Virtual Resource Allocation for Information-Centric Heterogeneous Networks with Mobile Edge Computing

Presenter: Prof. F. Richard Yu

Yuchen Zhou*, Fei Richard Yu[×], Jian Chen*, Yonghong Kuo*

*School of Telecomm. Eng., Xidian University, Xi'an 710071, P.R.China
[×]Dept. of Systems and Computer Eng., Carleton University, Ottawa, ON, Canada

2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)

IEEE INFOCOM

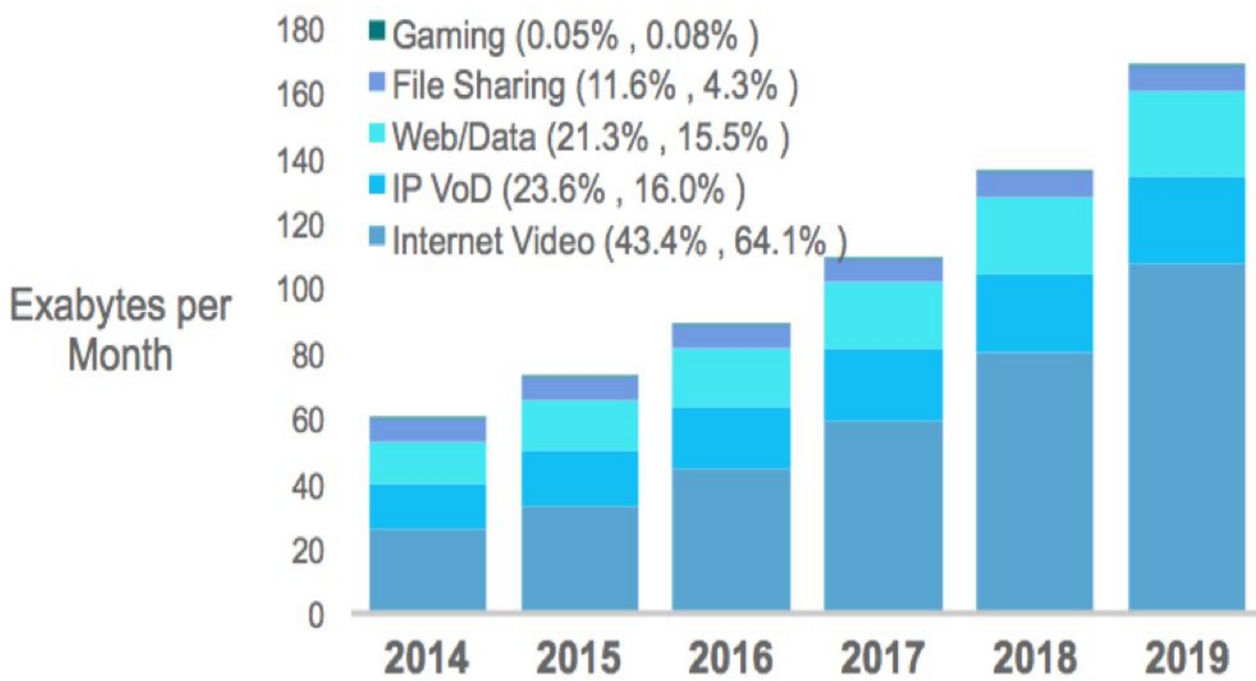
Outline

- Introduction
- System Model
- Problem Formulation
- Resource Allocation Process
- Simulation Results
- Conclusion

Global IP Video Traffic Growth

IP Video Will Account for 80% of Global IP Traffic by 2019

23% CAGR 2014–2019



- Prodigious amount of video traffic results in large-scale distribution of video contents calling for tremendous resources.
- Video contents need to be transcoded to fit the network condition and the usage of different mobile devices with different formats, resolutions, and bitrates.

* Figures (n) refer to 2014, 2019 traffic shares

Source: Cisco VNI Global IP Traffic Forecast, 2014–2019

Information-Centric Networking

- ICN is characterized by receiver-driven information-level delivery and *in-network caching*.

  - ICN is capable of converting the sender-driven end-to-end network paradigm into the receiver-driven content retrieval paradigm [1].
  - *In-network caching* can help efficient distribution of contents in wireless networks [1].
  - Compared to the traditional network paradigms with a general lack of content distribution information, the cache-enabled network is able to reduce the backhaul cost of the popular contents, increase the content delivery probability to mobile users, and support a highly efficient scalable content retrieval.

[1] 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, Secondquarter 2014.

Mobile Edge Computing

- MEC is recognized as a promising paradigm in next generation wireless networks, enabling the cloud-computing capabilities in close proximity to mobile devices.

  - With the physical proximity, MEC realizes a low-latency connection to a large-scale resource-rich computing infrastructure by offloading the computation task to an adjacent computing sever/cluster instead of relying on a remote cloud [2].

[2] N. Kumar, S. Zeadally, and J. J. Rodrigues, "Vehicular delay-tolerant networks for smart grid data management using mobile edge computing," *IEEE Commun. Mag.*, vol. 54, no. 10, pp. 60–66, Oct. 2016.

Wireless Network Virtualization

- In this paper, wireless network virtualization is considered as a candidate technique for simplifying network management.

  - Through virtualization, wireless network infrastructure can be decoupled from the provided services, and various users with differentiated services requirements can dynamically share the same infrastructure, thereby maximizing the system utilization [3].

[3] 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.

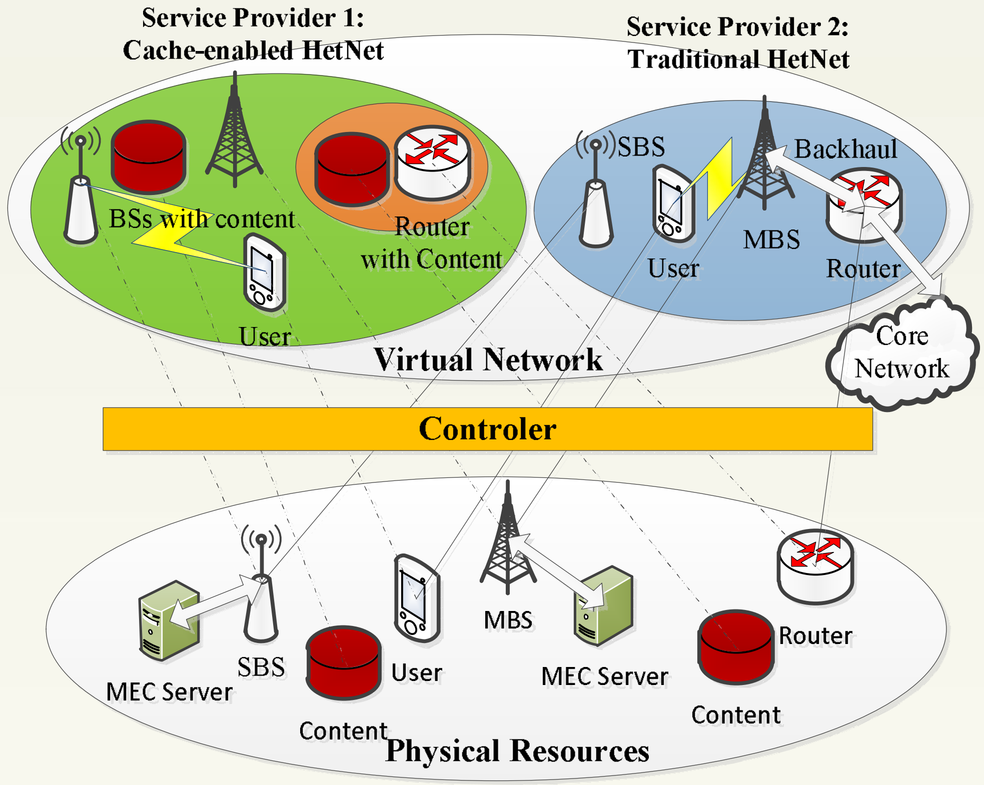
The distinctive features of this paper

- We design a novel virtualized HetNets framework aiming at enabling content caching and computing, in which the resources of communication, computing, and caching can be shared among users from different virtual networks.
- In this framework, we formulate the virtual resource allocation strategy as a joint optimization problem, where the gains of not only virtualization but also caching and computing are taken into consideration.
- A distributed algorithm, based on alternating direction method of multipliers (ADMM), is presented to solve the formulated problem with a lower computational complexity and a reduced signaling overhead.

Outline

Virtual Heterogeneous Networks Model

- Virtual radio resources are generated according to the requests of services providers (SPs) by mobile virtual network operator (MVNO).
- MVNO leases radio resource (e.g., spectrum) and backhaul bandwidth (e.g., data rate) from infrastructure providers (InPs), and slices them to virtual SPs.



Computing Model

- Assume each offloading computation task of user u_s can be described into four terms as $T_{u_s} = \{z_{u_s}, z'_{u_s}, R^{cm}_{u_s}, R^{cp}_{u_s}\}$.

  - The first two terms respectively represent the sizes of the contents before and after the computation, which can be used to calculate the computation energy and the computation rate.
  - The last two terms respectively represent the minimum requirements of the offloading rate (e.g., communication rate) and the computation rate.
  - The computation energy: $E_{u_s,n} = c_{u_s} e_n$ The computation rate: $R_{u_s,n} = \frac{z_{u_s}}{t_{u_s,n}} = \frac{f_{u_s,n} z_{u_s}}{c_{u_s}}$

c_{u_s} the computing ability required for accomplishing this task, which can be quantized by the amount of CPU cycles

e_n the energy consumption for one CPU cycle at BS *n*

$f_{u_s,n}$ the computation capability of BS *n* assigned to the user, which is quantized by the total number of CPU cycles per second

Caching Model

- The popularity distribution of each content can be described as a vector $P = \{p_1, p_2, ..., p_F\}$, if there exist *F* types of contents.
  - *P* can be modeled as the Zipf distribution, $p_f = \frac{1/f^{\varepsilon}}{\sum_{f=1}^{F} 1/f^{\varepsilon}}$.
- For each BS, they can determine whether to cache the content sent by users before or after the computation, according to the popularity distribution of each content.
  - The expected saved backhaul bandwidth through caching the contents can be calculated as

    $g_{z_{u_s} or\ z'_{u_s}} = p_{z_{u_s} or\ z'_{u_s}} \overline{R}_n$, where $\overline{R}_n$ is assumed to be the backhaul bandwidth usage, which can be viewed as the average data rate of each BS.

Outline

Decision Variables

- Association indicators $a_{u_s,n}$: give instructions that the base station of which SP will schedule the next serving users.
- Bandwidth assignment indicators $b_{u_s,n}$: indicate the appropriate allocated bandwidth to each subscribed user according to the QoS requirements.
- Caching indicators $x_{u_s,n}$: determine whether or not to cache the contents before or after the computation at base stations.

Objective Function

$$U_{u_s,n} = a_{u_s,n}(\alpha_{u_s} b_{u_s,n} r_{u_s,n} - \beta_n b_{u_s,n}) + a_{u_s,n}(\phi_{u_s} R_{u_s,n} - \varphi_n E_{u_s,n})$$
$$+ a_{u_s,n} x^1_{u_s,n}(\gamma_n g_{z_{u_s}} - \Psi^n_{z_{u_s}} z_{u_s}) + a_{u_s,n} x^2_{u_s,n}(\gamma_n g_{z'_{u_s}} - \Psi^n_{z'_{u_s}} z'_{u_s})$$

- The designed uplink virtual resources management strategy aims at maximizing the total utility of MVNO.

  - Communication Revenue: *the gain of user data rate – the cost of consumed radio bandwidth*
  - Computation Revenue: *the gain of computation rate – the cost of consumed computation energy*
  - Caching Revenue: *the gain of saved backhaul bandwidth – the cost of consumed caching space*

Problem Formulation

$$OP1: \max_{\{a_{u_s,n}, b_{u_s,n}, x^1_{u_s,n}, x^2_{u_s,n}\}} \sum_{s\in\mathcal{S}} \sum_{u_s\in\mathcal{U}_s} \sum_{n\in\mathcal{N}} U_{u_s,n}$$

$$s.t.: C1: \sum_{n\in\mathcal{N}} a_{u_s,n} = 1, \forall s \in \mathcal{S}, u_s \in \mathcal{U}_s$$

$$C2: \sum_{s\in\mathcal{S}} \sum_{u_s\in\mathcal{U}_s} a_{u_s,n} b_{u_s,n} \leq B_n, \forall n \in \mathcal{N}$$

$$C3: \sum_{n\in\mathcal{N}} a_{u_s,n} b_{u_s,n} r_{u_s,n} \geq R^{\text{cm}}_{u_s}, \forall s \in \mathcal{S}, u_s \in \mathcal{U}_s$$

$$C4: \sum_{n\in\mathcal{N}} a_{u_s,n} R_{u_s,n} \geq R^{\text{cp}}_{u_s}, \forall s \in \mathcal{S}, u_s \in \mathcal{U}_s$$

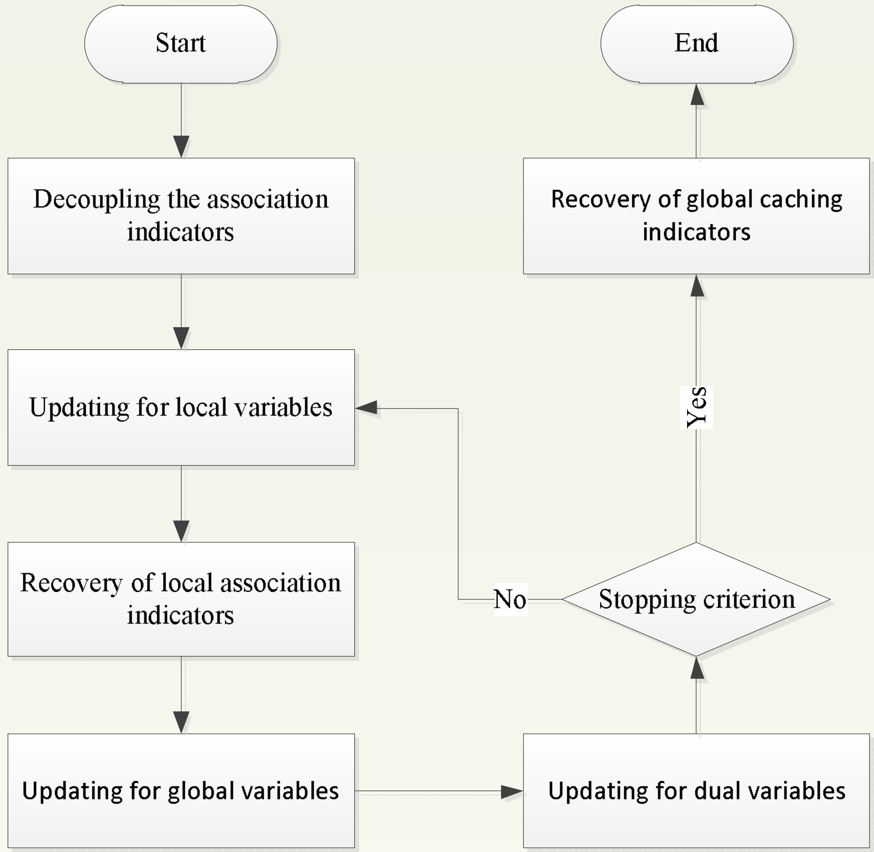
$$C5: \sum_{s\in\mathcal{S}} \sum_{u_s\in\mathcal{U}_s} a_{u_s,n} \leq D_n, \forall n \in \mathcal{N}$$

$$C6: \sum_{s\in\mathcal{S}} \sum_{u_s\in\mathcal{U}_s} a_{u_s,n}(x^1_{u_s,n} z_{u_s} + x^2_{u_s,n} z'_{u_s}) \leq Z_n, \forall n \in \mathcal{N}$$

- The strategy should guarantee the communication rate requirement and the computation rate requirement of each user, and that the total amount of the allocated virtual resources to all users cannot exceed the total resources.
- Here, the virtual resources include the communication resource (e.g., *C2: the total bandwidth of each base station*), computation resource (e.g., *C5: the maximum amount of tasks simultaneously executed on the MEC server of each base station*), and caching resource (e.g., *C6: the storage space of each base station*).

Outline

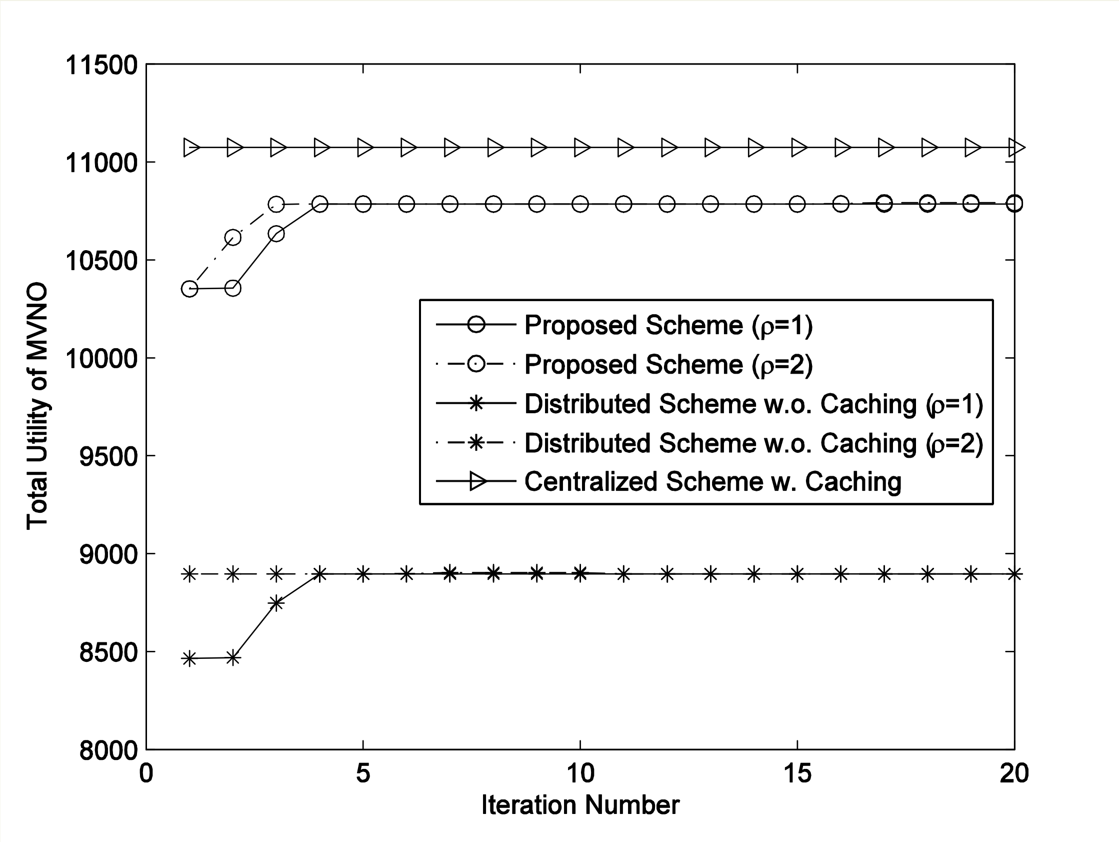
RA Process



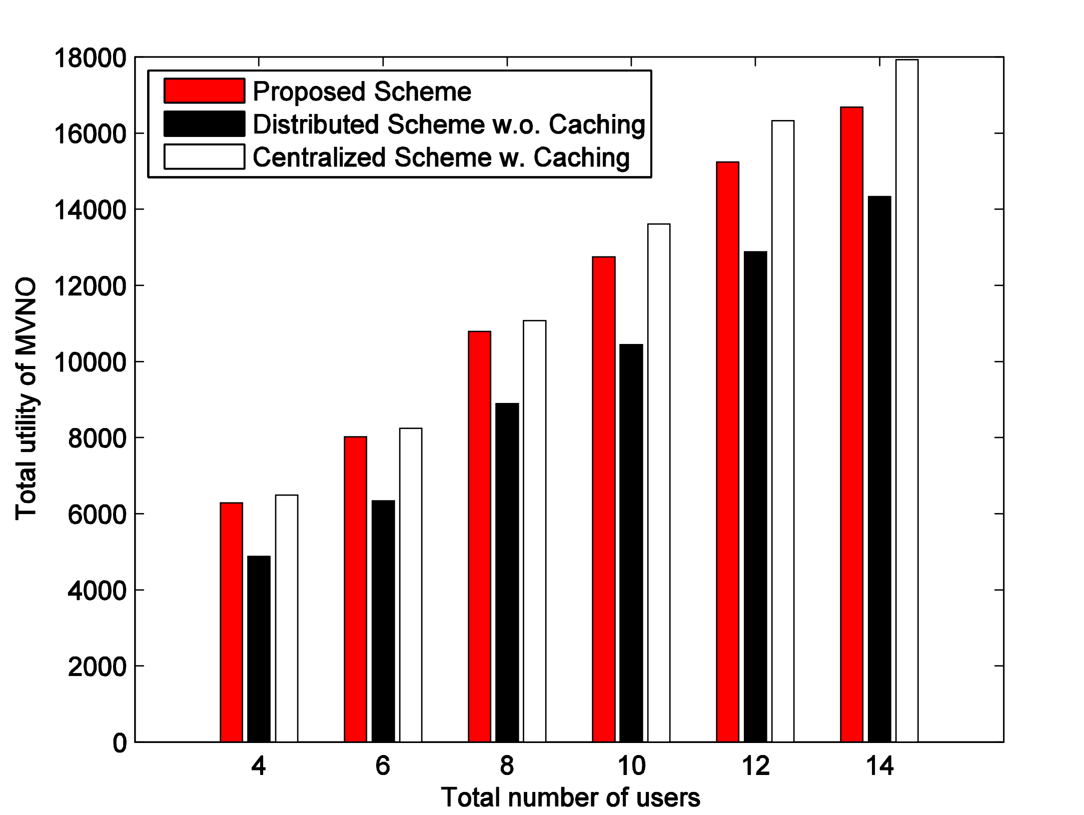
Start
Decoupling the association indicators
Updating for local variables
Recovery of local association indicators
Updating for global variables
Updating for dual variables
Stopping criterion
No
Yes
Recovery of global caching indicators
End

Outline

Convergence



Performance Comparison



Outline

- Introduction
- System Model
- Problem Formulation
- Resource Allocation Process
- Simulation Results
- Conclusion

Conclusion

- The virtual resource allocation for communication, computing and caching was studied in the designed virtualized HetNets framework.
- The allocation strategy was formulated as a joint optimization problem, considering the gains of not only virtualization but also caching and computing.
- A distributed ADMM-based algorithm was introduced to decouple the coupling variables and the split the optimization problem into several subproblems.
- Simulation results were presented to show the convergence and performance of the proposed scheme.

Future Work

- Future work is in progress to consider software-defined networking (SDN) in the proposed framework.

THANK YOU