Mixed Timescale Online PHY Caching and Content Delivery for Content-Centric Wireless Networks

An Liu[1], Vincent Lau[1], Wenchao Ding[1] and Edmund Yeh[2]
[1]Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology
[2]Department of Electrical and Computer Engineering, Northeastern University

***Abstract*—In content-centric wireless networks, physical layer (PHY) caching has been proposed to exploit the *dynamic side information* induced by base station (BS) cache to support Coordinated Multi-Point (CoMP) and achieve huge capacity gain. The performance of PHY caching depends heavily on the cache content placement algorithm. In existing algorithms, the cache content placement is adaptive to the long-term popularity distribution in an offline manner. We propose an online PHY caching framework based on the concept of virtual interest packet (VIP) in a virtual network. The VIP captures *microscopic spatial and temporary popularity variations*, and thus the VIP-based online PHY caching can adapt the cached content to the microscopic popularity variations to fully exploit the benefits of PHY caching. The joint optimization of online caching and content delivery is formulated as a mixed timescale drift minimization problem and a low complexity algorithm is proposed to find the optimal solution. Simulations show that the proposed solution achieves large gain over existing solutions.**

***Index Terms*—Content-centric wireless networks, Online PHY caching, CoMP, Resource Control**

I. Introduction

In content-centric wireless networks, the cache resource can be exploited to substantially improve the performance of content delivery [1], [2]. Recently, cache-enabled opportunistic CoMP (PHY caching) is proposed in [3], [4] to improve the spectral efficiency of wireless networks with limited backhaul. Specifically, when the requested content of users exists in the cache of several BSs, the BS caches induce *dynamic side information* to the BSs, which can be used to cooperatively transmit the requested packets to the users and thus achieving huge DoF gains. In practice, the cache storage capacity is limited and hence, the cache content placement algorithm plays a key role in determining the overall performance of the PHY caching schemes. The existing cache content placement algorithms can be classified into the following two types.

The *offline cache content placement* is adaptive to the long-term popularity distribution [3], [4]. Once the cache content placement phase is finished, the cached content cannot be changed during the content delivery phase. As a result, these offline caching algorithms cannot capture the *microscopic popularity variations* that occur in practice.

The *online cache content placement* is adaptive to the microscopic popularity variations. It has more refined control on the limited cache resource and thus can achieve a better performance. In [5], a VIP-based online caching framework is proposed for wired Name Data Networks (NDN). However, the solution in [5] cannot be applied to wireless networks with cache-induced CoMP.

This work was supported by the NSFC under Project No. 61571383.

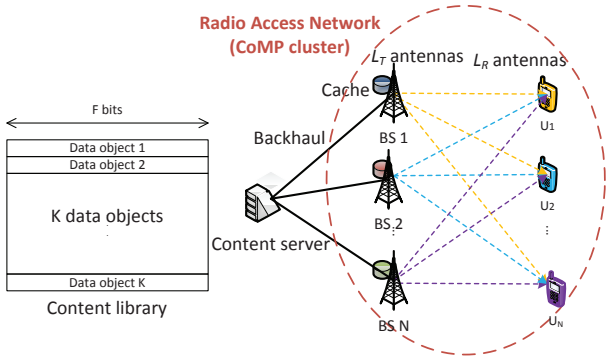


Figure 1: Architecture of cached MIMO interference networks.

In this paper, we propose an online PHY caching framework to fully exploit the benefits of cache-induced opportunistic CoMP. The main contributions are summarized as follows.

- **Dual-mode-VIP-based Online PHY Caching with Cache Placement Cost:** In the VIP-based online caching framework in [5], the cache content placement cost is ignored. However, for practical consideration, it is very important to model the cache content placement cost because cache content update will incur communication overhead. Moreover, the existing VIP-based framework cannot support the CoMP transmission mode in wireless networks. We propose a new dual-mode-VIP-based online PHY caching framework which considers cache content placement cost and can support the CoMP mode to enhance the capacity of radio access network (RAN).
- **Mixed Timescale Optimization of Online PHY Caching and Content Delivery:** Due to the CoMP and slow timescale cache content placement, the joint optimization of online PHY caching and content delivery is formulated as a mixed timescale drift minimization problem and the existing solutions cannot be applied. We propose a low complexity algorithm to find the optimal solution by exploiting specific structure of the problem.

II. System Model

A. Network Architecture

Consider a cached MIMO interference network with N BS-user pairs, as illustrated in Fig. 1. Each BS has L_T antennas

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

and each user has L_R antennas. Each BS has transmit power P and a cache with storage capacity of $L_C F$ bits. There is a content server providing a content library $\mathcal{K}$ that contains K data objects. For simplicity, we assume that all data objects have the same size of F bits. The users request data objects from the content server via the RAN. Each BS in the RAN is connected to the content server via a backhaul. The content server also serves as a central node for resource control.

For convenience, let $\mathcal{B}$ denote the set of BSs, $\mathcal{U}$ denote the set of users, $\mathcal{N} = \mathcal{B} \cup \mathcal{U}$ denote the set of all nodes, and g denote the content server. The serving BS of user $j \in \mathcal{U}$ is denoted as n_j and the associated user of BS $n \in \mathcal{B}$ is denoted as j_n. A user always sends its data object request to its serving BS. However, it may receive the requested data object from the serving BS only or all BSs depending on the PHY mode as will be elaborated in the next subsection.

Time is divided into frames indexed by i, and each frame consists of T time slots. The fast timescale resource control (such as PHY mode selection and rate allocation) is updated at the beginning of each time slot. The slow timescale cache content placement control is updated at the beginning of each frame. Unless otherwise specified, t is used to index a time slot in frame i, i.e., $t \in [1+(i-1)T, iT]$ and $i = \lceil \frac{t}{T} \rceil$.

B. Dual Mode Physical Layer Model

We assume that the content server has knowledge of the global channel state information (CSI) $\boldsymbol{H}(t) = \{\boldsymbol{H}_{jn}(t), \forall j \in \mathcal{U}, n \in \mathcal{B}\}$ of the RAN, where $\boldsymbol{H}_{jn}(t) \in \mathbb{C}^{L_R \times L_T}$ is the channel matrix between the n-th BS and the j-th user at the t-th time slot, and $\boldsymbol{H}(t)$ is quasi-static within a time slot and i.i.d. between time slots. The time index t in $\boldsymbol{H}(t)$ will be omitted when there is no ambiguity. There are two PHY modes as elaborated below.

CoMP transmission mode (PHY Mode A): In this mode, the BSs form a virtual transmitter and cooperatively serve all users. The RAN is a virtual MIMO broadcast channel (BC). Specifically, let $B_j(t)$ denote the data scheduled for delivery to user j at time slot t. In CoMP mode, $B_j(t)$ must be available at all BSs. The amount of scheduled data $|B_j(t)|$ (measured by the number of data objects) is limited by the data rate of the CoMP mode PHY, i.e., $|B_j(t)| = c_j^A(t)$, where $c_j^A(t)$ (data objects/slot) is the data rate of user j under CoMP mode with CSI $\boldsymbol{H}(t)$. The set of achievable rate vectors $\boldsymbol{c}^A(t) = \left[c_j^A(t)\right]_{j\in\mathcal{U}} \in \mathbb{R}_+^N$ forms the capacity region $C^A(\boldsymbol{H}(t)) \in \mathbb{R}_+^N$ under the CoMP transmission mode.

Coordinated transmission mode (PHY Mode B): In this case, user j can only be served by the serving BS n_j and the RAN is a MIMO interference channel (IFC). Similarly, we have $|B_j(t)| = c_j^B(t)$, where $c_j^B(t)$ (data objects/slot) is the data rate of user j under coordinated mode with CSI $\boldsymbol{H}(t)$. The set of achievable rate vectors $\boldsymbol{c}^B(t) = \left[c_j^B(t)\right]_{j\in\mathcal{U}} \in \mathbb{R}_+^N$ forms the capacity region $C^B(\boldsymbol{H}(t)) \in \mathbb{R}_+^N$ under the coordinated transmission mode.

There exist many CoMP/coordinated transmission schemes. In this paper, we do not restrict the PHY to be any specific CoMP/coordinated transmission scheme but consider an abstract PHY model represented by the capacity regions $C^A(\boldsymbol{H})$ and $C^B(\boldsymbol{H})$. Note that $C^A(\boldsymbol{H})$ and $C^B(\boldsymbol{H})$ depend on the transmit power P at each BS. Moreover, we assume that $C^B(\boldsymbol{H}) \subseteq C^A(\boldsymbol{H})$.

III. Mixed Timescale Online PHY Caching and Content Delivery Scheme

In this section, we first present an online PHY caching scheme that determines which data objects are cached/removed at each frame in an online manner. Then we propose a dual mode content delivery scheme in which the BSs obtain the requested data objects from either the local cache or backhaul and send them to the users using the dual mode PHY.

A. Slow Timescale Online PHY Caching Scheme

In the proposed online PHY caching scheme, the cached data objects at each BS are updated once every frame (T time slots). Since the local popularity variations at each BS usually changes at a timescale much slower than the instantaneous CSI (slot interval), in practice, we may choose $T \gg 1$ to reduce the cache content placement cost without losing the ability to track microscopic popularity variations.

Let $s_n^k(i) \in \{0,1\}$ denote the *cache state* of data object k at BS n, where $s_n^k(i) = 1$ means that data object k is in the cache of BS n at frame i and $s_n^k(i) = 0$ means the opposite. The *cache placement control action* at the beginning of the i-th frame is denoted by $\left\{p_n^k(i) \in \{-1,0,1\}, \forall n \in \mathcal{B}, k \in \mathcal{K}\right\}$, where $p_n^k(i) = -1$ and $p_n^k(i) = 1$ mean that the data object k is removed from and added to the cache of BS n at the beginning of the i-th frame respectively, and $p_n^k(i) = 0$ means that the cache state is unchanged. Then the cache state dynamics is

$$s_n^k(i) = \left[\min\left(s_n^k(i-1) + p_n^k(i), 1\right)\right]^+, \forall n \in \mathcal{B}, k \in \mathcal{K}.$$

Note that the cache placement control action $\left\{p_n^k(i)\right\}$ must satisfy the following cache size constraint:

$$\sum_{k\in\mathcal{K}} \left[\min\left(s_n^k(i-1) + p_n^k(i), 1\right)\right]^+ \le B_C, \forall n \in \mathcal{B}. \quad (1)$$

Moreover, there is no need to add an existing data object in the cache, i.e.,

$$p_n^k(i) \neq 1, \text{when } s_n^k(i-1) = 1. \quad (2)$$

Let $\boldsymbol{s}(i) = \left\{s_n^k(i), \forall n \in \mathcal{B}, k \in \mathcal{K}\right\}$ denote the *aggregate cache state*. When $p_n^k(i) = 1$, BS n needs to obtain data object k from the backhaul, which induces some *cache content placement cost*. To accommodate the traffic caused by cache content placement, the available backhaul capacity R (data objects/slot) at each BS is divided into a *data sub-channel* and a *control sub-channel* as $R = R_c + R_d$, where the data sub-channel with rate R_d (data objects/slot) is used for transmitting the data objects requested by users, and the control sub-channel is used for transmitting the data objects induced by the cache placement control action $p_n^k(i)$ as well as other

control signalings. The cache content placement cost for BS n to cache data object k at frame i is given by

$$\Gamma_n^k(i) = \gamma \mathbf{1}_{\{p_n^k(i)=1\}},$$

where $\mathbf{1}$ denotes the indication function, and γ is the price of transmitting one data object using the control sub-channel. The total cost function at frame i is then given by

$$\Gamma(i) = \sum_{n\in\mathcal{B},k\in\mathcal{K}} \Gamma_n^k(i) = \sum_{n\in\mathcal{B},k\in\mathcal{K}} \gamma \mathbf{1}_{\{p_n^k(i)=1\}},$$

and the average cache content placement cost is

$$\overline{\Gamma} = \limsup_{J\to\infty} \frac{1}{J} \sum_{i=1}^{J} \mathbb{E}\left[\Gamma\left(i\right)\right].$$

B. Fast Timescale Dual Mode Content Delivery Scheme

We consider a *dual mode content delivery scheme* which can support the CoMP mode to enhance the capacity of RAN. Specifically, each data object is divided into D data chunks and each data chunk is allocated with an unique ID. The content delivery operates at the level of data chunks using two types of packets: *Interest Packets* (IPs) and *Data Packet*s (DPs). To request a data chunk, a user sends out an *Interest Packet* (IP), which carries the ID of the desired data chunk, to the content server (via its serving BS). Therefore, a request for a data object consists of a sequence of IPs which request all the data chunks of the object, where the sequence starts with the IP requesting the starting chunk, and ends with the IP requesting the ending chunk. For each IP from user j, the content server determines its mode (*Coordinated Mode IP or CoMP Mode IP*) according to an *IP mode selection* policy. If it is marked as a CoMP Mode IP, the content server will ensure that the corresponding DP is delivered to all BSs. If it is marked as a Coordinated Mode IP, the content server will ensure that the corresponding DP is delivered to the serving BS n_j only. Upon receiving a *CoMP Mode DP* of user j at BS n, BS n will store it in a *CoMP mode data buffer* with queue length denoted by Q_{nj}^A. On the other hand, upon receiving a Coordinated Mode DP of user j at the serving BS n_j, BS n_j will store it in a *coordinated mode data buffer* with queue length $Q_{n_jj}^B$. Note that the IP mode selection policy only determines whether the corresponding DP is stored in the CoMP mode data buffers of all BSs (CoMP mode) or is stored in the coordinated mode data buffer of the serving BS (coordinated mode). At each time slot, the content server still needs to determine the PHY mode $M_a\left(t\right) \in \{0,1\}$ according to a *PHY mode selection* policy. If $M_a\left(t\right) = 0$ ($M_a\left(t\right) = 1$), the BSs will employ the coordinated (CoMP) transmission mode to transmit some data from the coordinated (CoMP) mode data buffers to the users. The mode selection policies will be elaborated in Section IV-C.

The dual mode content delivery has four components.

Component 1 (IP mode selection at the content server): Let $Da_j^k(t)$ denote the number of IPs of data object k received by the content server from user j at time slot t, where $a_j^k(t)$ can be interpreted as the instantaneous arrival rate of IPs in the unit of data object/slot since each data object corresponds to D IPs. The content server will mark all these $Da_j^k(t)$ IPs using the same *IP mode* denoted by $m_j^k\left(t\right) \in \{0,1\}$.

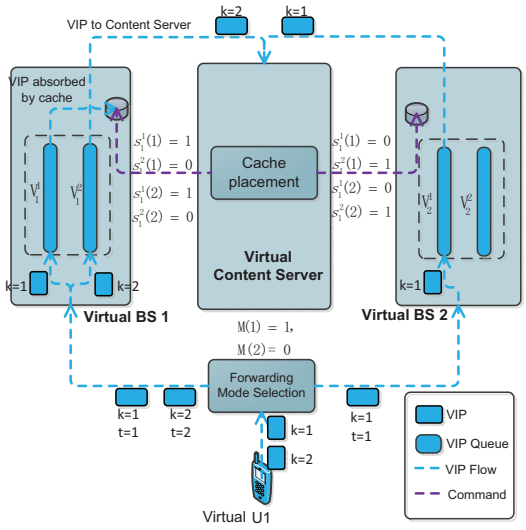
Component 2 (Coordinated mode DPs delivery to each BS): If $m_j^k\left(t\right) = 0$ (coordinated mode), the corresponding $Da_j^k(t)$ DPs are called *coordinated mode DPs* which will be delivered to the serving BS n_j only. Specifically, if BS n_j has data object k in the local cache (i.e., $s_{n_j}^k\left(i\right) = 1$), it creates $Da_j^k(t)$ DPs containing the requested data chunks indicated by the $Da_j^k(t)$ IPs. Otherwise ($s_{n_j}^k\left(i\right) = 0$), the content server will create $Da_j^k(t)$ DPs containing the requested data chunks and store them in the n_j-th data buffer with queue length Q_{gn_j} at the content server, which will be send to BS n_j via backhaul when they become the head-of-the-queue DPs. In both cases, after obtaining the $Da_j^k(t)$ DPs, BS n_j will store them in the coordinated mode data buffer $Q_{n_jj}^B$.

Component 3 (CoMP mode DPs delivery to each BS): If $m_j^k\left(t\right) = 1$ (CoMP mode), the corresponding $Da_j^k(t)$ DPs are called *CoMP mode DPs* which will be delivered to all BSs. Specifically, for any $n \in \mathcal{B}$, if BS n has data object k in the local cache (i.e., $s_n^k\left(i\right) = 1$), it creates $Da_j^k(t)$ DPs containing the requested data chunks. Otherwise ($s_n^k\left(i\right) = 0$), the content server will create $Da_j^k(t)$ DPs containing the requested data chunks and store them in the n-th data buffer Q_{gn} at the content server, which will be send to BS n via backhaul when they become the head-of-the-queue DPs. In both cases, after obtaining the $Da_j^k(t)$ DPs, BS n will store the $Da_j^k(t)$ DPs in the j-th CoMP mode data buffer Q_{nj}^A.

Component 4 (PHY mode determination at the content server): At time slot t, the content server first determines the *PHY mode* $M_a\left(t\right)$. If $M_a\left(t\right) = 0$, coordinated transmission mode will be used to send the data in $Q_{nj_n}^B$ to user j_n, $\forall j_n \in \mathcal{U}$, at rate $c_{j_n}^B\left(t\right)$ (data objects/slot). If $M_a\left(t\right) = 1$, CoMP transmission mode will be used to send the data in Q_{nj}^A to user j, $\forall j \in \mathcal{U}$, at rate $c_j^A\left(t\right)$ (data objects/slot). For convenience, let $\boldsymbol{c}^B\left(t\right) = \left[c_j^B\left(t\right)\right]_{j\in\mathcal{U}} \in \mathbb{R}_+^N$ and $\boldsymbol{c}^A\left(t\right) = \left[c_j^A\left(t\right)\right]_{j\in\mathcal{U}} \in \mathbb{R}_+^N$ denote the *PHY rate allocation* for coordinated and CoMP transmission modes respectively.

IV. Dual-Mode-VIP-based Resource Control

In the proposed scheme, the slow timescale control includes the cache content placement policy $\left\{p_n^k(i)\right\}$ which is adaptive to the spatial and temporal variations of the content popularity. The fast timescale controls include the IP mode selection $\left\{m_j^k\left(t\right)\right\}$, backhaul rate allocation $\{c_{ng}(t)\}$, PHY mode selection $M_a\left(t\right)$, and PHY rate allocation $\left\{\boldsymbol{c}^A\left(t\right), \boldsymbol{c}^B\left(t\right)\right\}$, which are adaptive to the cache state $\left\{s_n^k(i)\right\}$ and global CSI $\boldsymbol{H}\left(t\right)$. It is challenging to design the resource control policy in the actual network because the DPs of different data objects are mixed in DP queues, and thus the DP queue dynamics cannot reflect the popularity variations. To overcome this challenge, we consider a *dual mode VIP framework* which transforms the original network into a virtual network and formulate the resource control design in the virtual network.



U1 request data object 1 at time slot 1 and data object 2 at time slot 2
→ forwarding mode selection. Time slot 1 CoMP mode, slot 2 Coordinated mode.
→ k=1 VIP is forwarded to both BSs, and k=2 VIP is forwarded to BS1 only
→ BSs queue the VIP based on data object index
→ BS 1 forwards VIP of data object 2 to content server since data object 2 is not stored at BS1 at time slot 1

Figure 2: Illustration of the VIP flow in the virtual network.

A. Transformation to Virtual Network

The dual-mode VIP framework relies on the concept of Virtual Interest Packet (VIP) flowing over a *virtual network*, as illustrated in Fig. 2. The virtual network is simulated at the content server and it has exactly the same topology, cache state $\left\{s_n^k(i)\right\}$ and global CSI $\boldsymbol{H}\left(t\right)$ as the actual network. Whenever user j requests the starting chunk of data object k in the actual plane, a VIP of data object k will be generated by virtual user j in the virtual network. The VIPs generated by the virtual users are forwarded to the virtual BSs, where depending on the cache state, some VIPs are removed by the local cache and the remaining VIPs are forwarded to the virtual content server, as illustrated in Fig. 2. Each virtual node $n \in \mathcal{N}$ maintains a VIP queue $V_n^k(t)$ for each data object k, which is implemented as a counter in the content server. The VIP queue $V_n^k(t)$ captures the local popularity at each (virtual) node, and the set of all VIP queues $\mathbf{V}(t) = \left\{V_n^k(t), \forall n \in \mathcal{N}, k \in \mathcal{K}\right\}$ captures microscopic popularity variations.

Initially, all VIP queues are set to 0, i.e., $V_n^k(1) = 0, \forall n \in \mathcal{N}, k \in \mathcal{K}$. As the content server receives data object requests (IPs requesting the starting chunk of data objects) from users, the corresponding VIP queues $V_j^k(t), j \in \mathcal{U}$ are incremented accordingly. After some amount of VIPs in $V_j^k(t), j \in \mathcal{U}$ being "forwarded" to the virtual BSs (in the virtual network), the VIP queues $V_j^k(t), j \in \mathcal{U}$ are decreased and the VIP queues $V_n^k(t), n \in \mathcal{B}$ are increased by the same amount accordingly. Similarly, after some amount of VIPs in $V_n^k(t), n \in \mathcal{B}$ being "forwarded" to the virtual content server (content source) and local cache, the VIP queues $V_n^k(t), n \in \mathcal{B}$ are decreased by the same amount accordingly. Specifically, there are two modes for "forwarding" the VIPs from the virtual users to virtual BSs, corresponding to the two PHY modes in the actual plane. In the *CoMP forwarding mode*, VIPs in $V_j^k(t)$ are forwarded to all virtual BSs, and thus at time $t+1$, the VIP queues become

$$V_j^k(t+1) = \left(V_j^k(t) - \mu_j^{Ak}(t)\right)^+ + A_j^k(t),$$
$$V_n^k(t+1) = \left(\left(V_n^k(t) - \mu_{ng}^k(t)\right)^+ + \sum_{j\in\mathcal{U}} \mu_j^{Ak}(t) - r_n s_n^k\left(i\right)\right)^+$$

$\forall j \in \mathcal{U}, n \in \mathcal{B}$, where $A_j^k(t)$ is the number of exogenous data object request arrivals at the VIP queue $V_j^k(t)$ during slot t, $\mu_j^{Ak}(t)$ is the allocated transmission rate of VIPs for data object k from virtual user j to all virtual BSs during time slot t with CoMP forwarding mode, $\mu_{ng}^k(t)$ is the allocated transmission rate of VIPs for data object k from virtual BS n to the virtual content server during time slot t, and r_n is the maximum rate at which BS n can produce copies of cached object k (e.g., the maximum rate r_n may reflect the I/O rate of the storage disk). On the other hand, in the coordinated forwarding mode, VIPs in $V_j^k(t)$ are "forwarded" to the serving virtual BS n_j only, and thus at time $t+1$, the VIP queues become

$$V_j^k(t+1) = \left(V_j^k(t) - \mu_j^{Bk}(t)\right)^+ + A_j^k(t),$$
$$V_n^k(t+1) = \left(\left(V_n^k(t) - \mu_{ng}^k(t)\right)^+ + \mu_{j_n}^{Bk}(t) - r_n s_n^k\left(i\right)\right)^+,$$

$\forall j \in \mathcal{U}, n \in \mathcal{B}$, where $\mu_j^{Bk}(t)$ is the allocated transmission rate of VIPs for data object k from virtual user j to virtual BS n_j during time slot t with coordinated forwarding mode. Combining the above two cases, the VIP queue dynamics can be expressed in a compact form as:

$$V_j^k(t+1) = \left(V_j^k(t) - \mu_j^{Ak}(t)M\left(t\right) - \mu_j^{Bk}(t)\overline{M}\left(t\right)\right)^+ + A_j^k(t)$$
$$V_n^k(t+1) = \Bigg(\left(V_n^k(t) - \mu_{ng}^k(t)\right)^+ + \sum_{j\in\mathcal{U}} \mu_j^{Ak}(t)M\left(t\right) + \mu_{j_n}^{Bk}(t)\overline{M}\left(t\right) - r_n s_n^k\left(i\right)\Bigg)^+, \forall j \in \mathcal{U}, n \in \mathcal{B}, k \in \mathcal{K}$$

where $M\left(t\right) \in \{0,1\}$ is the forwarding mode at time slot t in the virtual plane ($M\left(t\right) = 0$ stands for the coordinated forwarding mode and $M\left(t\right) = 1$ stands for the CoMP forwarding mode), and $\overline{M}\left(t\right) = \mathbf{1}_{M(t)=0}$.

B. Mixed Timescale Resource Control Algorithm

1) Slow Timescale Cache Content Placement Solution: Our algorithm design relies on the Lyapunov optimization framework [6]. Define $\mathcal{L}(\mathbf{V}(t)) \triangleq \sum_{n\in\mathcal{N},k\in\mathcal{K}} \left(V_n^k(t)\right)^2$ as the Lyapunov function, which is a measure of unsatisfied requests in the network. The slow timescale cache content placement $\left\{p_n^k(i)\right\}$ is designed to minimize the *T-step VIP-drift-plus-penalty* defined as

$$\Delta_T\left(i\right) \triangleq \mathbb{E}\left[\mathcal{L}(\mathbf{V}(t_0^i+T)) - \mathcal{L}(\mathbf{V}(t_0^i))|\boldsymbol{X}(t_0^i)\right] + W\mathbb{E}\left[\sum_{n\in\mathcal{B},k\in\mathcal{K}} \gamma 1_{\{p_n^k(i)=1\}}|\boldsymbol{X}(t_0^i)\right],$$

Algorithm 1 Slow timescale cache content placement solution at frame i

For each base station n,

- Let $\mathcal{C}=\left\{k|s_n^k(i-1)=1,k\in\mathcal{K}\right\}$.
- Let $\mathcal{C}'=\left\{k|s_n^{k\star}(i)=1\right\}$, where $\left\{s_n^{k\star}(i)\right\}$ is the optimal solution of $\max_{\left\{s_n^k\right\}}\sum_{k\in\mathcal{K}}V_n^k s_n^k$, $\sum_{k\in\mathcal{K}}s_n^k\le B_C$.
- Let $\mathcal{O}'=\mathcal{C}'/\mathcal{C}$, $\mathcal{O}=\mathcal{C}/\mathcal{C}'$.

  Sort the queue backlogs as, $V_n^{k_1^{'}}\ge\cdots\ge V_n^{k_{|\mathcal{O}'|}^{'}}$, $0\le 0\le\cdots\le V_n^{k_{|\mathcal{O}'|-|\mathcal{O}|+1}}\le\cdots\le V_n^{k_{|\mathcal{O}'|}}$.
- For $i=1:(|\mathcal{O}'|-|\mathcal{O}|)^+$, if $V_n^{k_i^{'}}r_n\ge\frac{W}{2T}\gamma$, then let $p_n^{k_i^{'}}(i)=1$.
- For $i=(|\mathcal{O}'|-|\mathcal{O}|)^+ +1:|\mathcal{O}'|$, if $\left(V_n^{k_i^{'}}-V_n^{k_i}\right)r_n\ge\frac{W}{2T}\gamma$, then let $p_n^{k_i^{'}}(i)=1$ and $p_n^{k_i}(i)=-1$.

where t_0^i is the starting time slot of the i-th frame, and $\boldsymbol{X}(t_0^i)=\left[\mathbf{V}(t_0^i)\,\boldsymbol{s}(i-1)\right]$ is the observed system state at the beginning of the frame. Intuitively, if the first term in $\Delta_T(i)$ is negative, the VIP lengths tend to decrease. On the other hand, the second term in $\Delta_T(i)$ is the weighted cache content placement cost, and W is a price factor. Therefore, minimizing $\Delta_T(i)$ helps to strike a balance between stabilizing the VIP queues and cache content placement cost. Following similar analysis as in proof of Lemma 3 in [6], we obtain an upper bound of $\Delta_T(i)$.

Theorem 1 (T-step Drift-Plus-Penalty Upper Bound)**.** *An upper bound of $\Delta_T(i)$ is given by $\widetilde{\Delta}_T(i)\triangleq\mathbb{E}\left[\Delta_T^U(i)\,|\boldsymbol{X}(t_0^i)\right]$, where*

$$\Delta_T^U(i)=W\sum_{n\in\mathcal{B},k\in\mathcal{K}}\gamma 1_{\{p_n^k(i)=1\}}+\overline{\Delta}-2T\sum_{n\in\mathcal{B},k\in\mathcal{K}}V_n^k(t_0^i)r_n\min\left(s_n^k(i-1)+p_n^k(i),1\right)^+.$$

and $\overline{\Delta}$ is a term independent of $\left\{p_n^k(i)\right\}$.

The slow timescale drift minimization problem is given by

$$\min_{\left\{p_n^k(i)\right\}}\Delta_T^U(i)\ \textit{s.t.}\ (1-2)\ \textit{are satisfied}. \tag{3}$$

The detailed steps to find the optimal solution of (3) are summarized in Algorithm 1, which only has linear complexity w.r.t. the number of data objects K. In Algorithm 1, $V_n^k=V_n^k\left(t_0^i\right)$, $\mathcal{C}$ is the set of currently cached data objects at BS n, $\mathcal{C}'$ is a set of B_C data objects with the highest VIP counts (popularity), $\mathcal{O}'=\mathcal{C}'/\mathcal{C}$ is the set of the most popular data objects which have not been cached, $\mathcal{O}$ is the set of currently cached data objects which are not in $\mathcal{C}'$. Each data object $k_i^{'}$ in $\mathcal{O}'$ will be added in the cache (i.e., $p_n^{k_i^{'}}(i)=1$) if the benefit of caching it (indicated by the backlog difference $\left(V_n^{k_i^{'}}-V_n^{k_i}\right)r_n$) exceeds the cache content placement cost threshold $\frac{W}{2T}\gamma$. If data object $k_i^{'}$ is added in the cache and $i>(|\mathcal{O}'|-|\mathcal{O}|)^+$, data object k_i in $\mathcal{O}$ will be removed (i.e., $p_n^{k_i}(i)=-1$) to save space for caching data object $k_i^{'}$.

Algorithm 2 Fast timescale control solution at slot t

1. Backhaul rate allocation

Let $\mu_{ng}^k(t)=\begin{cases}R_d & k=k_n^*(t)\\ 0 & otherwise\end{cases},\forall n\in\mathcal{B}$, where $k_n^*(t)\triangleq\arg\max_k V_n^k(t)$.

2. Forwarding mode selection and rate allocation

Let

$$\left\{\mu_j^{Ak*}(t)\right\}=\underset{\left\{\mu_j^{Ak}(t)\right\}}{\text{argmax}}\sum_{j\in\mathcal{U},k\in\mathcal{K}}\mu_j^{Ak}(t)\left(V_j^k(t)-\sum_{n\in\mathcal{B}}V_n^k(t)\right)\quad\text{s.t. }\boldsymbol{\mu}^A(t)\in C^A(\boldsymbol{H}) \tag{6}$$

$$\left\{\mu_j^{Bk*}(t)\right\}=\underset{\left\{\mu_j^{Bk}(t)\right\}}{\text{argmax}}\sum_{j\in\mathcal{U},k\in\mathcal{K}}\mu_j^{Bk}(t)\left(V_j^k(t)-V_{n_j}^k(t)\right)\quad\text{s.t. }\boldsymbol{\mu}^B(t)\in C^B(\boldsymbol{H}) \tag{7}$$

Let $\Delta_1^A=\sum_{j\in\mathcal{U},k\in\mathcal{K}}\mu_j^{Ak*}(t)\left(V_j^k(t)-\sum_{n\in\mathcal{B}}V_n^k(t)\right)$ and $\Delta_1^B=\sum_{j\in\mathcal{U},k\in\mathcal{K}}\mu_j^{Bk*}(t)\left(V_j^k(t)-V_{n_j}^k(t)\right)$.

If $\Delta_1^A\ge\Delta_1^B$, **let** $M(t)=1$, $\left\{\mu_j^{Ak}(t)\right\}=\left\{\mu_j^{Ak*}(t)\right\}$, $\mu_j^{Bk}(t)=0,\forall j,k$;

Else, **let** $M(t)=0$, $\mu_j^{Ak}(t)=0,\forall j,k$, $\left\{\mu_j^{Bk}(t)\right\}=\left\{\mu_j^{Bk*}(t)\right\}$.

2) Fast Timescale Control Solution : Similarly, the fast timescale control solution$\{M(t)\}$ and $\left\{\mu_j^{Ak}(t),\mu_j^{Bk}(t),\mu_{ng}^k(t)\right\}$ is obtained by solving the following 1-*step VIP-drift minimization problem:*

$$\min_{\substack{\left\{M(t),\mu_{ng}^k(t)\right\}\\ \left\{\mu_j^{Ak}(t),\mu_j^{Bk}(t)\right\}}}\ M(t)\sum_{j\in\mathcal{U},k\in\mathcal{K}}\mu_j^{Ak}(t)\left(\sum_{n\in\mathcal{B}}V_n^k(t)-V_j^k(t)\right)+\overline{M}(t)\sum_{j\in\mathcal{U},k\in\mathcal{K}}\mu_j^{Bk}(t)\left(V_{n_j}^k(t)-V_j^k(t)\right)-\sum_{n\in\mathcal{B},k\in\mathcal{K}}V_n^k(t)\mu_{ng}^k(t) \tag{4}$$

$$\text{s.t.}\sum_{k\in\mathcal{K}}\mu_{ng}^k(t)\le R_d,\forall n;\boldsymbol{\mu}^A(t)\in C^A(\boldsymbol{H});\boldsymbol{\mu}^B(t)\in C^B(\boldsymbol{H}) \tag{5}$$

where $\boldsymbol{\mu}^A(t)=\left[\sum_{k\in\mathcal{K}}\mu_j^{Ak}(t)\right]_{j\in\mathcal{U}}\in\mathbb{R}_+^N$, $\boldsymbol{\mu}^B(t)=\left[\sum_{k\in\mathcal{K}}\mu_j^{Bk}(t)\right]_{j\in\mathcal{U}}\in\mathbb{R}_+^N$. Note that (5) is the link capacity constraint in the virtual plane. The detailed steps to find the optimal solution of (4) are summarized in Algorithm 2. In step 1 and 2, we need to solve two weighted sum-rate maximization (WSRM) problems in MIMO BC and IFC, respectively. There are many existing WSRM algorithms and the details are omitted for conciseness.

C. Virtual-to-Actual Control Policy Mapping

In the following, we propose a *virtual-to-actual control policy mapping* which can generate a resource control policy for the actual network from that in the virtual network.

Mapping for cache placement control policy $\left\{p_n^k(i)\right\}$: The cache placement control action in the actual network is the same as that in the virtual network.

Mapping for IP mode selection policy $\left\{m_j^k(t)\right\}$: The content server maintains a set of *virtual CoMP queues* whose dynamics are

$$U_j^{Ak}(t+1) = U_j^{Ak}(t) - m_j^k(t)\, a_j^k(t) + M(t)\, \mu_j^{Ak}(t), \forall j, k.$$

Then the forwarding mode at time slot t in the actual plane is

$$m_j^k(t) = \mathbf{1}_{U_j^{Ak}(t+1)>0}, \forall t.$$

Mapping for PHY mode selection and rate allocation policy: For the PHY mode selection and rate allocation policy in the actual plane, we simply let $c_{ng}(t) = \sum_{k\in\mathcal{K}} \mu_{ng}^k(t)$, $M_a(t) = M(t)$, $\boldsymbol{c}^A(t) = \boldsymbol{\mu}^A(t)$ and $\boldsymbol{c}^B(t) = \boldsymbol{\mu}^B(t)$.

V. Simulation Results

Consider a cached MIMO interference network with 7 BS-user pairs placed in 7 wrapped-around hexagon cells. Each BS is equipped with 2 antennas and each user is equipped with 1 antenna. The backhaul capacity per BS is 30Mbps. The channel bandwidth is 10MHz, the slot size is 2ms and the frame size is 0.5s. The pathloss exponent is 3.67. Zero-forcing beamforming is used at the PHY for both CoMP and coordinated modes. There are $K = 1000$ data objects in the content server. The Data chunk size is 50 KB and the data object size is 1 MB. The cache size is $L_C = 80$ data object. At each user, object requests arrive according to a Poisson process with a total average arrival rate of λ Mbps. To verify the performance under both spatial and temporary popularity variations, we assume user j only requests a subset $\mathcal{F}_j$ of 100 data objects whose indices are randomly generated. The average arrival rate of data object $\mathcal{F}_j(k) \in \mathcal{F}_j$ at user j is $\lambda_j^{\mathcal{F}_j(k)} = \lambda\rho_k$, where $\mathcal{F}_j(k)$ is the k-th data object in $\mathcal{F}_j$ and ρ_k's follows the Zipf distribution $\rho_k = \frac{k^{-\varsigma}}{\sum_{k=1}^{|\mathcal{F}_j|} k^{-\varsigma}}, k = 1, ..., |\mathcal{F}_j|$. A larger *popularity skewness* ς means that the user requests concentrate more on a few popular files. The following baselines are considered.

Baseline 1 (Offline Caching with dual mode PHY [3], [4]**):** Each BS caches the most popular L_C data objects in an offline manner. Dual mode PHY is employed at the RAN.

Baseline 2 (LFU with dual mode PHY): In Least Frequently Used (LFU), the nodes record how often each data object has been requested and choose to cache the new data object if it is more frequently requested than the least frequently requested cached data object (which is replaced). Dual mode PHY is employed at the RAN.

Baseline 3 (VIP caching with single mode PHY [5]**):** The cache placement is determined by the VIP framework in [5] and only coordinated transmission mode is considered at the PHY.

For fair comparison, the data sub-channel R_d and control sub-channel R_c are assumed to share the same $R = 30$ Mbps backhaul capacity for all schemes. In Fig. 3, we plot the delay performance versus the average arrival rate of each user λ. The delay for an IP request is the difference between the fulfillment time (i.e., time of arrival of the requested DP) and the creation time of the IP request. It can be seen that the delay of all algorithms increases with the average arrival rate λ. The proposed has significant gain for practical scenario when the cache size is limited and the popularity is not too concentrated.

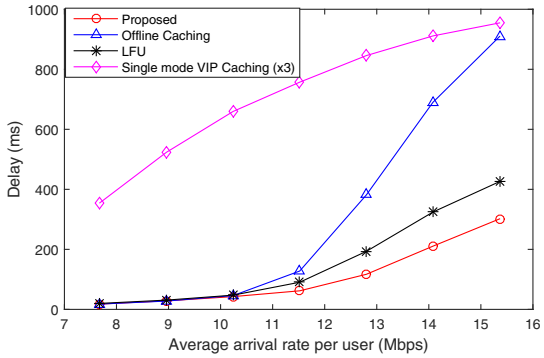


Figure 3: Delay versus per user average arrival rate with skewness $\varsigma = 0.5$.

VI. Conclusion

We propose a mixed timescale online PHY caching and content delivery scheme for wireless NDN networks with dual mode PHY. The cache content placement is performed once per frame (T time slots) to avoid excessive cache content placement cost. The PHY mode selection and rate allocation is performed once per time slot to fully exploit the cached content at the BS to enhance the system performance. To facilitate efficient resource control design, we introduce a *dual mode VIP framework* which transforms the original network into a virtual network and formulate the resource control design in the virtual network. The proposed solution can strike a balance between inducing MIMO cooperation gain and reducing the backhaul consumption. Simulations show that the proposed solution outperforms the existing offline PHY caching solution [3], [4] and online caching solutions [5].

References

[1] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*. IEEE, 2012, pp. 1107–1115.

[2] J. Dai, F. Liu, B. Li, B. Li, and J. Liu, "Collaborative caching in wireless video streaming through resource auctions," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 2, pp. 458–466, 2012.

[3] A. Liu and V. K. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, 2013.

[4] ——, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 390–402, 2014.

[5] E. Yeh, T. Ho, Y. Cui, M. Burd, R. Liu, and D. Leong, "VIP: A framework for joint dynamic forwarding and caching in named data networks," in *Proceedings of the 1st International Conference on Information-centric Networking*, 2014, pp. 117–126.

[6] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89–103, 2005.