Integrated Resource Management in Software Defined Networking, Caching and Computing

Qingxia Chen*†, F. Richard Yu[‡], Tao Huang*†, Renchao Xie*†, Jiang Liu*†, and Yunjie Liu*†
*State Key Laboratory of Networking and Switching Technology, Beijing University of Post and
Telecommunications, Beijing 100876, China

[†]Beijing Advanced Innovation Center for Future Internet Technology, Beijing 100876, China [‡]Depart. of Systems and Computer Eng., Carleton University, Ottawa, ON, Canada

Abstract—Recently, there are significant advances in the areas of networking, caching and computing. Nevertheless, these three important areas have traditionally been addressed separately in the existing research. In this paper, we present a novel framework that integrates networking, caching and computing in a systematic way and enables dynamic orchestration of these three resources to improve the end-to-end system performance and meet the requirements of different applications. Then, we consider the bandwidth, caching and computing resource allocation issue and formulate it as a joint caching/computing strategy and servers selection problem to minimize the combination cost of network usage and energy consumption in the framework. To minimize the combination cost of network usage and energy consumption in the framework, we formulate it as a joint caching/computing strategy and servers selection problem. In addition, we solve the joint caching/computing strategy and servers selection problem using an exhaustive-search algorithm. Simulation results show that our proposed framework significantly outperforms the traditional network without in-network caching/computing in terms of network usage and energy consumption.

Index Terms—Networking, caching, computing, resource allocation, energy efficient.

I. INTRODUCTION

Recently, there are significant advances in the areas of networking, caching and computing, which can have profound impacts on our society though the developments of smart cities, smart transportation, smart homes, etc. Software-defined networking (SDN) has been considered as one of the most promising technologies on realization of programmable networking, and has been deployed well in existing IP networks, such as Internet service providers and data center networks [1], [2]. By separating the control plane (decision functions) from the data plane (forwarding functions) in networks, SDN enables the development of new routing and forwarding approaches without the need to replace hardware components in the core network, simplifying network management and facilitating network evolution [3]–[7].

In the area of caching, information-centric networking (ICN), which has been extensively studied in recent years [8], [9], enables in-network caching to reduce the duplicate content transmission in networks [10]. By focusing on the data's names instead of their locations, ICN provides native support for

secure communication, scalable and efficient content distribution, and the enhanced capability for mobility [11]–[15]. In the area of computing, cloud computing paradigm has been widely adopted to utilize the computing resources in remote provider's servers via the Internet, providing enterprises and end users with a range of application services and freeing them from the specification of many details [16]-[22]. Nevertheless, it is not feasible or economical to satisfy current applications requirements of mobility support, location awareness, low latency and big data analytics as the distance between the cloud and the edge device is usually large. To address these issues, fog computing has been proposed to extends cloud computing and services close to end devices [23], [24]. A similar technique, called mobile edge computing, is being standardized to allocate computing resources in wireless access networks [25].

How to manage, control and optimize network, cache and compute (the three important underlying resources) can have great impacts on the performance of system and applications. Currently, in the works of SDN and ICN, these three resources are traditionally addressed separately, which could result in suboptimal performance. We present a novel framework called SD-NCC (Software Defined Networking, Caching and Computing) that integrates networking, caching and computing in a systematic way to meet the requirements of different applications and improve the end-to-end system performance. As Fig. 1 shows, the SD-NCC framework can be divided in three planes of functionality: data, control, and management planes (similar to SDN). Different from SDN, the added innetwork caching and computing in the data plane become inherent fundamental capabilities in SD-NCC. Thus, besides the software defined networking, SD-NCC enables the software defined caching and computing. Additionally, each data packet carries a name and a signature (similar to ICN), thus enabling information centric paradigm.

Based on the SD-NCC framework, we present a comprehensive system model and formulate a joint caching/computing strategy and servers selection problem as a mixed-integer nonlinear programming (MINLP) problem to minimize the combination cost of network usage and energy consumption. Specifically, we formulate the caching/computing capacity allocation problem and derive the optimal deployment numbers

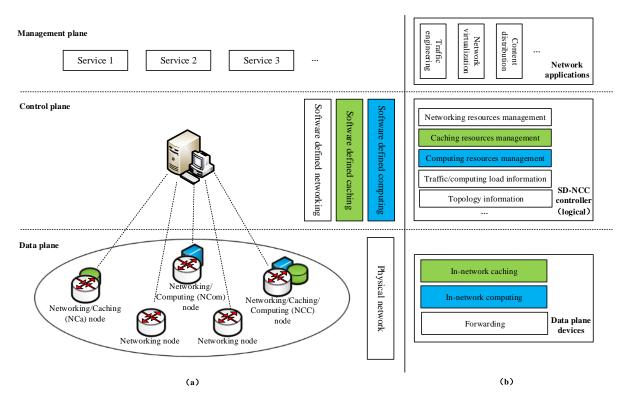


Fig. 1. Software-defined networking, caching, and computing in (a) planes and (b) system design architecture.

of service copies. In addition, we develop an exhaustive-search algorithm to find the optimal caching/computing strategy and servers selection strategy. Simulation results show that compared with traditional network, our proposed SD-NCC framework significantly reduces the traffic traversing the network and has performance advantage on energy consumption cost. We also show the impact of service popularity on optimal deployment number of service copies.

The rest of the article is organized as follows. Section II presents a system model for this framework. Section III formulates the joint bandwidth, caching and computing resource allocation problem and work out a near-optimal caching/computing strategy and servers selection solution. Simulation results are discussed in Section IV. Finally, we conclude this paper in Section V.

II. SYSTEM MODEL

In this section, we present the system model for the SD-NCC framework.

A. Network Model

Excellent works have been done on modeling networks [4], [26]–[52]. In this paper, Consider a network represented by graph $G = \{V, E\}$, where V denotes the set of nodes and E denotes the set of directed physical links. A node can be a router, a user, or a server. We introduce the following notation:

• $S = \{S_A, S_B\}$, the set of in-network server nodes; $S_A \subset V$, the set of caching nodes; $S_B \subset V$, the set of computing nodes.

- o, a single virtual origin node which is always able to serve the service requests.
- U, the set of users who issue service requests, $U \subset V$.
- $\mathbf{K} = {\mathbf{K}_A, \mathbf{K}_B}$, the set of service elements; \mathbf{K}_A , the set of content elements; \mathbf{K}_B , the set of computation services.
- $\mathbf{M}: \left\{m_u^k\right\}_{(k,u) \in \mathbf{K} \times \mathbf{U}}$, user demand matrix for duration $t; m_u^k$, the demand of service k from user $u, u \in \mathbf{U}$.
- $(\lambda_{k_1}^A, o_{k_1}^A)$, content k_1 's attribute, $k_1 \in \mathbf{K}_A$; $\lambda_{k_1}^A$, the number of content k_1 's requests for duration t; $o_{k_1}^A$, the size of content k_1 .
- $(\lambda_{k_2}^A, o_{k_2}^B, c_{k_2}^B)$, computation k_2 's attribute, $k_2 \in \mathbf{K}_B$; $\lambda_{k_2}^A$, the number of computation k_2 's requests for duration t; $o_{k_2}^B$, the amount of computation k_2 's data communications, $k_2 \in \mathbf{K}_B$; $c_{k_2}^B$, the amount of computation k_2 's workload (which can be measured by the number of clock cycles or execution time).

B. Caching/Computing Model

Let c_i^A , c_i^B denote the caching capacity and computing capacity of server node i. We assume the finite capacity of in-network server nodes and the infinite capacity of original server. We use $\mathbf{H}: \left\{h_i^k\right\}$ to denote the caching/computing strategy matrix, where $h_i^k=1$ if server i is able to provide the requested service k and $h_i^k=0$ if it is not. Specifically, $h_a^k=1$.

Thus, the caching/computing capacity constraints are as follows:

$$\sum_{k} h_i^k o_k^A \le c_i^A , \forall i \in \mathbf{S}_A$$
 (1)

and

$$\sum_{i} h_i^k c_k^B \le c_i^B \,, \forall i \in \mathbf{S}_B \tag{2}$$

C. Servers Selection Model

Let $\mathbf{P}: \left\{ \rho_{i,u}^k \right\}$ denote the servers selection matrix, where $\rho_{i,u}^k \in [0,\ 1]$ denotes the proportion of the service request k from user u served by server i. Node u can be viewed as an edge router that aggregates the traffic of many endhosts, which may be served by different servers. Let $\mathbf{X}: \left\{ x_{i,u}^k \right\}$ denote the traffic allocation matrix, where $x_{i,u}^k$ denotes the traffic of service k from user u to server i for duration t. Combining with the caching/computing strategy, we have

$$x_{i,u}^k = m_u^k \rho_{i,u}^k h_i^k, \forall i \in S \cup \{o\}$$

and

$$\sum_{i} \rho_{i,u}^{k} h_i^k = 1 \tag{4}$$

D. Routing Model

Let $\mathbf{R}:\left\{r_l^{i,u}\right\}$ be the routing matrix, where $r_l^{i,u}\in[0,1]$ denotes the proportion of traffic from user u to server i that traverses link l. $r_l^{i,u}$ equals 1 if the link l is on the path from server i to user u and 0 otherwise.

Then the routing selection strategy is given by

$$\sum_{l:l \in In(v)} r_l^{i,u} - \sum_{l:l \in Out(v)} r_l^{i,u} = I_{v=i}, \ \forall i \in \mathbf{S} \cup \{o\}, v \in \mathbf{V} \setminus \mathbf{U}$$

where $I_{v=i}$ is an indicator function which equals 1 if v=i and 0 otherwise, In(v) denotes the set of incoming links to node v, and Out(v) denotes the set of outgoing links from node v.

The capacity of a link l is c_l . Thus, the total traffic traversing link l denoted by x_l is given by

$$x_{l} = \sum_{k,i,u} x_{i,u}^{k} r_{l}^{i,u} \le c_{l}, \forall i \in \mathbf{S} \cup \{o\}$$
 (6)

E. Energy Model

The escalation of energy consumption in networks directly results in the increase of greenhouse gas emission, which has been recognized as a major threat for environmental protection and sustainable development [36], [45], [49], [51]–[53]. The energy is mainly consumed by content caching, data computing and traffic transmission in SD-NCC framework. We discuss these three energy models as follows.

1) Caching Energy: E_{ca}^{A} , we use an energy-proportional model, similar to [54]. If n_{k_1} copies of the content k_1 are cached for duration t, the energy consumed by caching $n_k o_k^A$ bits is given by

$$E_{ca,k_1}^A = n_{k_1} o_{k_1}^A p_{ca}^A t (7)$$

where p_{ca}^{A} is the power efficiency of caching. The value of p_{ca}^{A} strongly depends on the caching hardward technology.

2) Computing Energy: We assume that computation applications are executed within the virtual machines (VMs) which are deployed on the computing nodes and the amount of incoming computation k_2 's workloads from all users is $\lambda_{k_2}^B c_{k_2}^B$

for duration t. The energy consumption of the computing nodes consists of two parts, including a dynamic energy consumption part E^B_{active} when the VMs are active (processing computing service requests) and a static energy consumption part E^B_{static} when the VMs are idle.

Thus, the dynamic energy consumption on processing the workloads is

$$E_{active}^{B,k_2} = \lambda_{k_2}^B c_{k_2}^B p_{active}^B \tag{8}$$

and the static energy consumption for m_{k_2} copies of computation k_2 's dedicated VM is

$$E_{static}^{B,k_2} = m_{k_2} p_{static}^B t (9)$$

where p_{active}^{B} , p_{static}^{B} are the average power efficiency of VMs in active and static states, respectively.

Note that the dynamic energy consumption is independent of the VMs' copies.

3) Transmission Energy: The transmission energy E_{tr} consumption mainly consists of the energy consumption at routers and energy consumption along the links.

The transmission energy of content E_{tr}^{A} and computation E_{tr}^{B} for duration t are given by

$$E_{tr}^{A,k_1} = \lambda_{k_1}^A o_{k_1}^A \left[p_{tr,link} \cdot d_{k_1}^A + p_{tr,node} \cdot \left(d_{k_1}^A + 1 \right) \right] \quad (10)$$

$$E_{tr}^{B,k_2} = \lambda_{k_2}^B o_{k_2}^B \left[p_{tr,link} \cdot d_{k_2}^B + p_{tr,node} \cdot \left(d_{k_2}^B + 1 \right) \right]$$
 (11)

where $p_{tr,link}$ and $p_{tr,node}$ are the energy efficiency parameters of the link and node respectively. $d_{k_1}^A$ and $d_{k_2}^B$ represent the average hop distance to the content server nodes for content k_1 's request and the computation server nodes for computation k_2 's request, respectively.

III. CACHING/COMPUTING/BANDWIDTH RESOURCE ALLOCATION

In this subsection, we consider the joint caching, computing and bandwidth resource allocation problem and formulate it as a joint caching/computing strategy and servers selection problem (CCS-SS) which we show is a mixed integer nonlinear programming (MINLP) problem. By relaxing the whole caching, computing, link constraints, we focus on the caching/computing capacity allocation for each service element and formulate it as a a nonlinear programming (NLP) problem. We derive the optimal caching/computing capacity allocation (i.e., the copies number for each service element) and then propose an exhaustive-search algorithm to find the optimal caching/computing strategy and servers selection.

A. Problem Formulation

- 1) Objective Function: Based on the known topology, traffic matrix, routing matrix and the energy parameters of network equipments, we introduce the following quantities:
 - $d_{i,u}$, the hop distance between server node i and user u;
 - $D_{i,u}$, the end-to-end latency between server node i and user u;
 - $a_{i,u} = p_{tr,link}d_{i,u} + p_{tr,node}(d_{i,u} + 1)$, the energy for transporting per bit from server node i to user u;

- $f_{e,tr} = \sum_{k \in \mathbf{K}} \sum_{i \in \mathbf{S} \cup \{o\}} \sum_{u \in \mathbf{U}} a_{i,u} m_u^k \rho_{i,u}^k h_i^k$, the transmission energy for duration t;
 $f_{e,ca} = \sum_{k_1 \in \mathbf{K}_A} \sum_{i \in \mathbf{S}_A} h_i^{k_1} o_{k_1}^A p_{ca}^A t$, the caching energy for duration t;
- $f_{e,com} = \sum_{k_2 \in \mathbf{K}_B} \sum_{i \in \mathbf{S}_B} h_i^{k_2} p_{static}^B t + \sum_{k_2 \in \mathbf{K}_B} \lambda_{k_2}^B c_{k_2}^B p_{active}^B$, the energy consumption on computing nodes for duration
- $f_{tr} = \sum_{k \in \mathbf{K}} \sum_{i \in \mathbf{S} \cup \{o\}} \sum_{u \in \mathbf{U}} D_{i,u} m_u^k \rho_{i,u}^k h_i^k$, the network traffic for duration t.

We select a combining objective which balances the energy costs and the network usage costs for duration t. The cost function is given by

$$f = f_{e,ca} \left(h_i^k \right) + f_{e,com} \left(h_i^k \right) + f_{e,tr} \left(\rho_{i,u}^k h_i^k \right)$$
$$+ \gamma f_{tr} \left(\rho_{i,u}^k h_i^k \right)$$
(12)

where γ is the weight between the two costs. The first three parts constitute the energy consumption for duration t and the last one is network traffic which denotes the network usage in this paper.

2) Formulation: In the following, we formulate the CCS-SS problem to minimize the combination cost function. The optimization problem is shown as follows:

Constraints (1) specifies that user u's demand rate for service k can be served by several servers simultaneously depending on caching/computing strategy and servers selection strategy. Constraints (2) and (3) specify the caching/computing resource capacity limits on in-network server nodes. Data rates are subject to link capacity constraint (4).

Problem (13) is difficult to solve based on the following observations:

- Both the objective function and feasible set of (13) are not convex due to the binary variables h_i^k and the product relationship between h_i^k and $\rho_{i,u}^k$.
- The size of the problem is very large. For instance, if there are F service elements and N network nodes, the number of variables h_i^k is N^F . In future networks, the number of the service elements and network nodes will rise significantly.

As is well known, a MINLP problem is expected to be NPhard [55], and a variety of algorithms are capable of solving instances with hundreds or even thousands of variables. However, it is very challenging to find the global optimum resource allocation in our problem, since the number of variables and the constraints grow exponentially. How to efficiently solve it is left for future research. Instead, in the next subsection, we propose an NLP formulation from a different view which can be solved analytically.

B. Caching/Computing Capacity Allocation

In this section, we focus on the optimal caching/computing capacity allocated for each service k, but not the optimal cache location. We denote n_{k_1} as the number of content k_1 's copies and m_{k_2} as the number of computation k_2 's dedicated VM copies. For simplicity, we remove the integer constraint of n_{k_1} and m_{k_2} . In a network with N nodes, if service k can be provided by n server nodes, N-n nodes have to access the service via one or more hops in a steady state. For the irregular and asymmetric network, the authors in [54] take a semi-analytical approach in deriving the average hop distance to the servers,d, as a power-law function of n:

$$d(n) = A(N/n)^{\alpha} \tag{14}$$

Thus, we have

$$d_{k_1}^A = A(N/n_{k_1})^{\alpha}, d_{k_2}^B = A(N/m_{k_2})^{\alpha}$$
 (15)

where $d_{k_1}^A$ and $d_{k_2}^B$ are the average hop distance to content k_1 's copies and computation k_2 's dedicated VM copies, respectively.

We assume the average end-to-end latency is proportional to the average hop distance and the scaling factor is η . Then the network traffic for duration t is

$$T_{total} = \sum_{k_1 \in K_A} T_{k_1}^A + \sum_{k_2 \in K_B} T_{k_2}^B$$

$$= \eta \left(\sum_{k_1 \in K_A} \lambda_{k_1}^A o_{k_1}^A d_{k_1}^A + \sum_{k_2 \in K_B} \lambda_{k_2}^B o_{k_2}^B d_{k_2}^B \right)$$
(16)

According to energy model in Subsection II-E, the total energy consumption for duration t is as follows:

$$E_{total} = E_{total}^{A} + E_{total}^{B}$$

$$= \sum_{k_{1} \in \mathbf{K}_{A}} \left(E_{ca,k_{1}}^{A} \left(n_{k_{1}} \right) + E_{tr}^{A,k_{1}} \left(n_{k_{1}} \right) \right)$$

$$+ \sum_{k_{2} \in \mathbf{K}_{B}} \left(E_{static}^{B,k_{1}} \left(m_{k_{2}} \right) + E_{active}^{B,k_{2}} + E_{tr}^{B,k_{2}} \left(m_{k_{2}} \right) \right)$$
(17)

Thus, the caching/computing capacity allocation problem can be formulated as:

$$\min E_{total} + \gamma T_{total}
s.t. \quad 1 \le n_{k_1} \le N
\quad 1 \le m_{k_2} \le N$$
(18)

We ignore the server capacity constraint by assuming sufficiently large server capacities, since we are more interested in the impact of caching/computing capacity allocation on total network cost than the decision of load balancing among congested servers.

By using the Lagrangian dual method, the optimal n_{k_1} and m_{k_2} , which are denoted as $n_{k_1}^*$ and $m_{k_2}^*$, can be derived as:

$$n_{k_1}^* = \max[1, \min[n_{k_1}^o, N]] m_{k_2}^* = \max[1, \min[m_{k_2}^o, N]]$$
 (19)

where $n_{k_1}^o$ and $m_{k_2}^o$ are given by

$$n_{k_1}^o = \left\lceil \frac{A\lambda_{k_1}^A \alpha \left(p_{tr,link} + p_{tr,node} + \gamma \eta \right)}{p_{ca}^A t} \right\rceil^{\frac{1}{\alpha+1}} N^{\frac{\alpha}{\alpha+1}}$$
 (20)

$$m_{k_2}^o = \left[\frac{A\lambda_{k_2}^A o_{k_2}^B \alpha \left(p_{tr,link} + p_{tr,node} + \gamma \eta \right)}{p_{static}^B t} \right]^{\frac{1}{\alpha+1}} N^{\frac{\alpha}{\alpha+1}}$$
(21)

C. The Exhaustive-search Algorithm

The CCS-SS problem (13) can be denoted by minimizing $f(\mathbf{H}, \mathbf{P}|\mathbf{M}, \mathbf{R})$ with the caching, computing, link constraints. Note that if the caching/computing strategy matrix \mathbf{B} is pre-known, the formulation (13) will turn into minimizing $f(\mathbf{P}|\mathbf{M}, \mathbf{R}, \mathbf{H})$ with the link constraints which is a linear optimization problem. In Subsection III-B, we derived the optimal deployment copies $n_{k_1}^*$ for each content k_1 , and $m_{k_2}^*$ for computation service k_2 without regard to copy locations in the network. We assume that the element numbers of \mathbf{K}_A , \mathbf{K}_B are F_1 , F_2 . Thus, the number of all possible combination subsets of caching/computing strategy (copy locations) Q are

$$Q = \prod_{k_1 \in \mathbf{K}_A} C_N^{n_{k_1}} \prod_{k_2 \in \mathbf{K}_B} C_N^{m_{k_2}} \tag{22}$$

which is significantly reduced in contrast with the original number $2^{N(F_1+F_2)}$. In the following, we propose an exhaustive-search algorithm to find optimal resource allocation solutions for each service.

We denote the set of all possible combination subsets as

$$\mathbf{\Phi} = \{\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_Q\} \tag{23}$$

Then, the resource allocation process is described as follows. After the controller selects a caching/computing strategy $\mathbf{H}_q \in \mathbf{\Phi}$, it minimizes $f(\mathbf{P}|\mathbf{M},\mathbf{R},\mathbf{H})$ with the link constraints: finding the optimal servers selection \mathbf{P} for the service requests to minimize the combination cost function f. By exhaustive searching, we choose the optimal caching/computing strategy \mathbf{H}^* :

$$\mathbf{H}^* = \arg\min_{\mathbf{H}_q \in \mathbf{\Phi}} (\mathbf{P} \mid \mathbf{M}, \mathbf{R}, \mathbf{H}_q)$$
 (24)

and the corresponding P^* and f^* .

IV. SIMULATION RESULTS AND DISCUSSIONS

In this section, we conduct simulations based on a hypothetical United States backbone network US64 [54], which is a representation of the topological characteristics of commercial networks. In the simulation, the parameters of the network are referred to [56]. Energy density of a link is $p_{tr,link} = 0.15 \times 10^{-8} J/bit$. Energy density of a router is in the order of $p_{tr,node} = 2 \times 10^{-8} J/bit$. Power density of caching is $p_{ca}^A = 0.25 \times 10^{-8} W/bit$. Power density of computation VM in the static state is $p_{static}^B = 50W$. We assume that both content and computation traffic demands are 1Gbps.

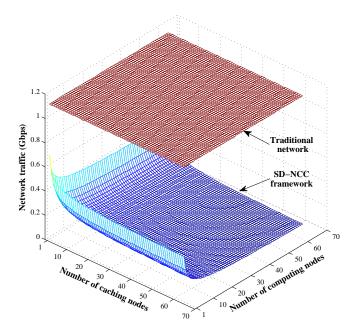


Fig. 2. The network traffic comparison between two schemes.

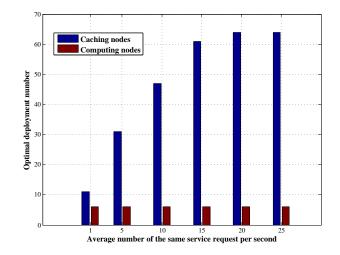


Fig. 3. The impact of service popularity on optimal deployment number of service copies.

A. Network Usage Cost

In Fig. 2, we show the impact of caching/computing nodes' numbers on the average network traffic per second. We can see that SD-NCC architecture significantly reduces the network traffic compared with the traditional network without innetwork caching/computing. This is due to the fact that with in-network caching and computing, large amount of content and computing requests are served on in-network nodes.

B. Optimal Deployment Numbers

In Subsection III-B, we derive the optimal deployment numbers on minimizing the combination costs of energy consumption and network usage. Fig. 3 shows that under the fixed traffic demand, the optimal deployment number of caching copies for each content is proportional to the service popularity but the optimal deployment number of computation copies for each computation is independent of it. Because the larger number of the same content request (from different users), the less energy on caching (less content need to be cached). In contrast, the energy consumption on computing is fixed. Since computation tasks are more complex and the computation result is difficult to be used for other computation requests, even for the same computation requests from different users.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to jointly consider networking, caching and computing in a systematic framework to improve the end-to-end system performance. We presented its system model and formulated the joint caching, computing and bandwidth resource allocation problem to minimize the energy cost and network usage cost. In addition, we derived the optimal deployment number of service copies and used an exhaustive-search algorithm to find the near-optimal caching/computing strategy and servers selection. Simulation showed that comparing the traditional network, our proposed SD-NCC framework significantly reduces the traffic traversing the network and has performance advantage on energy consumption. Future work is in progress to consider wireless networks in the proposed framework.

ACKNOWLEDGMENT

This work is jointly supported by the National High Technology Research and Development Program(863) of China (No. 2015AA016101), the National Natural Science Fund (No. 61302089, 61300184).

REFERENCES

- D. Kreutz, F. Ramos, P. Esteves Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, Jan. 2015.
- [2] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A survey on software-defined networking," *IEEE Commun. Surveys Tutorials*, vol. 17, no. 1, pp. 27–51, Firstquarter 2015.
- [3] F. Hu, Q. Hao, and K. Bao, "A survey on software-defined network and openflow: From concept to implementation," *IEEE Commun. Surveys Tutorials*, vol. 16, no. 4, pp. 2181–2206, Fourthquarter 2014.
- [4] Q. Yan, F. R. Yu, Q. Gong, and J. Li, "Software-defined networking (SDN) and distributed denial of service (DDoS) attacks in cloud computing environments: A survey, some research issues, and challenges," *IEEE Commun. Survey and Tutorials*, vol. 18, no. 1, pp. 602–622, 2016.
- [5] Y. Cai, F. R. Yu, C. Liang, B. Sun, and Q. Yan, "Software defined device-to-device (D2D) communications in virtual wireless networks with imperfect network state information (NSI)," *IEEE Trans. Veh. Tech.*, 2015, dOI:10.1109/TVT.2015.2483558.
- [6] L. Cui, F. R. Yu, and Q. Yan, "When big data meets software-defined networking (SDN): SDN for big data and big data for SDN," *IEEE Network*, vol. 30, no. 1, pp. 58–65, Jan. 2016.
- [7] Q. Yan and F. R. Yu, "Distributed denial of service attacks in software-defined networking with cloud computing," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 52–59, Apr. 2015.
- [8] G. Xylomenos, C. N. Ververidis, V. A. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. V. Katsaros, and G. C. Polyzos, "A survey of information-centric networking research," *IEEE Commun. Surveys Tutorials*, vol. 16, no. 2, pp. 1024–1049, Second Quarter 2014.
- [9] S. Lederer, C. Mueller, C. Timmerer, and H. Hellwagner, "Adaptive multimedia streaming in information-centric networks," *IEEE Network*, vol. 28, no. 6, pp. 91–96, Nov. 2014.

- [10] M. Zhang, H. Luo, and H. Zhang, "A survey of caching mechanisms in information-centric networking," *IEEE Commun. Surveys Tutorials*, vol. 17, no. 3, pp. 1473–1499, Thirdquarter 2015.
- [11] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, "A survey of green information-centric networking: Research issues and challenges," *IEEE Commun. Surveys Tutorials*, vol. 17, no. 3, pp. 1455–1472, Thirdquarter 2015.
- [12] C. Liang and F. R. Yu, "Virtual resource allocation in information-centric wireless virtual networks," in *Proc. IEEE ICC'15*, London, UK, June 2015.
- [13] C. Liang, F. R. Yu, H. Yao, and Z. Han, "Virtual resource allocation in information-centric wireless virtual networks," *IEEE Trans. Veh. Tech.*, 2016, to appear, available online.
- [14] 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.
- [15] K. Wang, F. R. Yu, and H. Li, "Information-centric virtualized cellular networks with device-to-device (D2D) communications," *IEEE Trans. Veh. Tech.*, 2016, accepted, online.
- [16] M. Armbrust, A. Fox, R. Griffith, and A. D. Joseph, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.
- [17] Z. Yin, F. R. Yu, S. Bu, and Z. Han, "Joint cloud and wireless networks operations in mobile cloud computing environments with telecom operator cloud," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 4020–4033, July 2015.
- [18] Y. Cai, F. R. Yu, and S. Bu, "Dynamic operations of cloud radio access networks (C-RAN) for mobile cloud computing systems," *IEEE Trans. Veh. Tech.*, 2015, doi: 10.1109/TVT.2015.2411739, online.
- [19] ——, "Cloud computing meets mobile wireless communications in next generation cellular networks," *IEEE Network*, vol. 28, no. 6, pp. 54–59, Nov. 2014.
- [20] —, "Joint dynamic cloud and wireless networks operations in the mobile cloud computing environment," in *Proc. IEEE Infocom'14 Workshop* on Mobile Cloud Computing, Toronto, ON, Apr. 2014.
- [21] Z. Yin, F. R. Yu, and S. Bu, "Joint dynamic cloud and wireless networks operations in the mobile cloud computing environment," in *Proc. IEEE Globecom'14*, Austin, TX, Dec. 2014.
- [22] F. R. Yu and V. C. M. Leung, Advances in Mobile Cloud Computing Systems. New York: CRC Press, 2015.
- [23] M. Zhanikeev, "A cloud visitation platform to facilitate cloud federation and fog computing," *Computer*, vol. 48, no. 5, pp. 80–83, May 2015.
- [24] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the suitability of fog computing in the context of Internet of things," *IEEE Trans. Cloud Computing*, vol. PP. no. 99, 2015, online.
- [25] ETSI, "Mobile-edge computing introductory technical white paper," ETSI White Paper, Sept. 2014.
- [26] L. Ma, F. Yu, V. C. M. Leung, and T. Randhawa, "A new method to support UMTS/WLAN vertical handover using SCTP," *IEEE Wireless Commun.*, vol. 11, no. 4, pp. 44–51, Aug. 2004.
- [27] F. Yu and V. C. M. Leung, "Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks," in *Proc. IEEE INFOCOM'01*, Anchorage, AK, Apr. 2001.
- [28] Z. Li, F. R. Yu, and M. Huang, "A distributed consensus-based cooperative spectrum sensing in cognitive radios," *IEEE Trans. Veh. Tech.*, vol. 59, no. 1, pp. 383–393, Jan. 2010.
- [29] F. Yu and V. Krishnamurthy, "Optimal joint session admission control in integrated WLAN and CDMA cellular networks with vertical handoff," *IEEE Trans. Mobile Computing*, vol. 6, no. 1, pp. 126–139, Jan. 2007.
- [30] R. Xie, F. R. Yu, H. Ji, and Y. Li, "Energy-efficient resource allocation for heterogeneous cognitive radio networks with femtocells," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 3910 –3920, Nov. 2012.
- [31] A. Attar, H. Tang, A. Vasilakos, F. R. Yu, and V. Leung, "A survey of security challenges in cognitive radio networks: Solutions and future research directions," *Proceedings of the IEEE*, vol. 100, no. 12, pp. 3172–3186, 2012.
- [32] Y. Fei, V. W. S. Wong, and V. C. M. Leung, "Efficient QoS provisioning for adaptive multimedia in mobile communication networks by reinforcement learning," *Mob. Netw. Appl.*, vol. 11, no. 1, pp. 101–110, Feb. 2006.

- [33] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tutorials*, vol. 17, no. 1, pp. 358–380, Firstquarter 2015.
- [34] Y. Wei, F. R. Yu, and M. Song, "Distributed optimal relay selection in wireless cooperative networks with finite-state Markov channels," *IEEE Trans. Veh. Tech.*, vol. 59, no. 5, pp. 2149 –2158, June 2010.
- [35] Q. Guan, F. R. Yu, S. Jiang, and G. Wei, "Prediction-based topology control and routing in cognitive radio mobile ad hoc networks," *IEEE Trans. Veh. Tech.*, vol. 59, no. 9, pp. 4443 –4452, Nov. 2010.
- [36] S. Bu, F. R. Yu, Y. Cai, and P. Liu, "When the smart grid meets energy-efficient communications: Green wireless cellular networks powered by the smart grid," *IEEE Trans. Wireless Commun.*, vol. 11, pp. 3014–3024, Aug. 2012.
- [37] R. Xie, F. R. Yu, and H. Ji, "Dynamic resource allocation for heterogeneous services in cognitive radio networks with imperfect channel sensing," *IEEE Trans. Veh. Tech.*, vol. 61, pp. 770–780, Feb. 2012.
- [38] F. R. Yu, H. Tang, M. Huang, Z. Li, and P. C. Mason, "Defense against spectrum sensing data falsification attacks in mobile ad hoc networks with cognitive radios," in *Proc. IEEE Military Commun. Conf.* (MILCOM)'09, Oct. 2009.
- [39] F. R. Yu, M. Huang, and H. Tang, "Biologically inspired consensusbased spectrum sensing in mobile ad hoc networks with cognitive radios," *IEEE Network*, vol. 24, no. 3, pp. 26 –30, May 2010.
- [40] C. Luo, F. R. Yu, H. Ji, and V. C. M. Leung, "Cross-layer design for TCP performance improvement in cognitive radio networks," *IEEE Trans. Veh. Tech.*, vol. 59, no. 5, pp. 2485–2495, 2010.
- [41] F. Yu, V. Krishnamurthy, and V. C. M. Leung, "Cross-layer optimal connection admission control for variable bit rate multimedia traffic in packet wireless CDMA networks," *IEEE Trans. Signal Proc.*, vol. 54, no. 2, pp. 542–555, Feb. 2006.
- [42] F. R. Yu, P. Zhang, W. Xiao, and P. Choudhury, "Communication systems for grid integration of renewable energy resources," *IEEE Network*, vol. 25, no. 5, pp. 22 –29, Sept. 2011.
- [43] G. Liu, F. R. Yu, H. Ji, V. C. M. Leung, and X. Li, "In-band full-duplex relaying for 5G cellular networks with wireless virtualization," *IEEE Network*, vol. 29, no. 6, pp. 54–61, Nov. 2015.
- [44] Q. Guan, F. R. Yu, S. Jiang, and V. Leung, "Capacity-optimized topology control for MANETs with cooperative communications," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 2162 –2170, July 2011.
- [45] S. Bu and F. R. Yu, "Green cognitive mobile networks with small cells

- for multimedia communications in the smart grid environment," *IEEE Trans. Veh. Tech.*, vol. 63, no. 5, pp. 2115–2126, June 2014.
- [46] J. Liu, F. R. Yu, C.-H. Lung, and H. Tang, "Optimal combined intrusion detection and biometric-based continuous authentication in high security mobile ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 806–815, 2009.
- [47] S. Bu, F. R. Yu, and H. Yanikomeroglu, "Interference-aware energy-efficient resource allocation for heterogeneous networks with incomplete channel state information," *IEEE Trans. Veh. Tech.*, vol. 64, no. 3, pp. 1036–1050, Mar. 2015.
- [48] L. Zhu, F. R. Yu, B. Ning, and T. Tang, "Cross-layer handoff design in MIMO-enabled WLANs for communication-based train control (CBTC) systems," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 719–728, May 2012.
- [49] S. Zhang, F. R. Yu, and V. Leung, "Joint connection admission control and routing in IEEE 802.16-based mesh networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1370 –1379, Apr. 2010.
- [50] Z. Wei, H. Tang, F. R. Yu, M. Wang, and P. Mason, "Security enhancements for mobile ad hoc networks with trust management using uncertain reasoning," *IEEE Trans. Veh. Tech.*, vol. 63, no. 9, pp. 4647–4658, Nov. 2014.
- [51] S. Bu, F. R. Yu, and P. Liu, "Dynamic pricing for demand-side management in the smart grid," in *Proc. IEEE Online Conference on Green Communications (GreenCom)*'11, Sept. 2011, pp. 47–51.
- [52] S. Bu and F. R. Yu, "A game-theoretical scheme in the smart grid with demand-side management: Towards a smart cyber-physical power infrastructure," *IEEE Trans. Emerging Topics in Computing*, vol. 1, no. 1, pp. 22–32, June 2013.
- [53] G. Cili, H. Yanikomeroglu, and F. R. Yu, "Cell switch off technique combined with coordinated multi-point (CoMP) transmission for energy efficiency in beyond-LTE cellular networks," in *Proc. IEEE ICC'12*, June 2012.
- [54] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, "In-network caching effect on optimal energy consumption in content-centric networking," in *Proc. IEEE ICC'12*, June 2012, pp. 2889–2894.
- [55] S. Burer and A. N. Letchford, "Non-convex mixed-integer nonlinear programming: a survey," Surveys in Operations Research and Management Science, vol. 17, no. 2, pp. 97–106, 2012.
- [56] J. Baliga, R. Ayre, K. Hinton, and R. S. Tucker, "Architectures for energy-efficient iptv networks," in *Optical Fiber Communication Con*ference (OFC), Mar. 2009, pp. 1–3.