other two probabilistic solutions in most cases. The reason can be explained by follows, ς^*-OCP has higher ACHR than other three solutions which is shown in Fig. 10, namely a nearby cache may be hit with higher probability, resulting in lower ADT. Comparing with DPC and RND(0.5), GrIMS can achieve better ACHR by estimating the content supply and demand globally, also resulting lower ADT. However, GrIMS still uses a heuristic solution whose performance is unbounded, hence underperforms the ς^*-OCP. The fixed probability utilized by RND(0.5) has higher frequency of cache miss due to the ignorance of demand variation of different chunks, which in turn leads to long ADT.

***Control Overhead (CO)*:** In our simulation, we count the average occupied bandwidth per second of signalling used to optimize the global caching as the control overhead (CO). In our ς^*-OCP, the CO is mainly the traffic of state list and caching configuration information generated by users and coordinators, respectively. In GrIMS, every node is required to submit the request information and node capacity, the Cloud coordinator determines which to cache for each node individually. The traffic generated by such information and control are considered as the CO of GrIMS. DPC and RND(0.5) configure the caching locally, hence have no CO. As Fig. 13 shows, the CO of ς^*-OCP and GrIMS both experience a growing trend during the simulation. This is mainly because with the increase of simulation time, more users join and begin to request content, which enlarges the CO accordingly. GrIMS performs worse than ς^*-OCP in the sense that corresponding CO is almost 40% higher than that of ς^*-OCP when simulation time is 1000s. The main reason is the coordinator in GrIMS needs to control the caching decision of each node by unicast-based message exchange method. Instead, ς^*-OCP uses broadcasting-based method to control caching in all mobile nodes, hence results in a lower control overhead. Although DPC and RND(0.5) have no CO because they make caching decision according to local information, however, this comes at a price in terms of lower caching hit ratio and higher downloading latency and caching cost.

7 Conclusion and Future Work

This paper studied the problem of optimal caching in ICN 5G D2D environment. We modeled replica dissemination process in ICN 5G D2D as a fluid-based model, which captures the dynamic relationship between content replication and users behavior under controllable caching operations. Furthermore, our formulation has led to an optimal control problem to jointly minimize the caching cost and system load. We then proved the superiority of this time threshold caching control with respect to probabilistic-based methods and existence of a time threshold solution for above optimization problem. Additionally, we also designed a practical caching algorithm named ς^*-OCP which integrate with our time threshold-based optimal caching control. Simulation results showed our ς^*-OCP achieves higher caching hit ratio, lower caching redundancy and delivery latency when comparing with the state-of-art solutions.

Our work also opens some avenues for future work in this field. First, although RWP is an general mobility model, yet is not well suit for vehicular environment where vehicles are moving along the pre-given routes (such as streets). Hence, new models could be introduced in order to provide more comprehensive analysis. Second, in our fluid-based model, we consider two widely used ICN forwarding schemes: random unicast and broadcast, other forwarding strategies such as geographical-based forwarding [22] can be also adopted with some modification on fluid-based model. In this case, except for investigating the caching dynamic, the proposed fluid-based model can be also used to analyze the performance of request forwarding strategies. Third, as our model can be also used to describe the data dissemination under different forwarding strategies, our future work may also include designing efficient forwarding strategies that reduce the delivery latency and forwarding energy costs.

8 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61522103, 61372112; by the H2020 Bonvoyage project under Grant Nos. 635867; by the BUPT Excellent Ph.D. Students Foundation CX2017312.

References

[1] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models," in *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510-527, Feb. 2016.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" in *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065 - 1082, Nov. 2014.

[3] W. Hong; K.-H. Baek, Y. Lee, Y. Kim, and S.-T. Ko, "Study and prototyping of practically large-scale mmWave antenna systems for 5G cellular devices," in *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 63-69, Nov. 2014.

[4] Y. Zhou, and W. Yu, "Optimized Backhaul Compression for Uplink Cloud Radio Access Network," in *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295-1307, June 2014.

[5] A. Asadi, and V. Mancuso, "Network-assisted Outband D2D-clustering in 5G Cellular Networks: Theory and Practice," *IEEE Trans. on Mobile Comput.*, vol.PP, no.99, pp.1-1,Oct, 2016.

[6] G. S. Park, W. Kim, S.-H. Jeong, and H. Song, "Smart Base Station-Assisted Partial-Flow Device-to-Device Offloading System for Video Streaming Services ," *IEEE Trans. on Mobile Compu.*, vol. PP, no. 99, pp.1 - 1, Nov., 2016.

[7] A. Orsino, A. Samuylov, D. Moltchanov, S. Andreev, L. Militano, G. Araniti, Y. Koucheryavy, "Time-Dependent Energy and Resource Management in Mobility-Aware D2D-Empowered 5G Systems," *in IEEE Wirel. Commun.*, vol. 24, no. 4, pp. 14-22, Aug. 2017.
[8] Y. Wu, S. Wang, W. Liu, W. Guo, and X. Chu, "Iunius: A Cross-Layer Peer-to-Peer System With Device-to-Device Communications,"*IEEE Trans. on Wireless Comm.*, vol. 15, no. 10, pp. 7005-7017, July 2016.
[9] C. Xu, S. Jia, L. Zhong, and G. M. Muntean, "Socially aware mobile peer-to-peer communications for community multimedia streaming services,"*IEEE Commun. Mag.*, vol. 53, no. 10, pp. 150-156, Oct. 2015.
[10] A. Ioannou, and S. Weber, "A Survey of Caching Policies and Forwarding Mechanisms in Information-Centric Networking," *IEEE Commun. Surv. Tut.*, vol. 18, no. 4, pp. 2847-2886, May 2016.
[11] H. Liu,Z. Chen, X. Tian, X. Wang and M. Tao, "On content-centric wireless delivery networks," *IEEE Wireless Commun. Mag.*, vol. 21, no. 6, pp. 118-125, Dec. 2014.
[12] L. Zhang, A. Afanasyev, J. Burke, et al. "Named data networking," in *Proc. ACM SIGCOMM Compu. Commun. Rev.*, vol. 44, no. 3 pp. 66-73, Aug. 2014.
[13] K. Wang, F. R. Yu, H. Li, and Z. Li, "Information-Centric Wireless Networks with Virtualization and D2D Communications," *IEEE Wireless Commun.*, vol. PP, no. 99, pp. 2-9, Jan. 2017.
[14] C. Xu; P. Zhang; S. Jia; M. Wang, and G.-M. Muntean, "Video Streaming in Content-Centric Mobile Networks: Challenges and Solutions," *IEEE Wireless Commun.*, vol. PP, no. 99, pp. 2-10, Jan. 2017.
[15] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5g mobile wireless networks," *IEEE Network*, vol. 29, no. 3, pp. 68-74, May. 2015.
[16] A. Morelli, M. Tortonesi, C. Stefanelli, andp N. Suri, "Information-Centric Networking in next-generation communications scenarios", *J. Net. & Compu. App.*, vol. 80, pp. 232-250, 2017.
[17] D. Malak, M. Al-Shalash, and J. G. Andrews "Exploring synergy between communications, caching, and computing in 5G-grade deployments," *IEEE Trans. on Commun.*, vol. 64 no.10, pp.4365-4380, Aug, 2016.
[18] S. Wang, J. Bi, J. Wu, and A. V. Vasilakos, "CPHR: In-Network Caching for Information-Centric Networking With Partitioning and Hash-Routing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2742-2755, Oct. 2016.
[19] H. Wu, J. Li, and J. Zhi, "A Max-Benefit Probability-based caching strategy in Information-Centric Networking," in *Proc. IEEE Conf. Commun.*, vol.15, no. 4, pp. 5646 - 5651, June. 2015.
[20] M. Mangili, F. Martignon, S. Paris, and A. Capone, "Bandwidth and Cache Leasing in Wireless Information-Centric Networks: A Game-Theoretic Study," *IEEE Trans. on Veh. Technol.*, vol. 66, no. 1, pp. 679 - 695, Mar. 2016.
[21] K. Kvaternik, J. Llorca, D. Kilper, and L. Pavel, "A Methodology for the Design of Self-Optimizing, Decentralized Content-Caching Strategies," *IEEE/ACM Trans. on Net*, vol. 24, no. 5, pp. 2634-2647, Oct. 2016.
[22] G. Grassi, D. Pesavento, G. Pau, R. Vuyyuru, R. Wakikawa, and L. Zhang "VANET via Named Data Networking," in*Proc. IEEE Conf. Compu. Commun. Workshops*, pp. 410 - 415, July.2014.
[23] G. Deng, L. Wang, F. Li, and R. Li, "Distributed Probabilistic Caching strategy in VANETs through Named Data Networking," in *Proc. IEEE Conf. Compu. Commun. Workshops* , Sep, 2016.
[24] C. Xu, W. Quan, H. Zhang, L. A. Grieco, et al., "GrIMS: Green Information-centric Multimedia Streaming Framework in Vehicular Ad Hoc Networks," *IEEE Trans. on Cir. & Sys. for Vid. Technol.* , vol.PP, no.99, pp.1-1,Sep, 2016.
[25] C .Xu, W. Quan, V. A. Vasilakos, and H. Zhang, "Information-centric cost-efficient optimization for multimedia content delivery in mobile vehicular networks," *Comput. Commun.*, vol. PP, no. 99, pp. 1-1,2016.
[26] R. Groenevelt, P. Nain, and G. Koole, "The message delay in mobile ad hoc networks," *Perfor. Eva.*, vol. 62, no. 1, pp. 210-228, Oct. 2005.
[27] T. Camp, J. Boleng, B. Williams, L. Wilcox, and W. Navidi, "Performance comparison of two location based routing protocols for ad hoc networks," in *Proc. IEEE Conf. Compu. Commun.*, pp.1678-1687,2002.
[28] Hao Che, Y. Tung, and Z. Wang, "Hierarchical Web caching systems: modeling, design and experimental results, " *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305 - 1314, Nov. 2002.
[29] A. Barrat, M. Barthelemy, and A. Vespignani, "Dynamical processes on complex networks," *Cambridge University* Press, 2008.
[30] X. Fu, Z. Xu, Q. Peng, J.You, L. Fu, X. Wang and S. Lu,"ConMap: A Novel Framework for Optimizing Multicast Energy in Delay-constrained Mobile Wireless Networks" *Proc. ACM MobiHoc 2017*, Chenai, India.
[31] H. Gong, L. Fu , X. Fu, L. Zhao, K. Wang, X. Wang ,"Distributed Multicast Tree Construction in Wireless Sensor Networks," *IEEE Trans. on Info. Theo.*,2016.
[32] ndnSim in NS-3, *http://ndnsim.net/intro.html.*
[33] Network Simulator 3, *https://www.nsnam.org/.*
[34] NGMN 5G White Paper, *http://www.ngmn.org/home.html.*
[35] L. Richard Burden, J. Douglas Faires, and M. Annette Burden, " Numerical Analysis,"*Cengage Learning Press*, 2015.
[36] I. Ullah *et al.*, "A Survey and Synthesis of User Behavior Measurements in P2P Streaming Systems," *IEEE Commun. Surv. Tut.*, vol. 14, no. 3, 2012, pp. 734-749.
[37] H. Smith, "Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems," American Mathematical Society, 2008.
[38] M. Crowder, " Stochastic Approximation: A Dynamical Systems Viewpoint by Vivek S. Borkar, " *International Statistical Review*, vol. 77, no. 2, pp. 306-306, Jul. 2009.

Changqiao Xu [M'04, S'15] (cqxu@bupt.edu.cn) received the Ph.D. degree from Institute of Software, Chinese Academy of Sciences (ISCAS), Beijing, China, in 2009. He was an Assistant Research Fellow with ISCAS from 2002 to 2007. From 2007 to 2009, he was a Researcher with the Software Research Institute, Athlone Institute of Technology, Athlone, Ireland. He joined Beijing University of Posts and Telecommunications, Beijing, in 2009. He is currently a Professor with State Key Laboratory of Networking and Switching Technology, and the Director of the Next Generation Internet Technology Research Center, BUPT. He has authored over 100 technical papers in prestigious international journals and conferences. His research interests include wireless networking, multimedia communications, and next generation Internet technology. Dr. Xu served a Co-Chair and Technical Program Committee (TPC) member for a number of international conferences and workshops. He also served as the Co-Chair of the IEEE MMTC Interest Group, Green Multimedia Communications, and a Board Member of the IEEE MMTC Services and Publicity. He is Senior member of IEEE.

Mu Wang received his M.S. degree in computer technology from Beijing University of Posts and Telecommunications (BUPT) in 2015. He is currently working toward the Ph.D with the Institute of Network Technology, BUPT. His research interests include information centric networking, wireless communications, and multimedia sharing over wireless networks.

Xingyan Chen received the BE degree in Applied Physics from the College of Science, Beijing University of Posts and Telecommunications, in 2016. He is currently working toward the master degree under Prof. Xu at the Next generation Internet Lab. His research interests include information dissemination and content center network.

Lujie Zhong received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2013. She is an Associate Professor with the Information Engineering College, Capital Normal University, Beijing. Her research interests include communication networks, computer system and architecture, mobile Internet technology.

Luigi Alfredo Grieco received the Ph.D. degree in information engineering from the university di Lecce, Lecce, Italy, in December 2003. He has authored more than 100 scientific papers with a significant scientific impact. His main research interests include multimedia communications, quality of service in wireless networks, Internet of things (IoT), and future Internet. He is the Founder Editor-in-Chief of the Internet Technology Letters journal (Wiley) and serves as EiC of the WIley Transactions on Emerging Telecommunications Technologies and as an Editor of the IEEE Trans. on Vehicular Technology. Within the Internet Engineering Task Force and Internet Research Task Force, he is actively contributing to the definition of new standard protocols for industrial IoT applications and new standard architectures for tomorrow Information Centric Networking (ICN)-IoT systems.