Analysis and Design for Wireless Edge Networks with Caching, Computing, and Communication

Ming-Chun Lee
Institute of Communications Engineering
National Yang Ming Chiao Tung University

Contact Information

- Ming-Chun Lee
  - Assistant Professor
  - Institute of Communications Engineering
  - National Yang Ming Chiao Tung University

- Email: mingchunlee@nycu.edu.tw

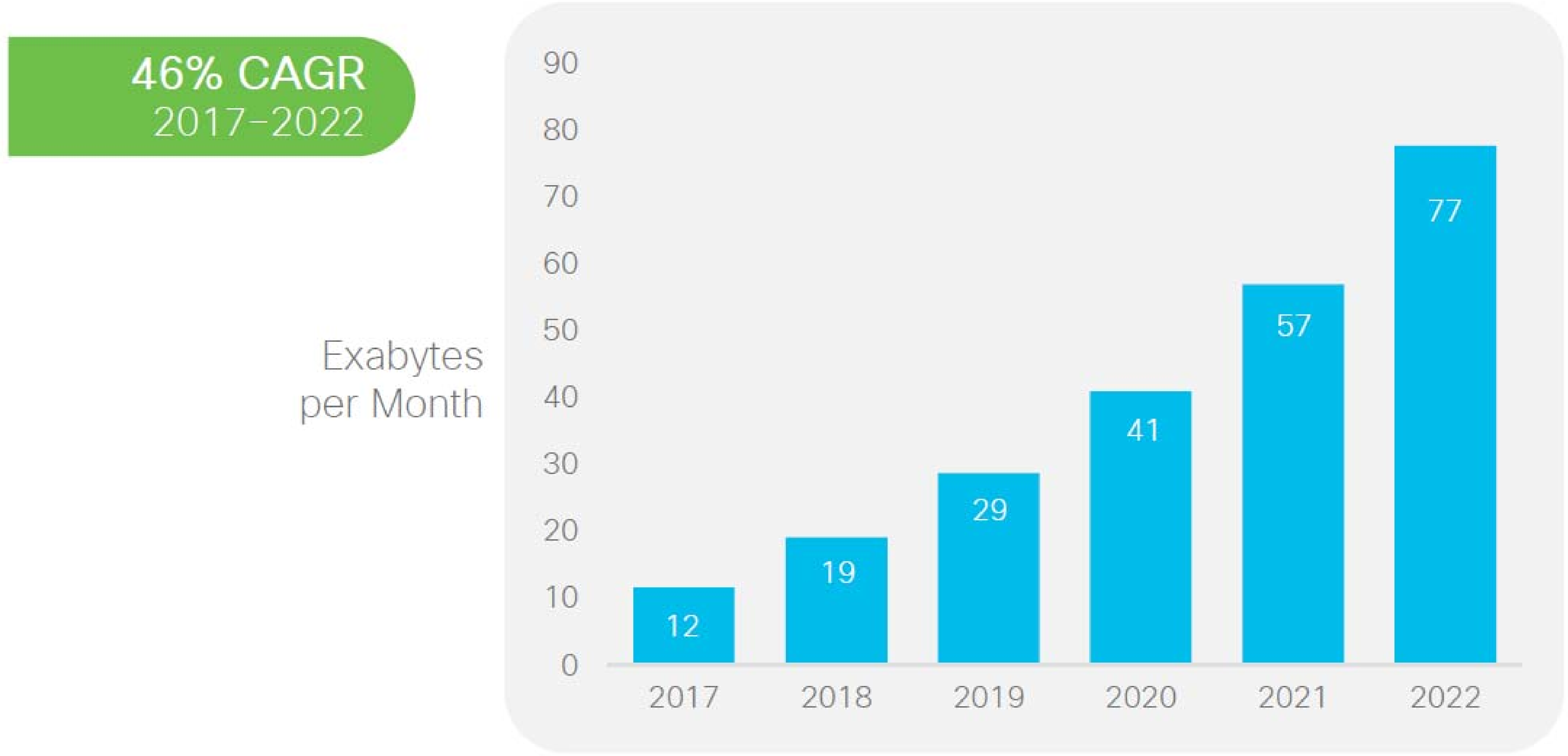
Outline

- Introduction to Edge-Caching and Edge-Computing
- Collaborative Caching, Computing, and Communication Supported Networks
- Research Examples
- Final Remarks

I Introduction

Challenges for Video Services

- Video traffic has increased significantly



Source: Cisco annual report

Source: Cisco VNI Global Mobile Data Traffic Forecast, 2017–2022

Video responsible for 66% of wireless traffic demand increase

- Operators need to decrease $/bit exponentially to prevent
  - Restriction on data usage
  - Lose money
  - Network collapse

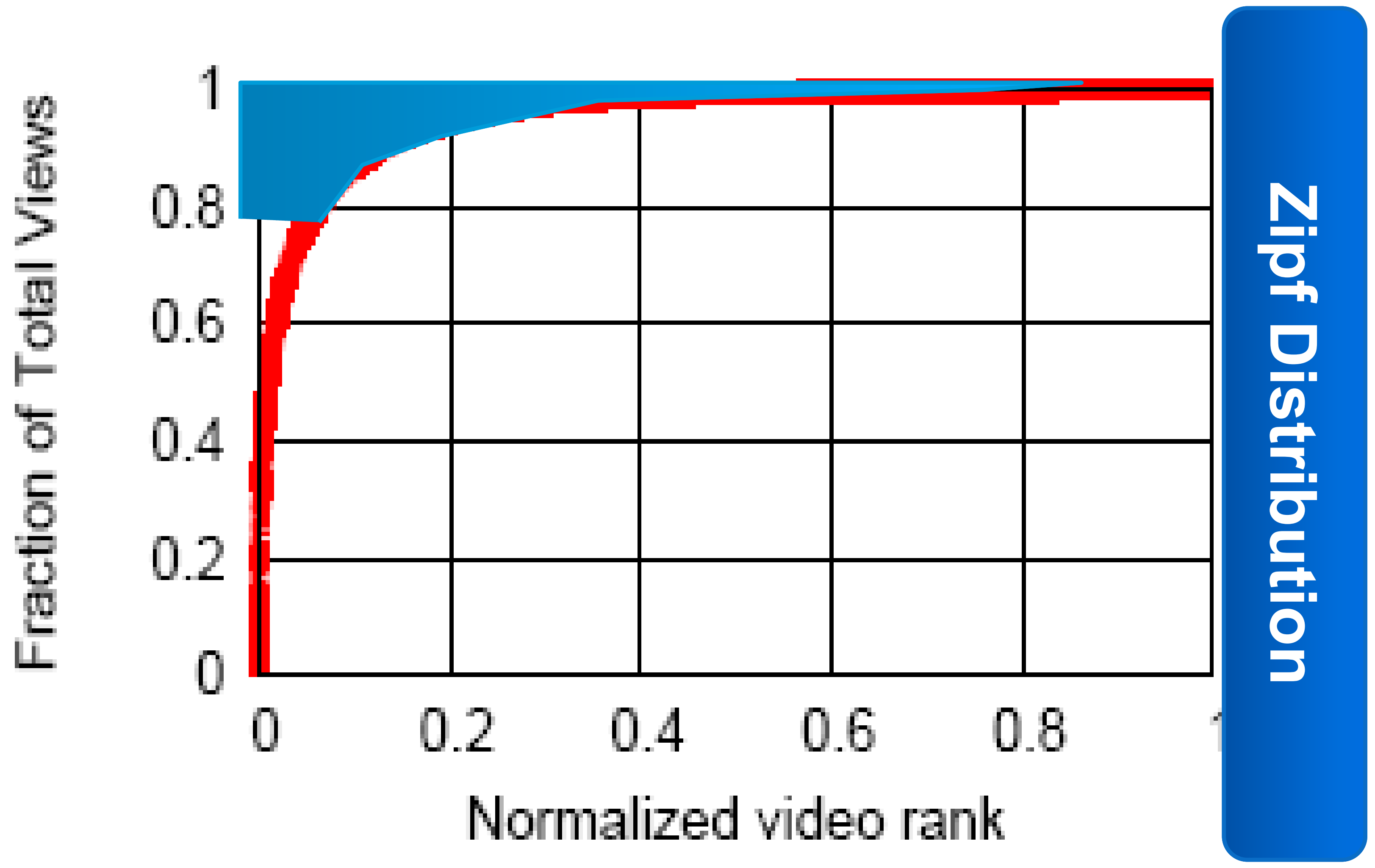
Physical Layer Aspect Limitations

- More spectrum
  - Expensive, limited to ~factor 2-5
- Better PHY layer
  - Single-link performance close to capacity
- Heterogeneous networks
  - Femtocells are promising but
    - Expensive
    - Backhaul can be bottleneck

-> NEW NETWORK STRUCTURE NEEDED

On-Demand Video Features

- Content reuse: a few popular files create a large percentage of video traffic
  - Viral YouTube videos
  - Popular movies on Netflix
  - Sport/News videos



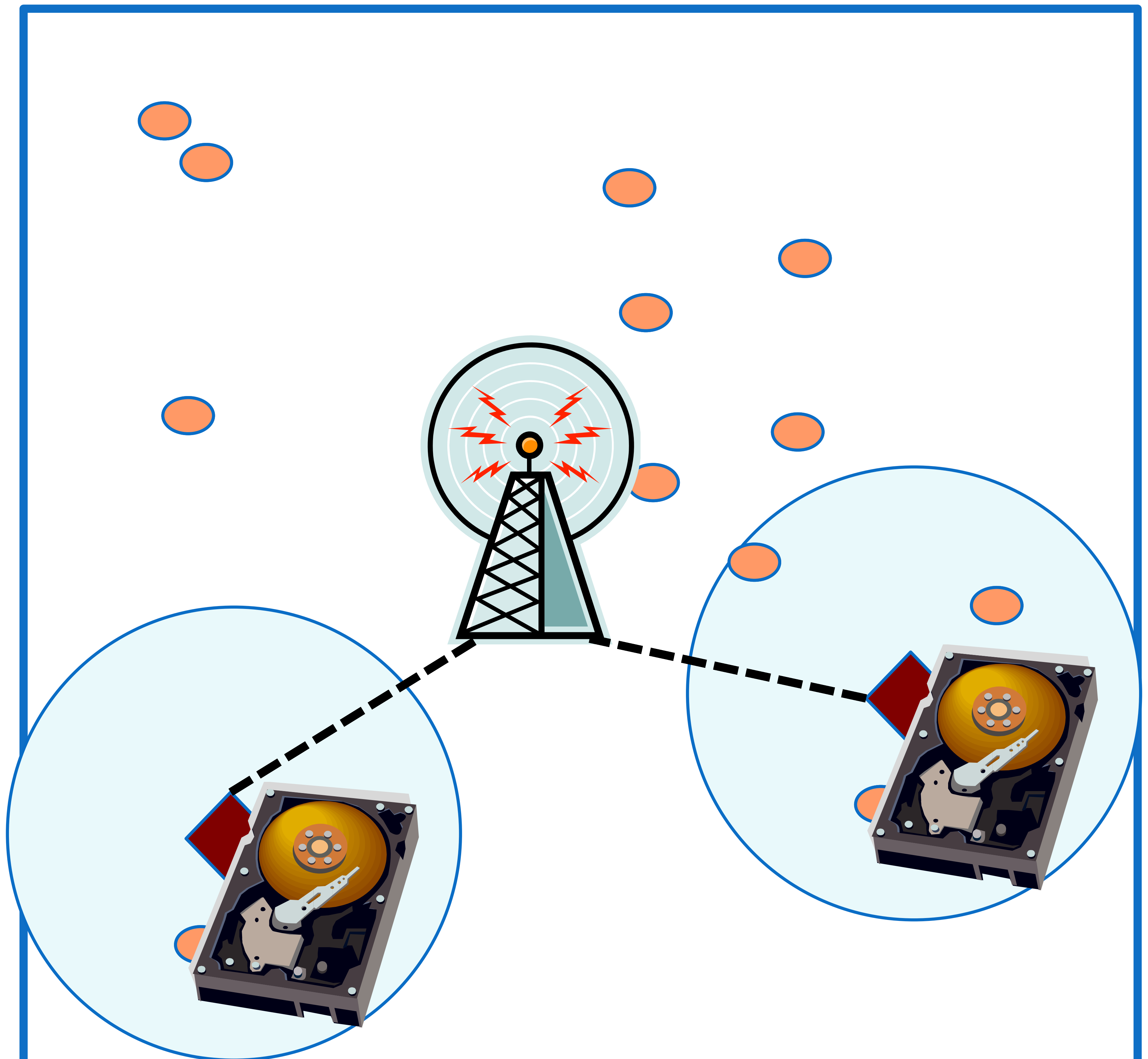
- Video broadcast/multicast failed

- *Is there a way to exploit video popularity while retaining on-demand capability?*

  YES: use storage (caching) close to requesting user!
  Known as Caching at the Wireless Edge (edge caching)

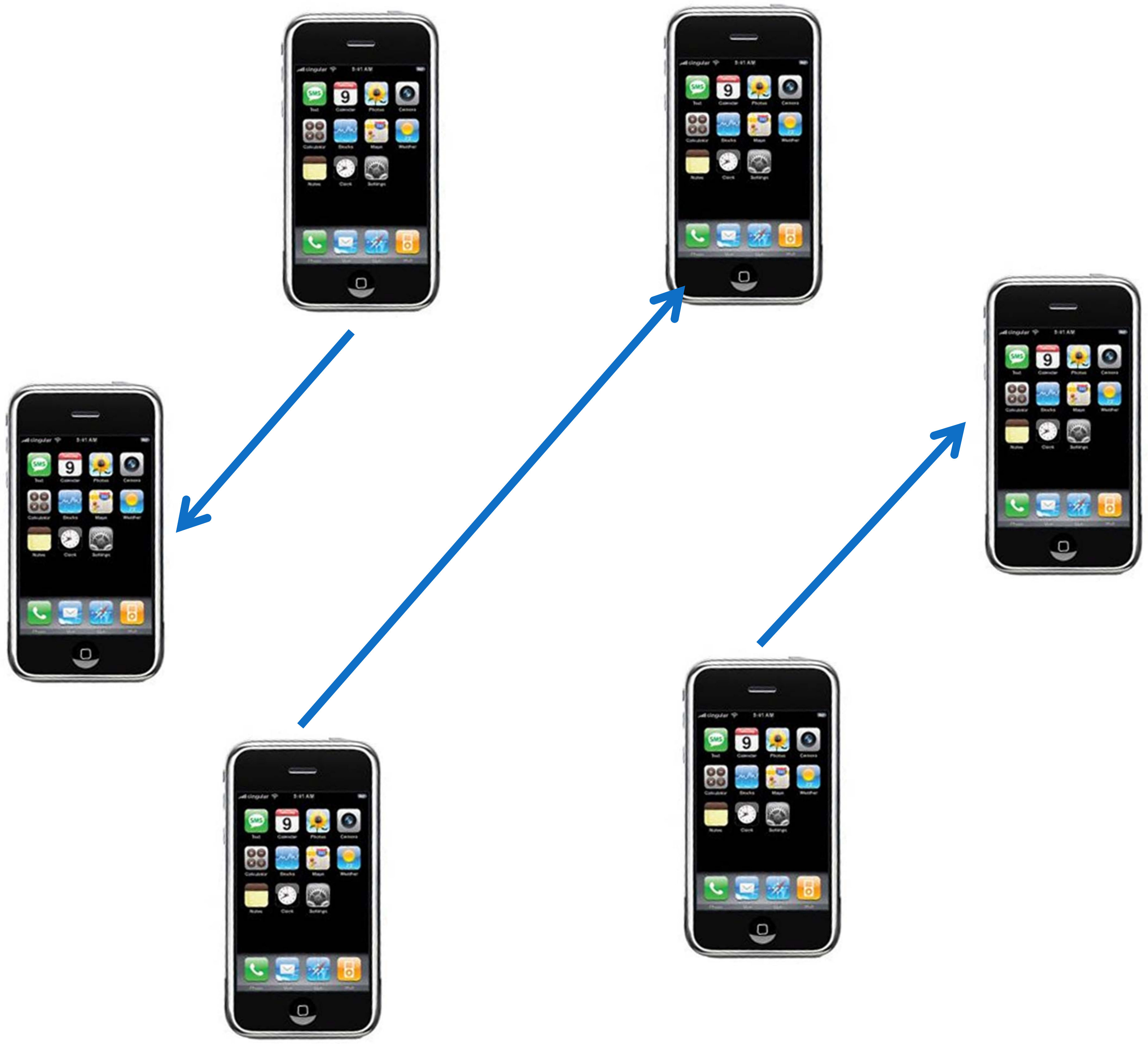
Edge-Caching Technologies

- Caching at the BSs



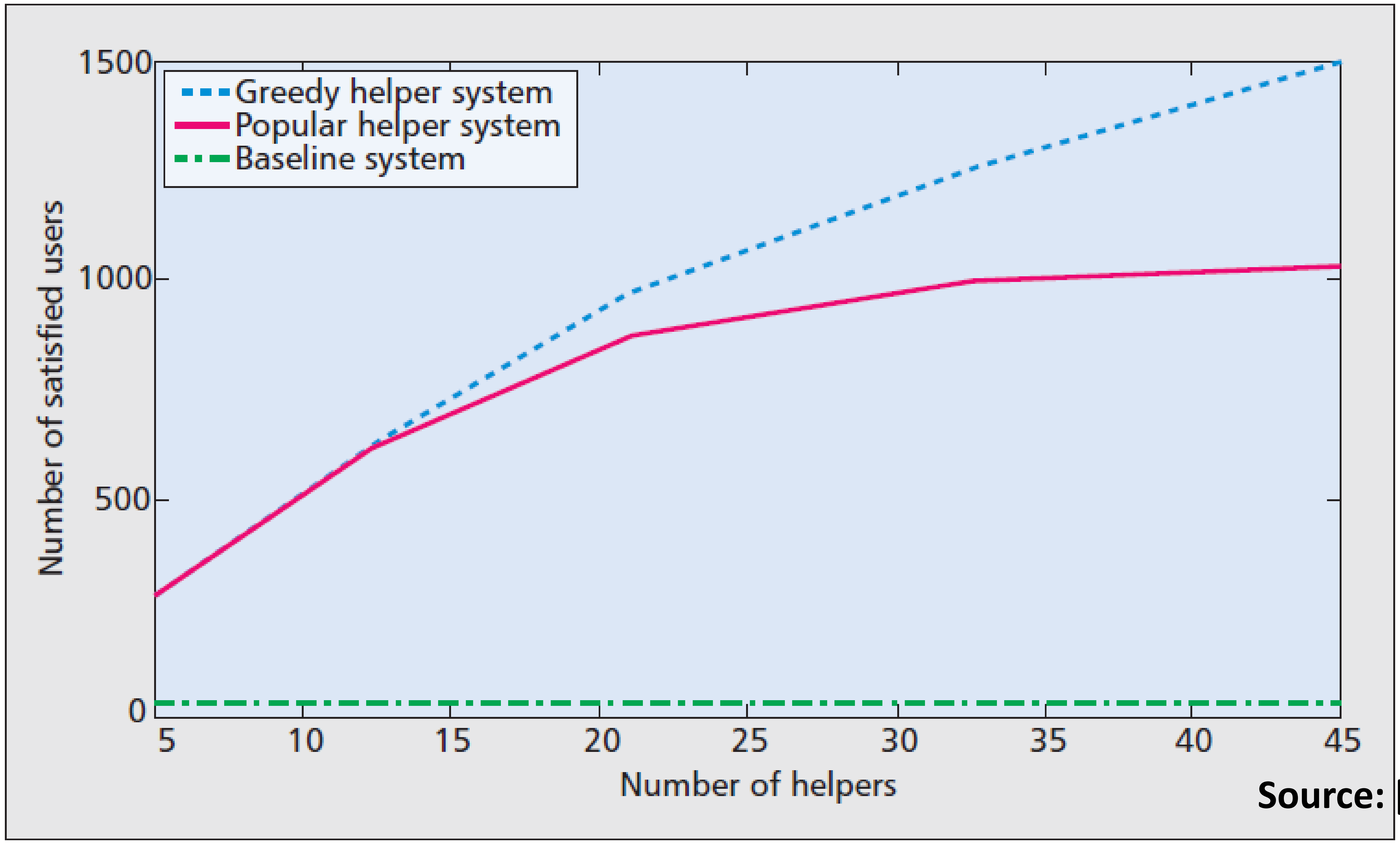
Source: ETC 2017, A. F. Molisch

- Caching at the Devices



Simulation Results

- Femtocaching
  - Replace backhaul with caching
  - Reduce network loading and latency



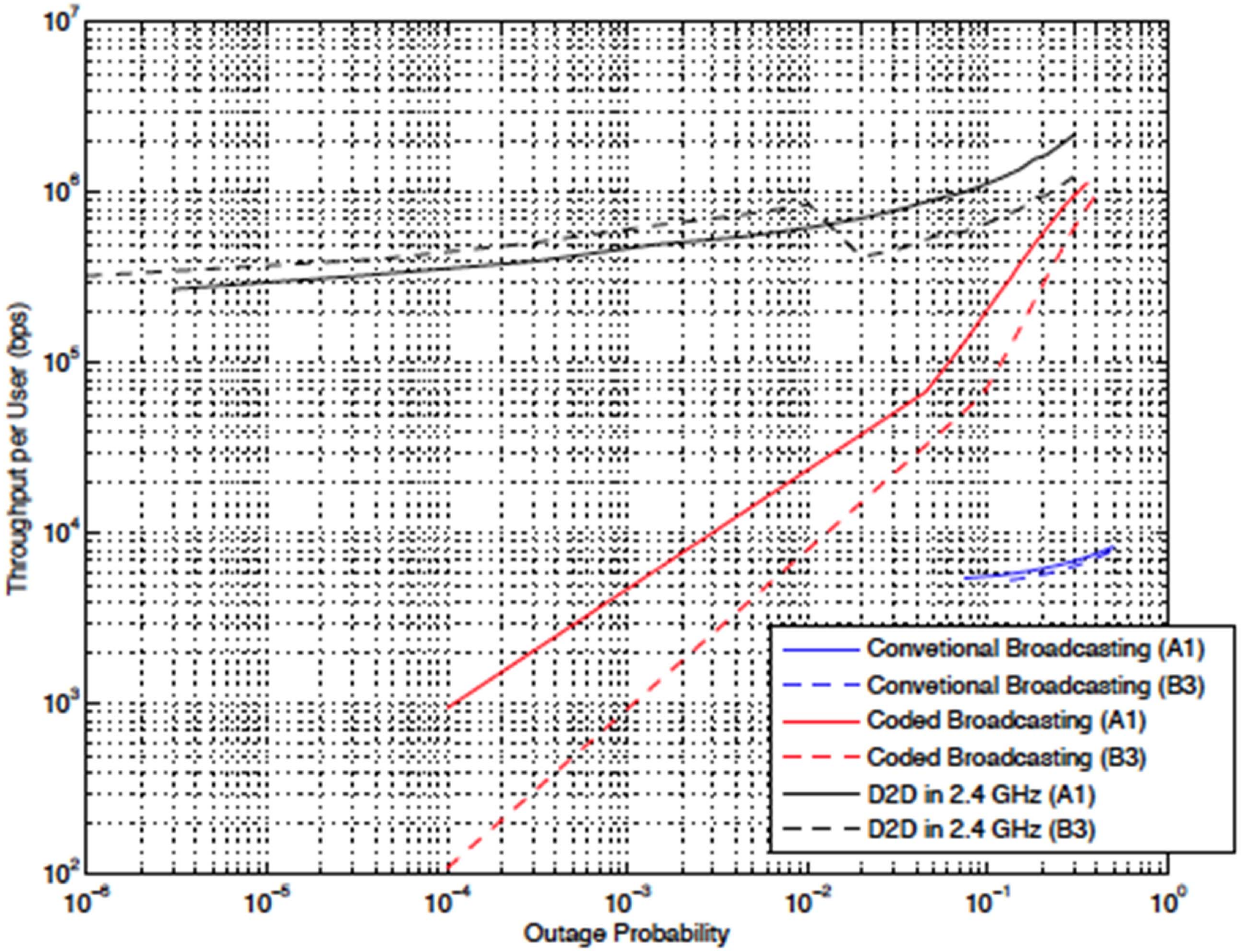
Source: [Gol]

The number of satisfied users versus the number of helpers, the cache capacity of each helper is 60GB, QOS=200 seconds.

[Gol]. N. Golrezai, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and Device-to-Device Collaboration: A New Architecture for Wireless Video Distribution", *IEEE Comm. Mag.*, 51, issue 4, 142-149 (2013).

Simulation Results

- Cache-aided D2D
  - Low-complexity of code construction
  - Uses high-capacity rate links (short distances)

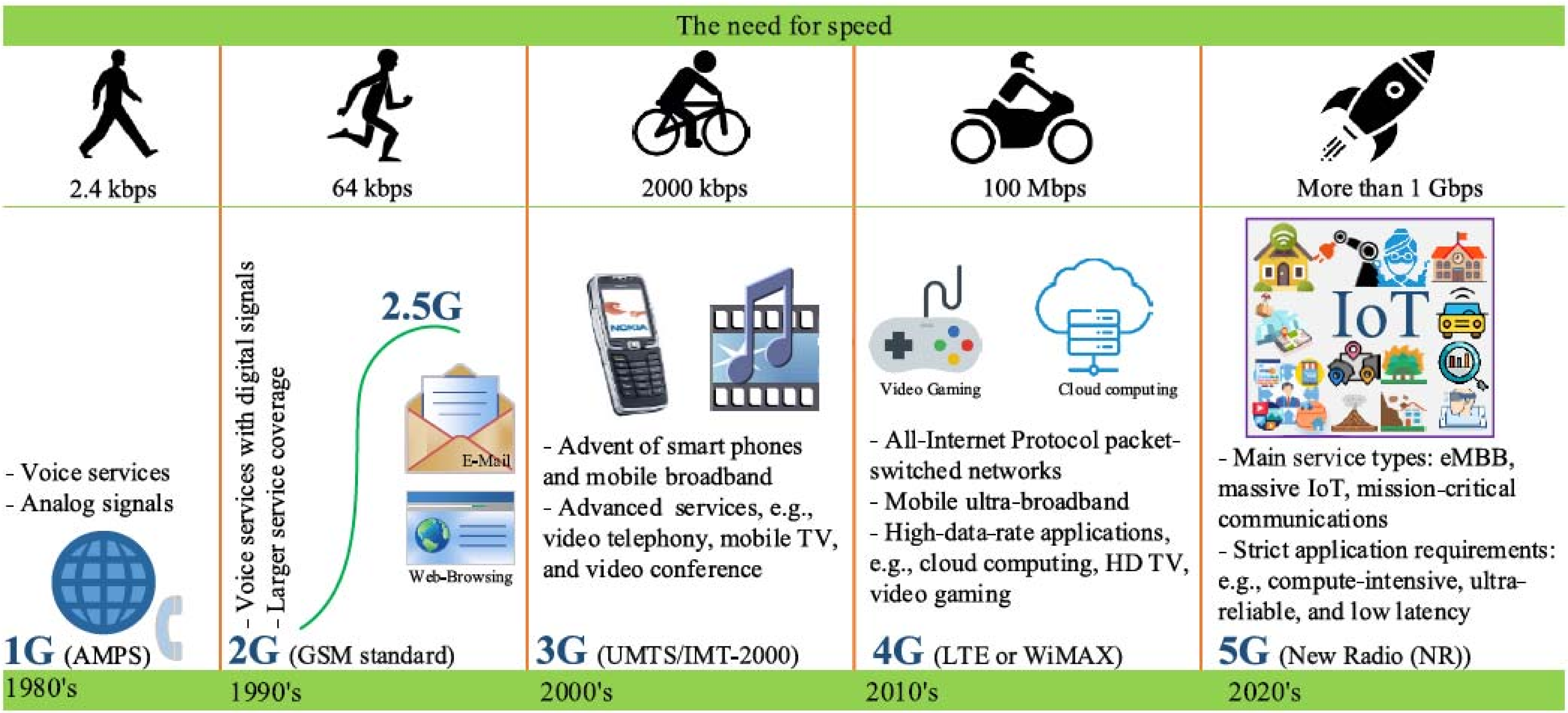


Source: [Ji, JSAC]

[Ji, JSAC]. M. Ji, G. Caire, and A. F. Molisch, "Wireless Device-to-Device Caching Networks: Basic Principles and System Performance", IEEE J. Selected Areas Comm., vol. 34, no. 1, pp. 176-189, Jan. 2016.

Evolution of Applications

- Compute-Intensive Tasks
  - AR/VR, gaming, AI-aided applications (e.g., face recognition), etc.

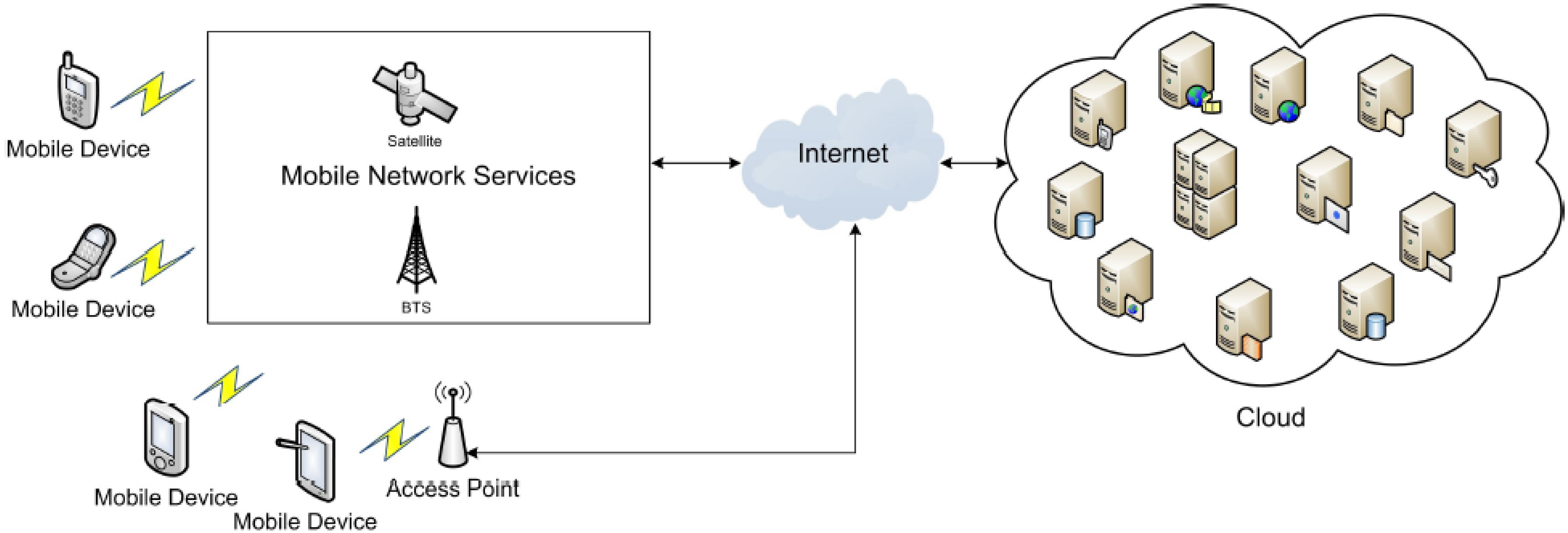


Source: [Pham, Access]

[Pham, Access]. Q. Pham *et al.*, "A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art," in *IEEE Access*, vol. 8, pp. 116974-117017, 2020, doi: 10.1109/ACCESS.2020.3001277.

Cloud Computing

- Resolve the computing power limitations for mobile devices
  - Computation offloading
- However, long latency due to travelling distance

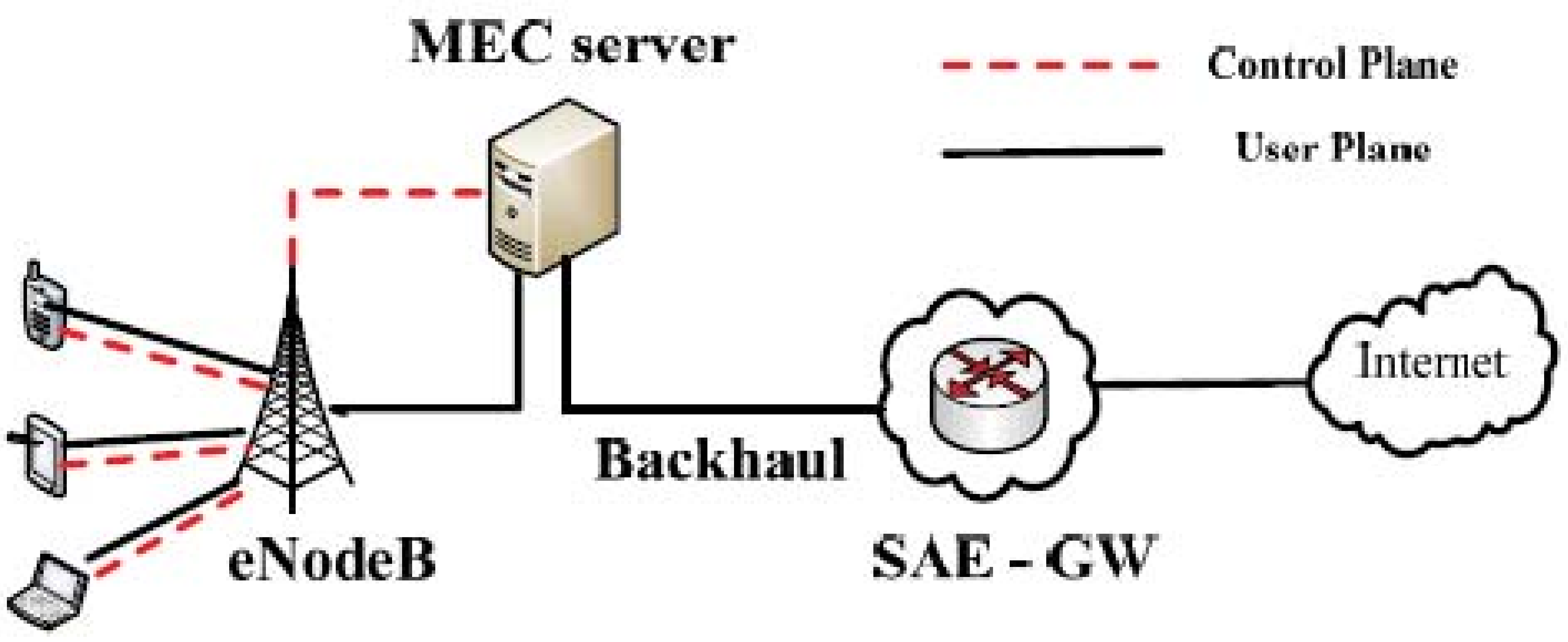


Source: [Khan, ComSuTu]

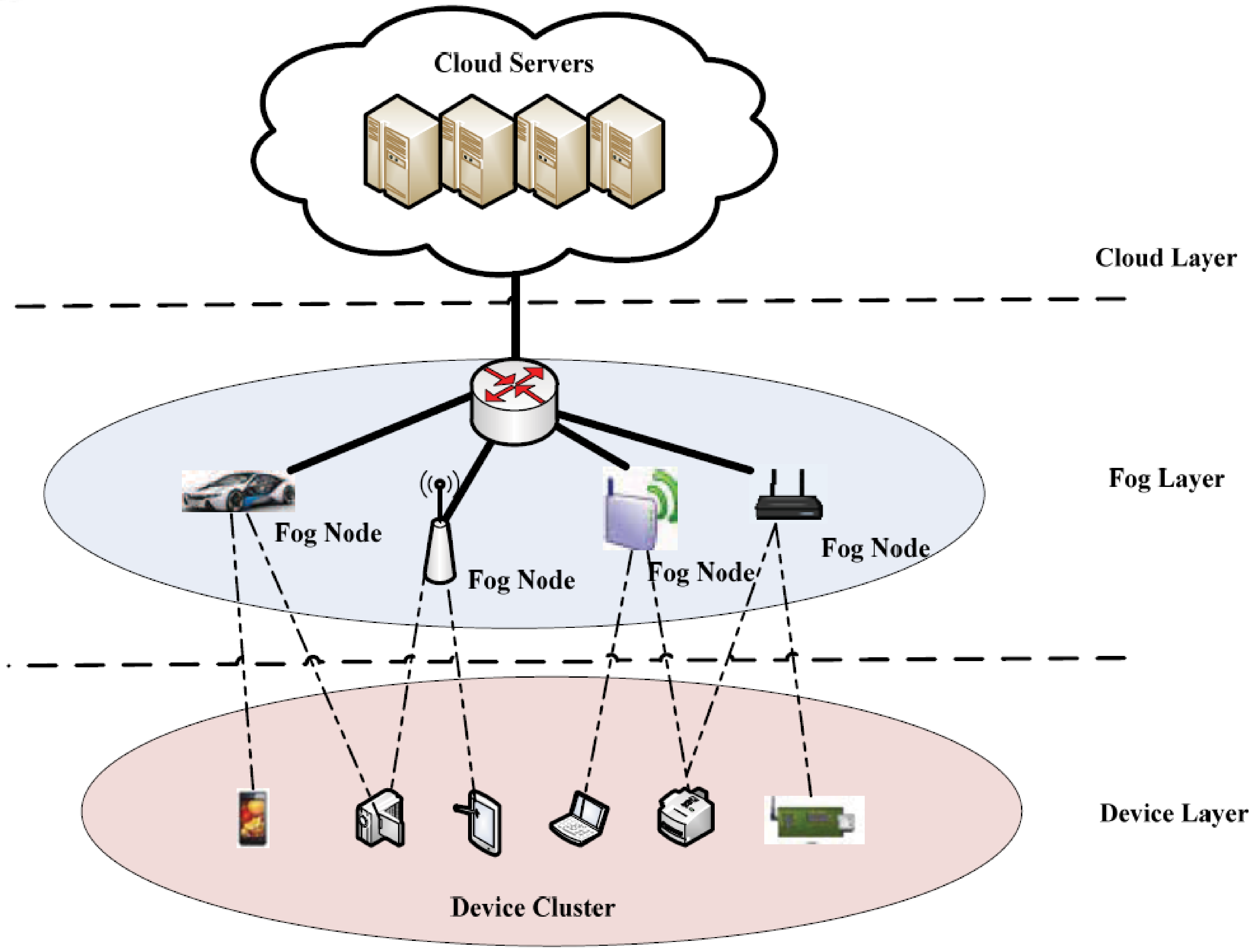
[Khan, ComSuTu]. A. u. R. Khan, M. Othman, S. A. Madani and S. U. Khan, "A Survey of Mobile Cloud Computing Application Models," in *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 393-413, First Quarter 2014, doi: 10.1109/SURV.2013.062613.00160.

Mobile Edge Computing

- Resolve the long latency by bringing the computing units closer to the users (at the edge)



Source: [Wang, Access]



[Wang, Access]. S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang and W. Wang, "A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications," in *IEEE Access*, vol. 5, pp. 6757-6779, 2017, doi: 10.1109/ACCESS.2017.2685434..

Case Study for Edge Computing

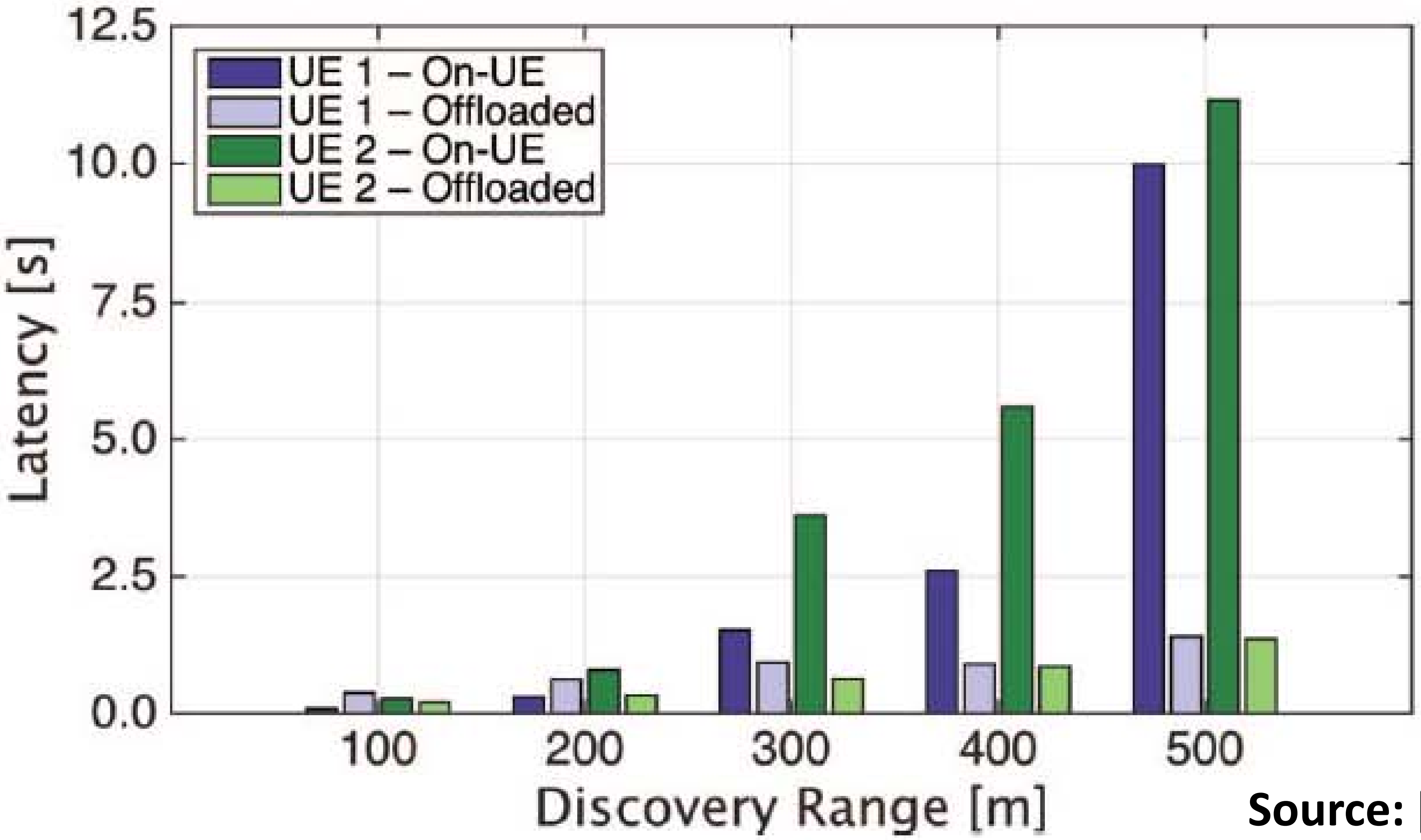
- AR application to discover and render visible places in user's cone of vision [Dolezal, CSCN]

Source: [Dolezal, CSCN]

[Dolezal, CSCN]. J. Dolezal, Z. Becvar and T. Zeman, "Performance evaluation of computation offloading from mobile device to the edge of mobile network," *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2016, pp. 1-7, doi: 10.1109/CSCN.2016.7785153.

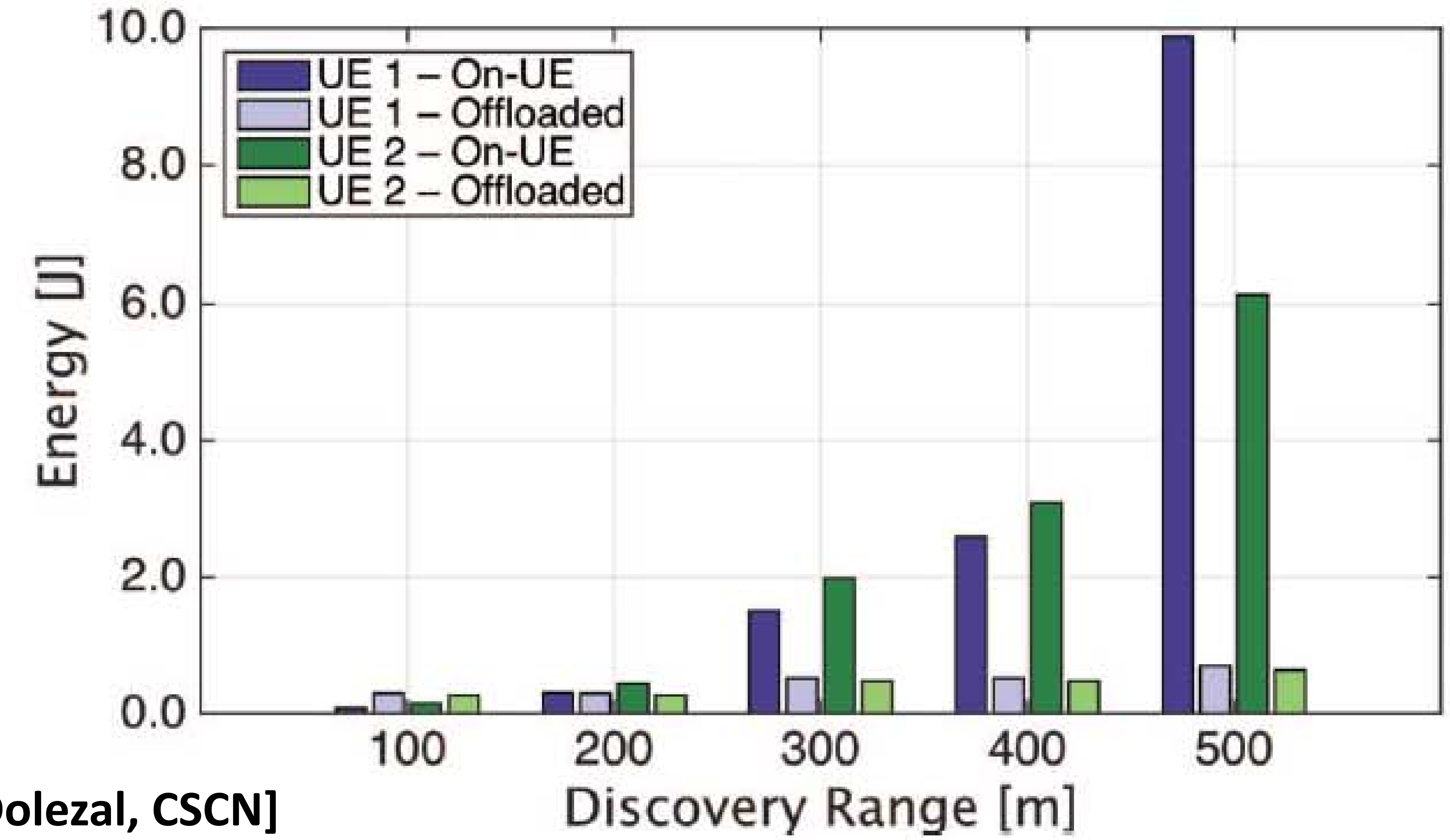
Case Study for Edge Computing

- Latency and Energy Consumption Improvements



(a)

Source: [Dolezal, CSCN]

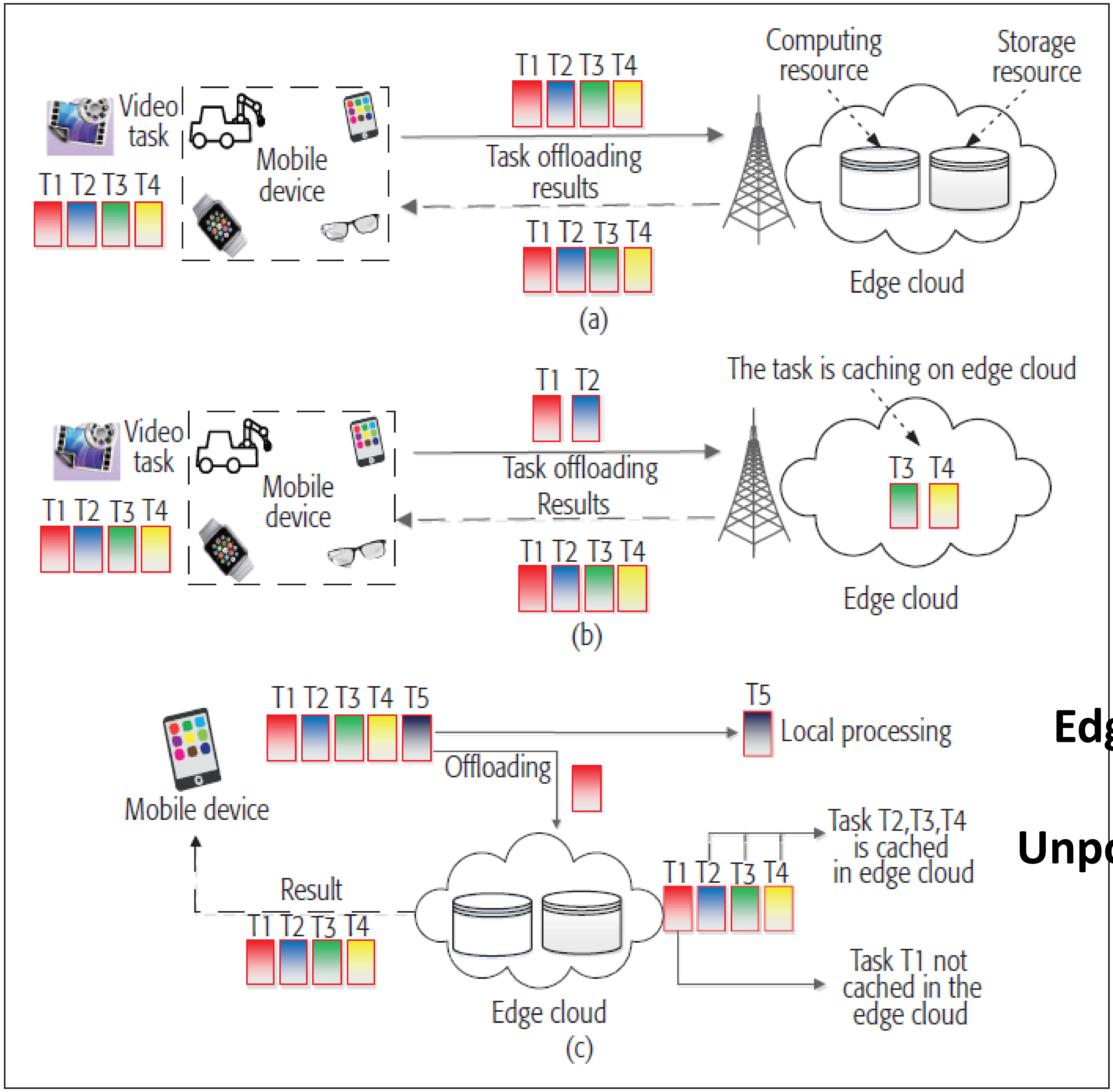


(b)

[Dolezal, CSCN]. J. Dolezal, Z. Becvar and T. Zeman, "Performance evaluation of computation offloading from mobile device to the edge of mobile network," *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2016, pp. 1-7, doi: 10.1109/CSCN.2016.7785153.

II Collaborative Caching, Computing, and Communication Supported Networks

Concept of 3C



Source: [Chen, WirelessMag]

Edge-Computing:
Tasks are done via data + computing

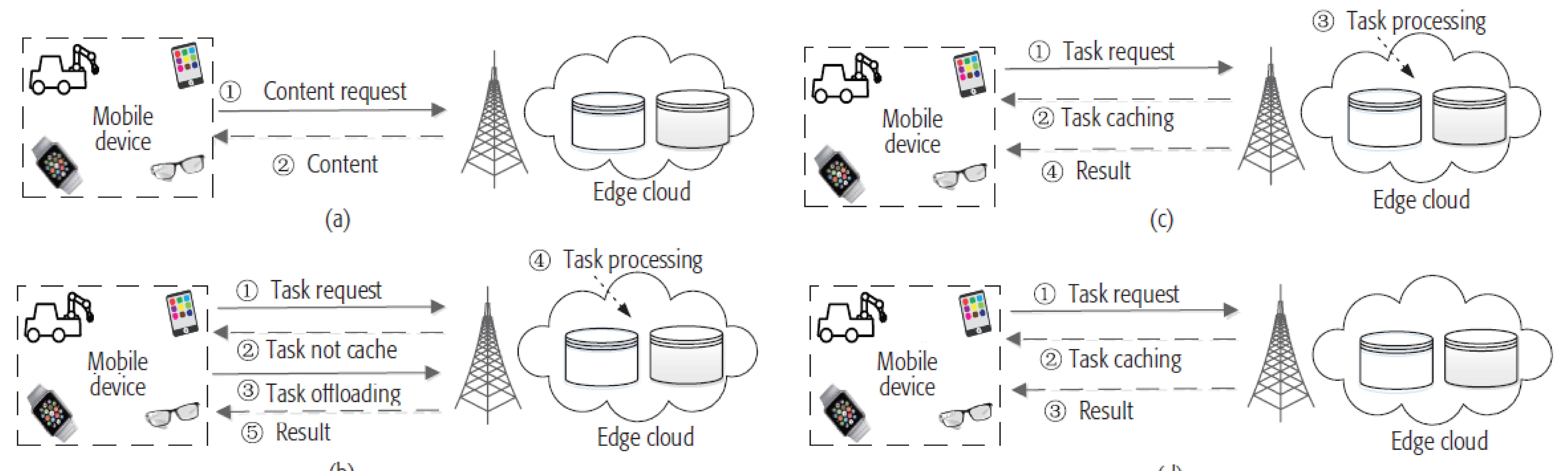
Edge-Caching:
Tasks are directly cached

Edge-Caching + Edge-computing:
Popular tasks are cached
Unpopular tasks can be retrieved via data + computing

[Chen, WirelessMag]. M. Chen, Y. Hao, L. Hu, M. S. Hossain and A. Ghoneim, "Edge-CoCaCo: Toward Joint Optimization of Computation, Caching, and Communication on Edge Cloud," in *IEEE Wireless Communications*, vol. 25, no. 3, pp. 21-27, JUNE 2018, doi: 10.1109/MWC.2018.1700308.

Concept of 3C

- General processing procedure
  - (a). Content cached
  - (b). Task not cached (intense computing + data retrieval)
  - (c). Task cached (minor computing)
  - (d). Task result cached

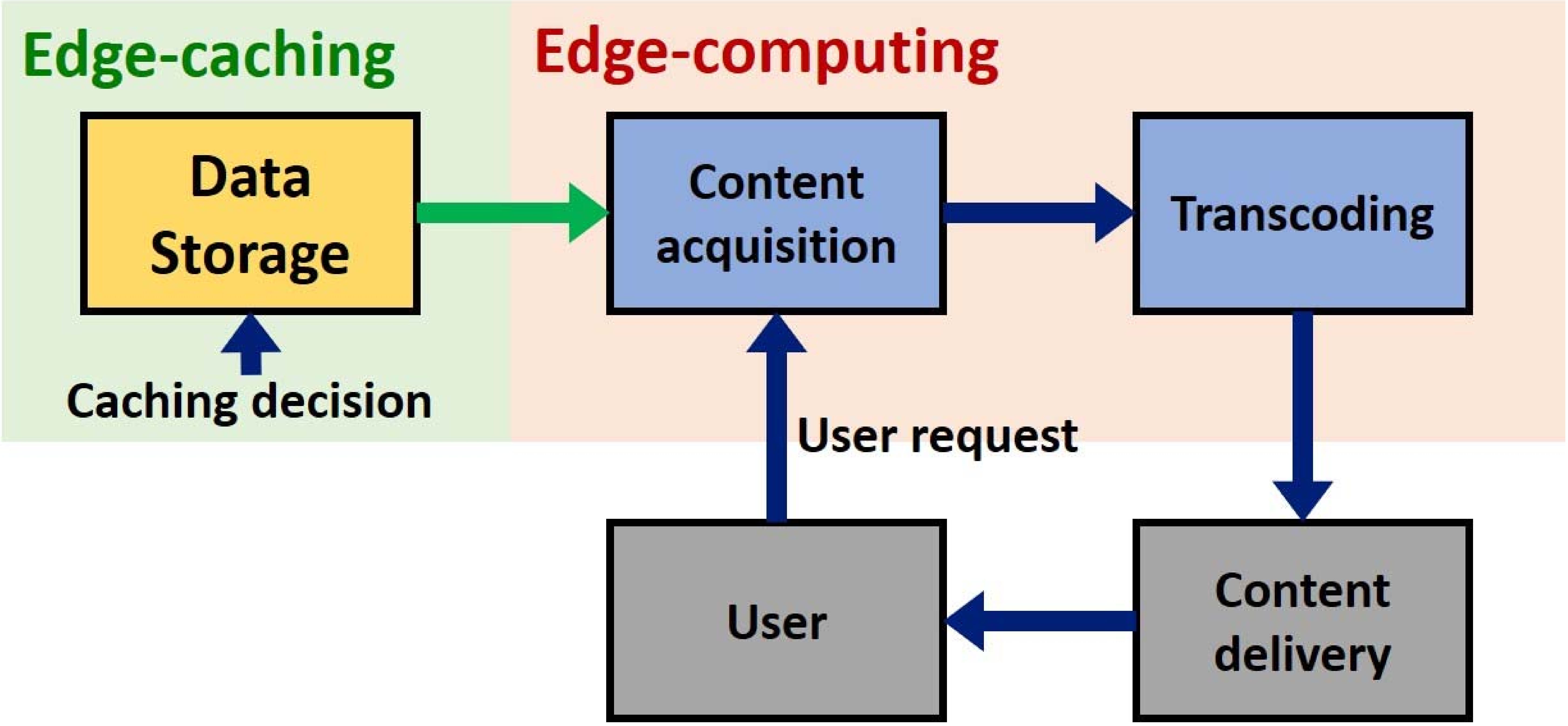


Source: [Chen, WirelessMag]

[Chen, WirelessMag]. M. Chen, Y. Hao, L. Hu, M. S. Hossain and A. Ghoneim, "Edge-CoCaCo: Toward Joint Optimization of Computation, Caching, and Communication on Edge Cloud," in *IEEE Wireless Communications*, vol. 25, no. 3, pp. 21-27, JUNE 2018, doi: 10.1109/MWC.2018.1700308.

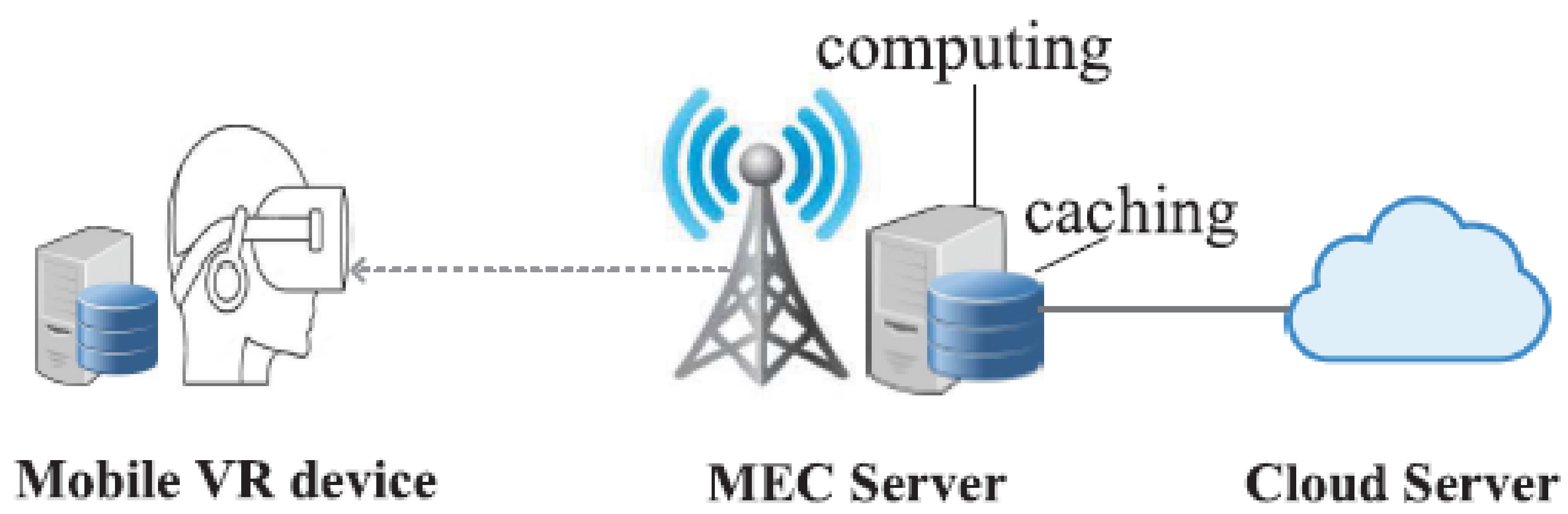
Motivating Applications

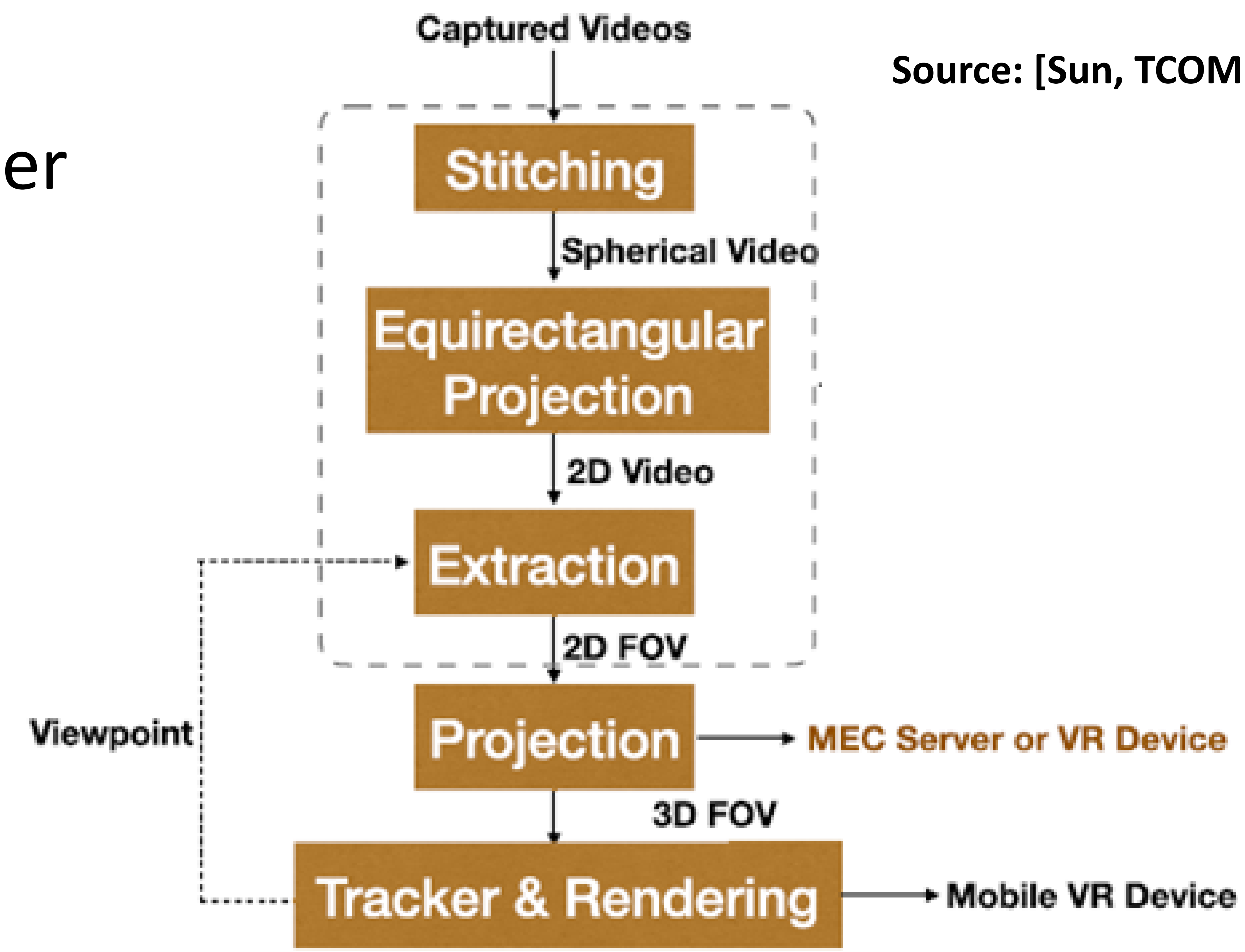
- High-definition video streaming
  - Video contents with different qualities are cached
  - When requested, the content is transcoded and then delivered
    - Low to high quality: super-resolution technique
    - High to low quality: video compression



Motivating Applications

- Virtual reality (VR)
  - Viewpoint is uploaded
  - 2D images are accessed
  - Viewpoint is projected to construct the 3D image
  - Result delivered to the user



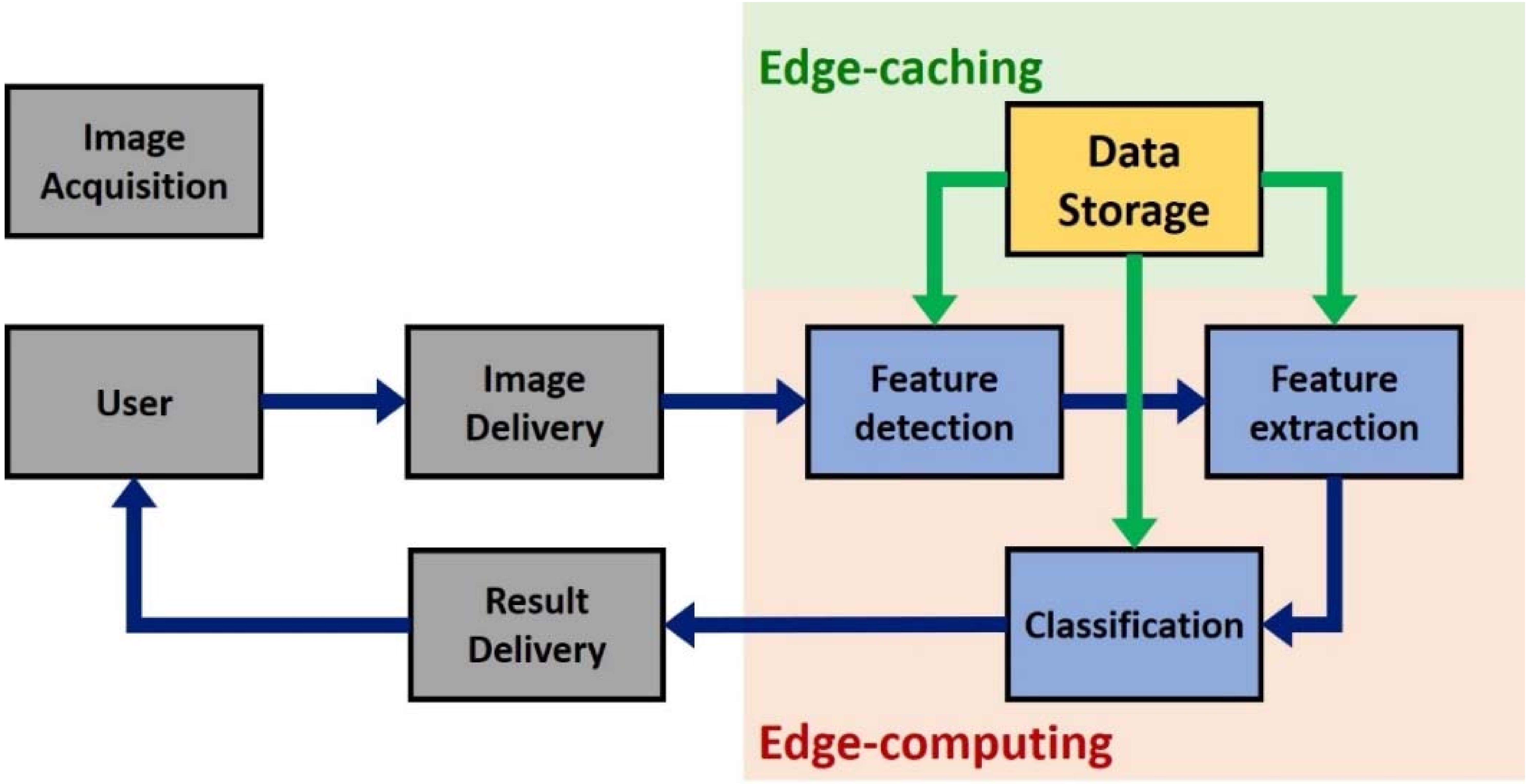


Source: [Sun, TCOM]

[Sun, TCOM]. Y. Sun, Z. Chen, M. Tao and H. Liu, "Communications, Caching, and Computing for Mobile Virtual Reality: Modeling and Tradeoff," in *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7573-7586, Nov. 2019, doi: 10.1109/TCOMM.2019.2920594.

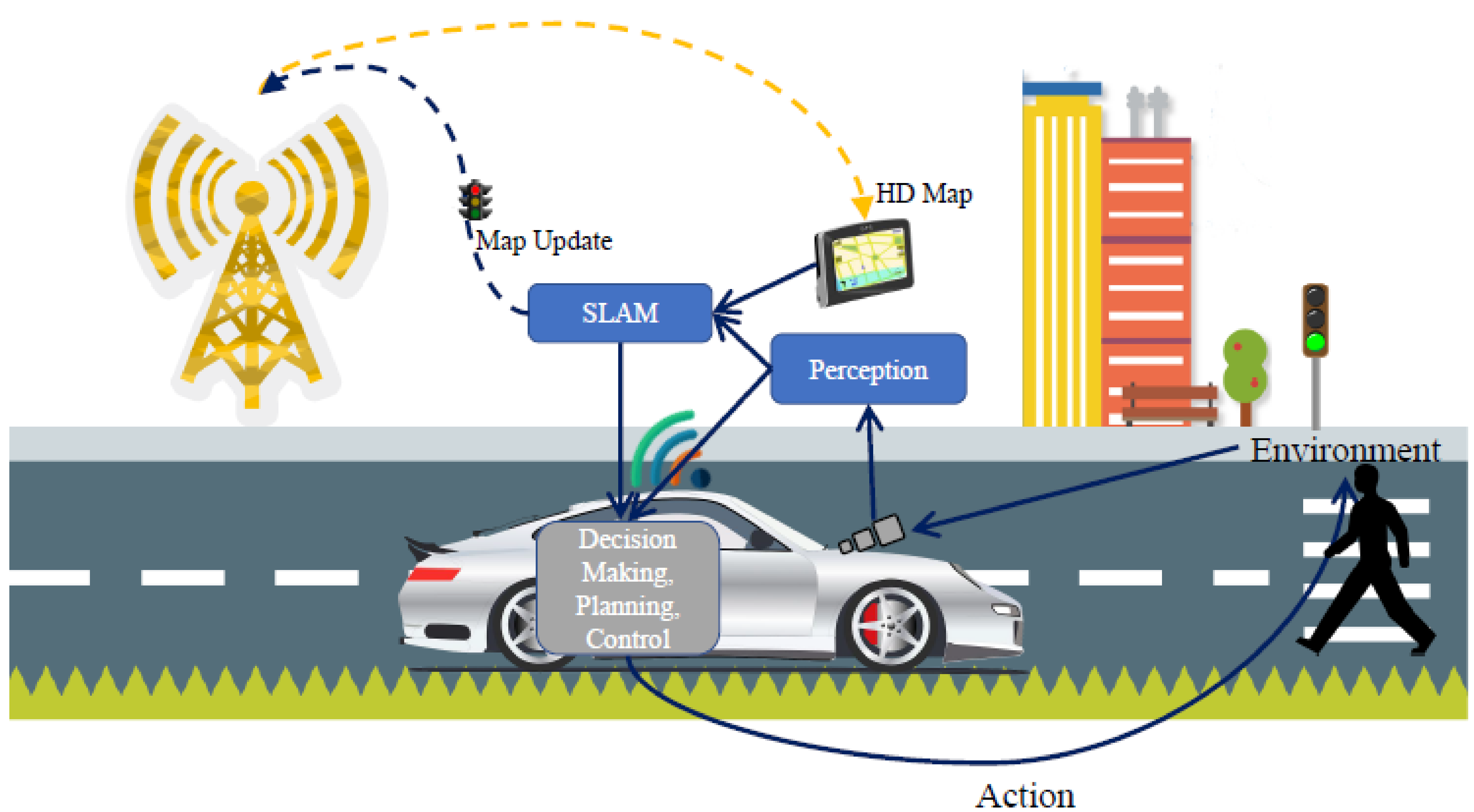
Motivating Applications

- Face recognition (AI-aided Inference)
  - Image upload for face recognition
  - AI-aided approach, e.g., NN, for conducting face recognition
  - Result delivered to the user



Motivating Applications

- Intelligent vehicles [Zhang, Proceeding]
  - Perceptron: Estimate the environment model with on-board sensors, e.g., object detection and tracking, lane detection
  - HD mapping: Three-dimensional representation of all crucial aspects of a roadway
  - SLAM: Simultaneous estimation of the location of a vehicle and the construction of the map

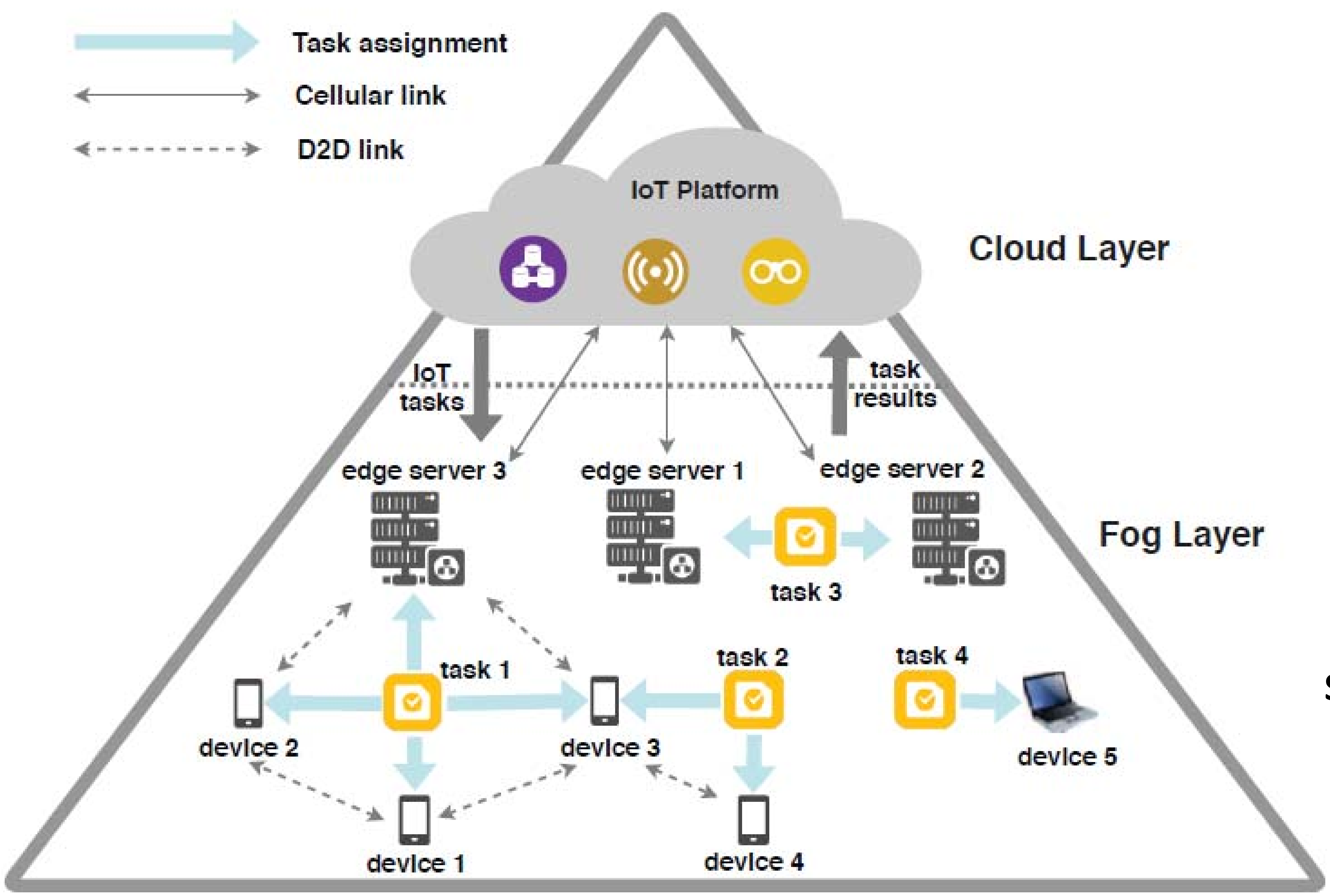


Source: [Zhang, Proceeding]

[Zhang, Proceeding]. J. Zhang and K. B. Letaief, "Mobile Edge Intelligence and Computing for the Internet of Vehicles," in *Proceedings of the IEEE*, vol. 108, no. 2, pp. 246-261, Feb. 2020, doi: 10.1109/JPROC.2019.2947490.

Motivating Applications

- 3C with IoTs
  - IoT devices are commonly with low computing and caching capabilities
  - Still need to conduct different tasks

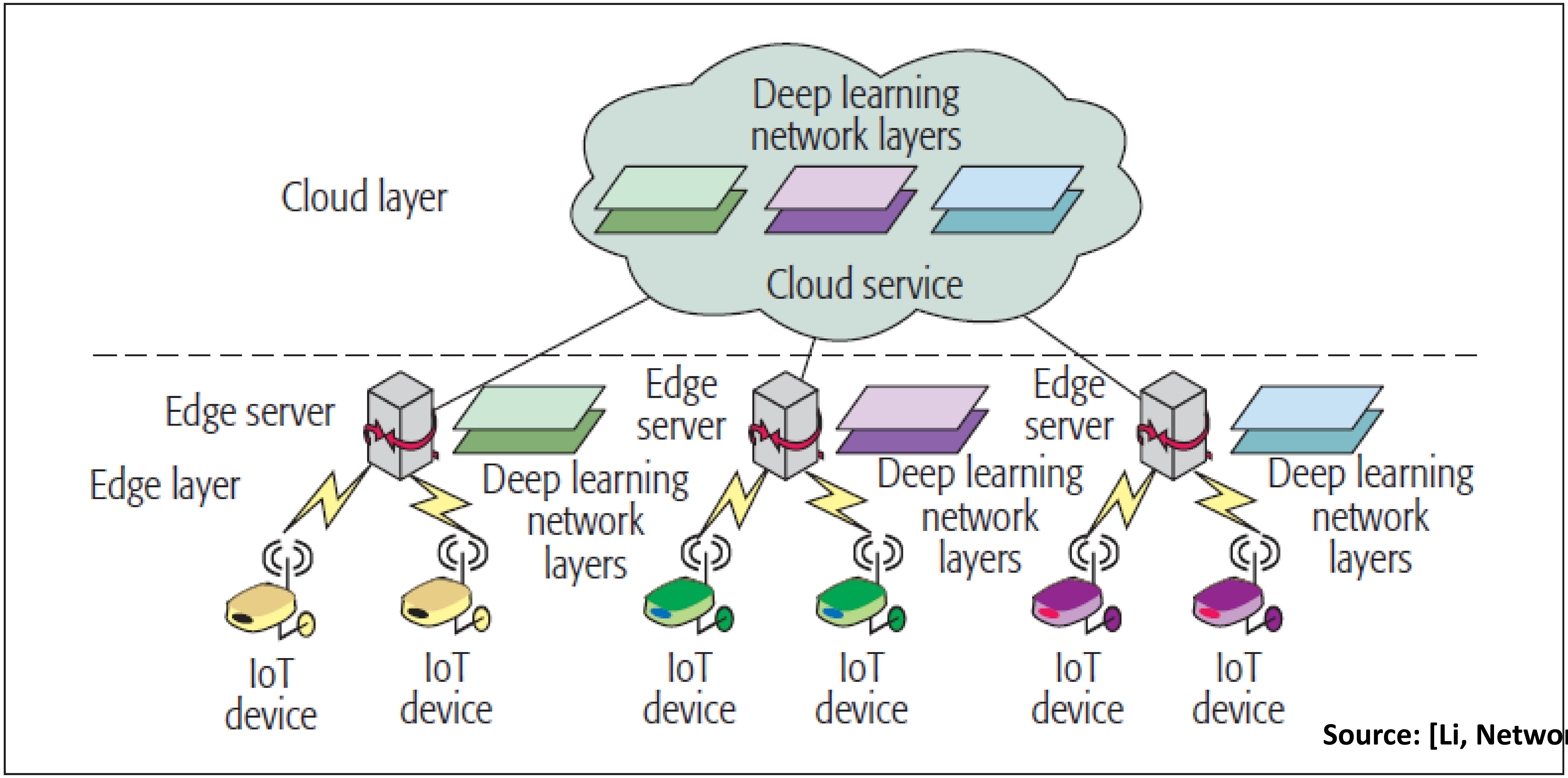


Source: [Luo, *ICDCS*]

[Luo, *ICDCS*]. S. Luo, X. Chen and Z. Zhou, "F3C: Fog-enabled Joint Computation, Communication and Caching Resource Sharing for Energy-Efficient IoT Data Stream Processing," *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 2019, pp. 1019-1028, doi: 10.1109/ICDCS.2019.00105.

Motivating Applications

- AI-aided IoTs
  - Use DNNs to conduct inference
  - Split the DNN layers and conduct computing at different levels

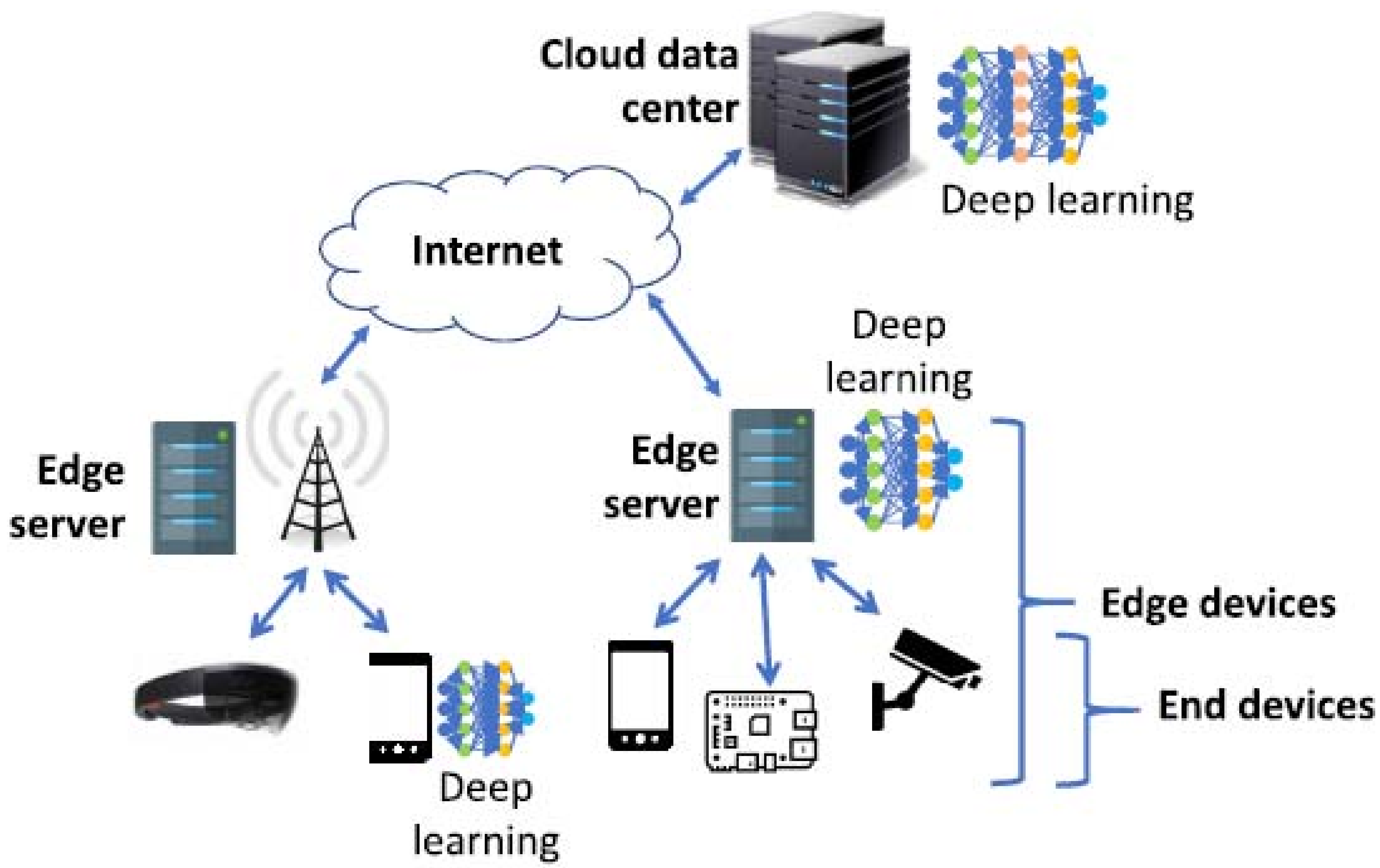


Source: [Li, NetworkMag]

[Li, NetworkMag]. H. Li, K. Ota and M. Dong, "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing," in *IEEE Network*, vol. 32, no. 1, pp. 96-101, Jan.-Feb. 2018, doi: 10.1109/MNET.2018.1700202.

Edge-AI

- Deep learning can help in many applications
  - It has been shown that deep learning can bring intelligence to the devices, systems, and networks
  - To have device intelligence, 3C is necessary

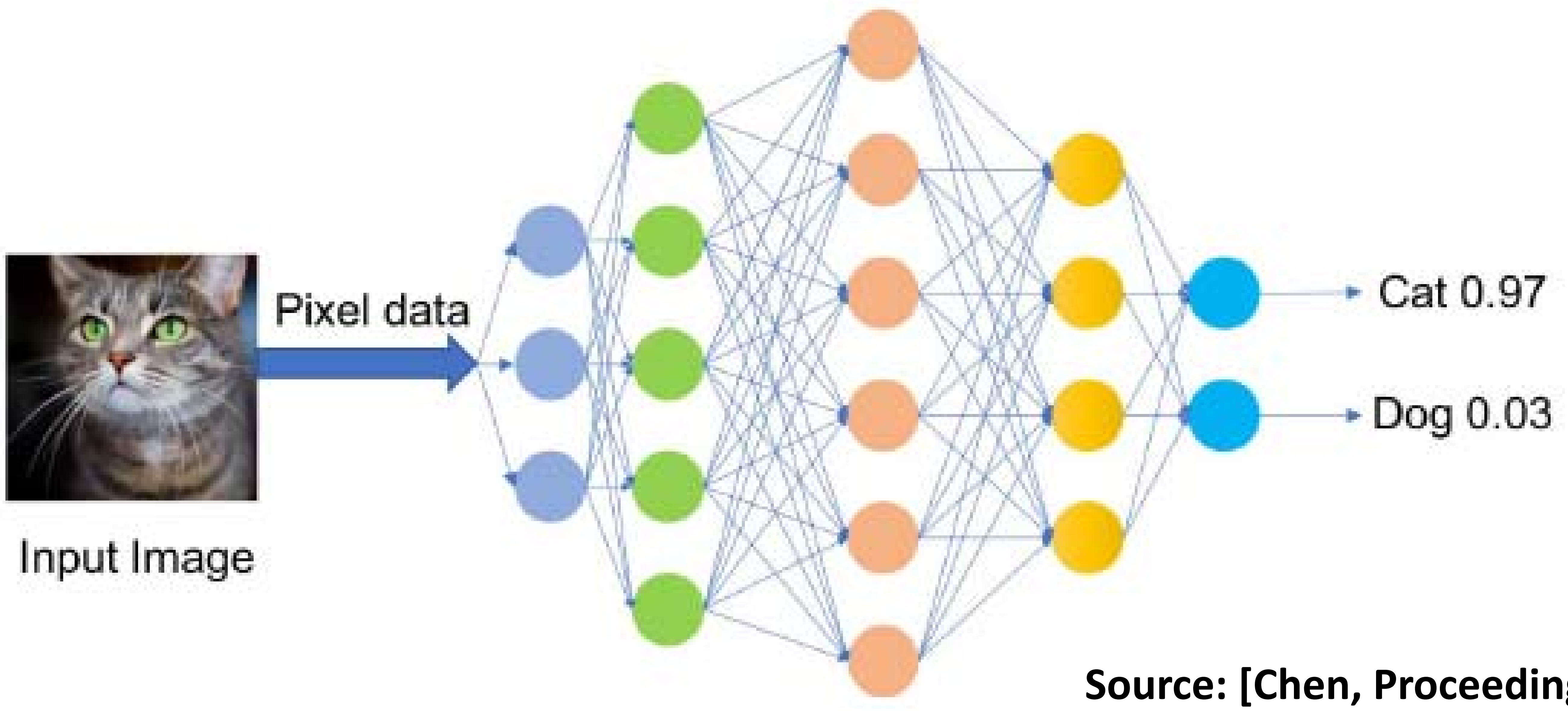


Source: [Chen, Proceeding]

[Chen, Proceeding]. J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review," in *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655-1674, Aug. 2019, doi: 10.1109/JPROC.2019.2921977.

Edge-AI

- Suppose we would like to conduct a classification task
  - Four-layer DNN

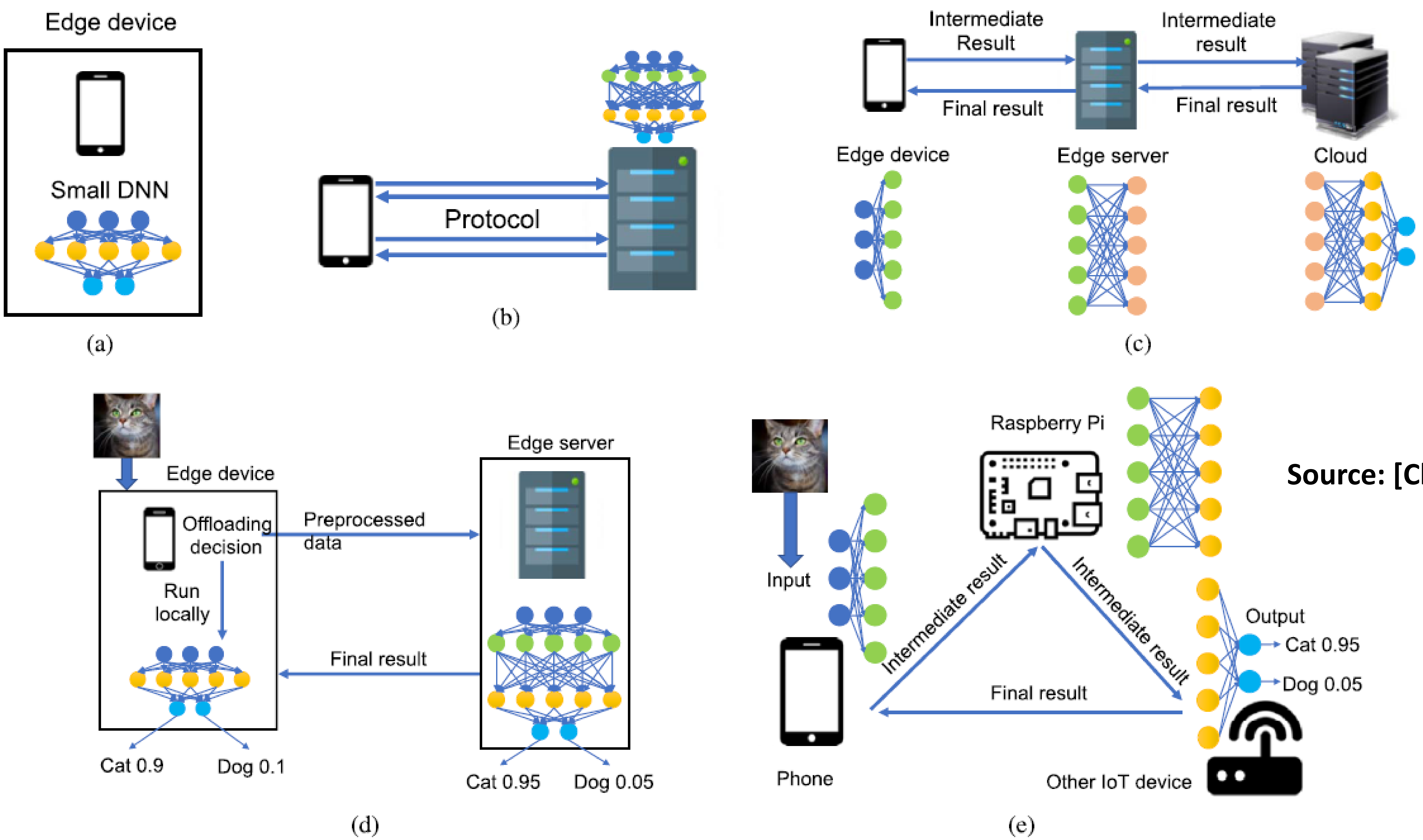


Source: [Chen, Proceeding]

[Chen, Proceeding]. J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review," in *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655-1674, Aug. 2019, doi: 10.1109/JPROC.2019.2921977.

Edge-AI

- Edge-AI modes
  - (a) On-device computation; (b) Edge computing; (c) Computing with DNN model splitting; (d) Offloading with model selection; (e) Distributed computing with model splitting



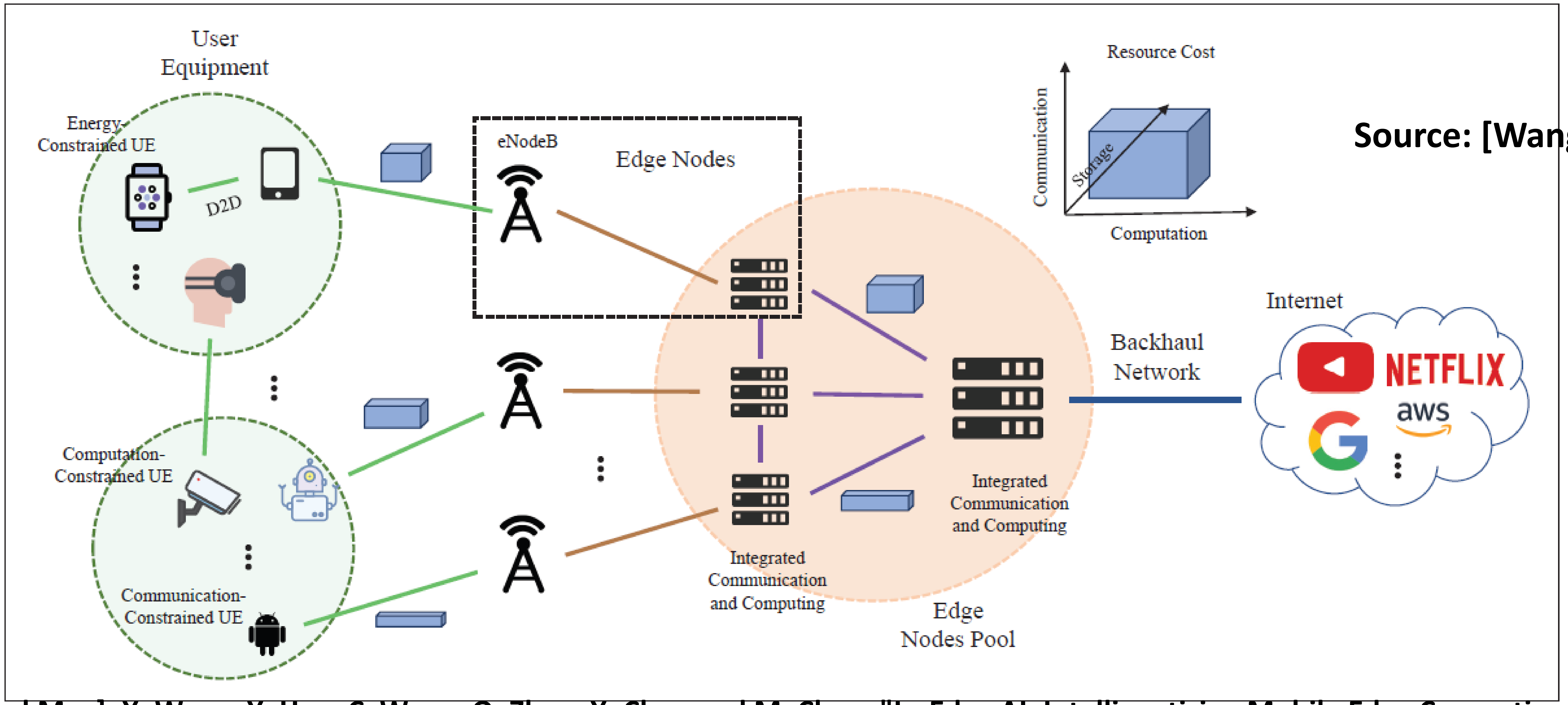
Source: [Chen, Proceeding]

Edge-AI

- Performance metrics
  - Latency, energy consumption, offloading probability, etc.
- Challenges:
  - 3C resource sharing design
    - Caching, computing, communications
  - Resource-Friendly Edge AI Model Design
    - Both design and selection are needed
  - Security and privacy
    - Model and data integrity
    - Secure communications
    - Personal information, e.g., location and activity records
  - Programming and platforms
    - Need to be flexibly used in different platforms

Edge-AI

- AI for communications
  - Previous discussion focused on realizing intelligent services
  - Now, we talk about intelligentized networks
  - We consider the AI-supported network
    - Learning-aided decision-making, sensing, etc.
    - Need collaborative 3C to realize both training and inference

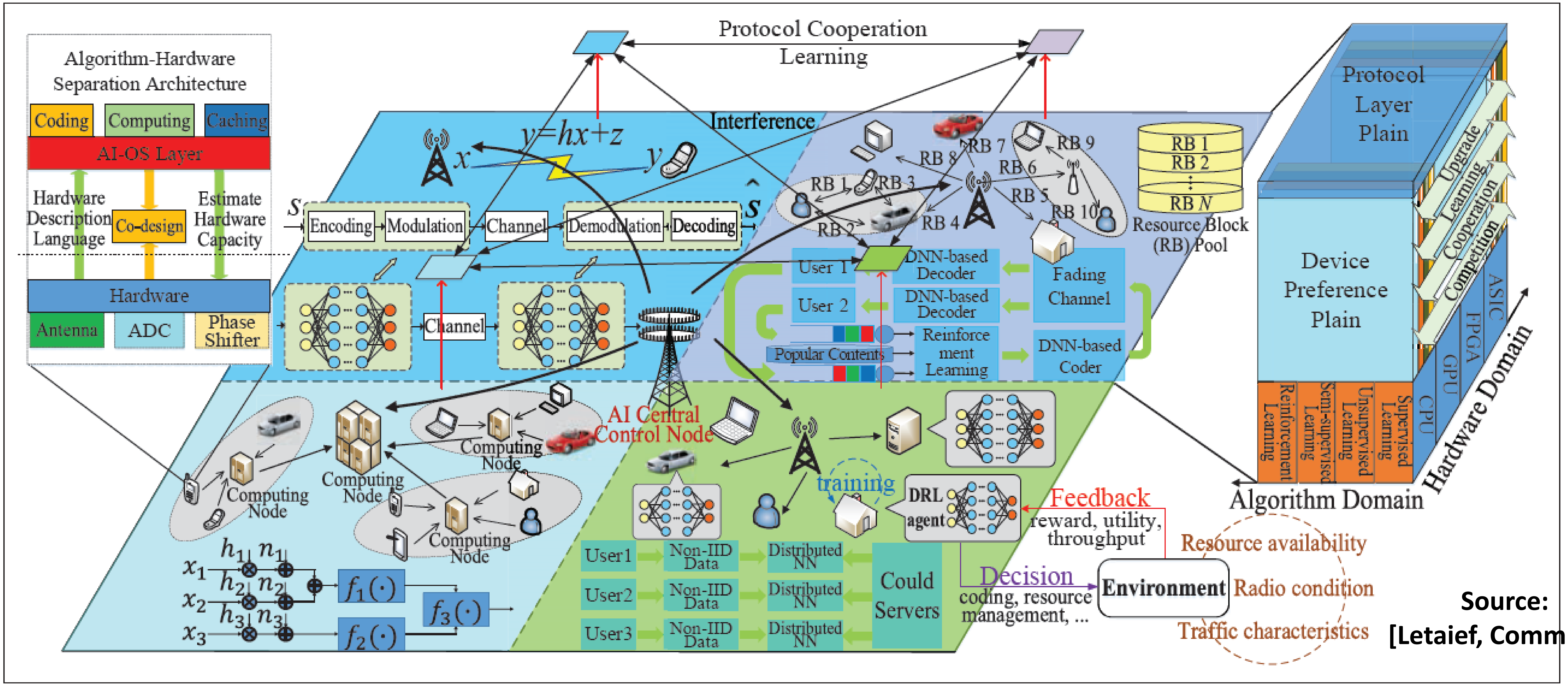


Source: [Wang, NetworkMag]

[Wang, NetworkMag]. X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen and M. Chen, "In-Edge AI: Intelligentizing Mobile Edge Computing, Caching and Communication by Federated Learning," in *IEEE Network*, vol. 33, no. 5, pp. 156-165, Sept.-Oct. 2019, doi: 10.1109/MNET.2019.1800286.

Edge-AI

- Future network architecture
  - AI to serve; AI to train; AI to infer; AI to control; AI to network
  - All involves caching, computing, and communications



Source: [Letaief, CommMag]

[Letaief, CommMag]. K. B. Letaief, W. Chen, Y. Shi, J. Zhang and Y. A. Zhang, "The Roadmap to 6G: AI Empowered Wireless Networks," in *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84-90, August 2019, doi: 10.1109/MCOM.2019.1900271.

III

Research Examples

Two Works to Demo

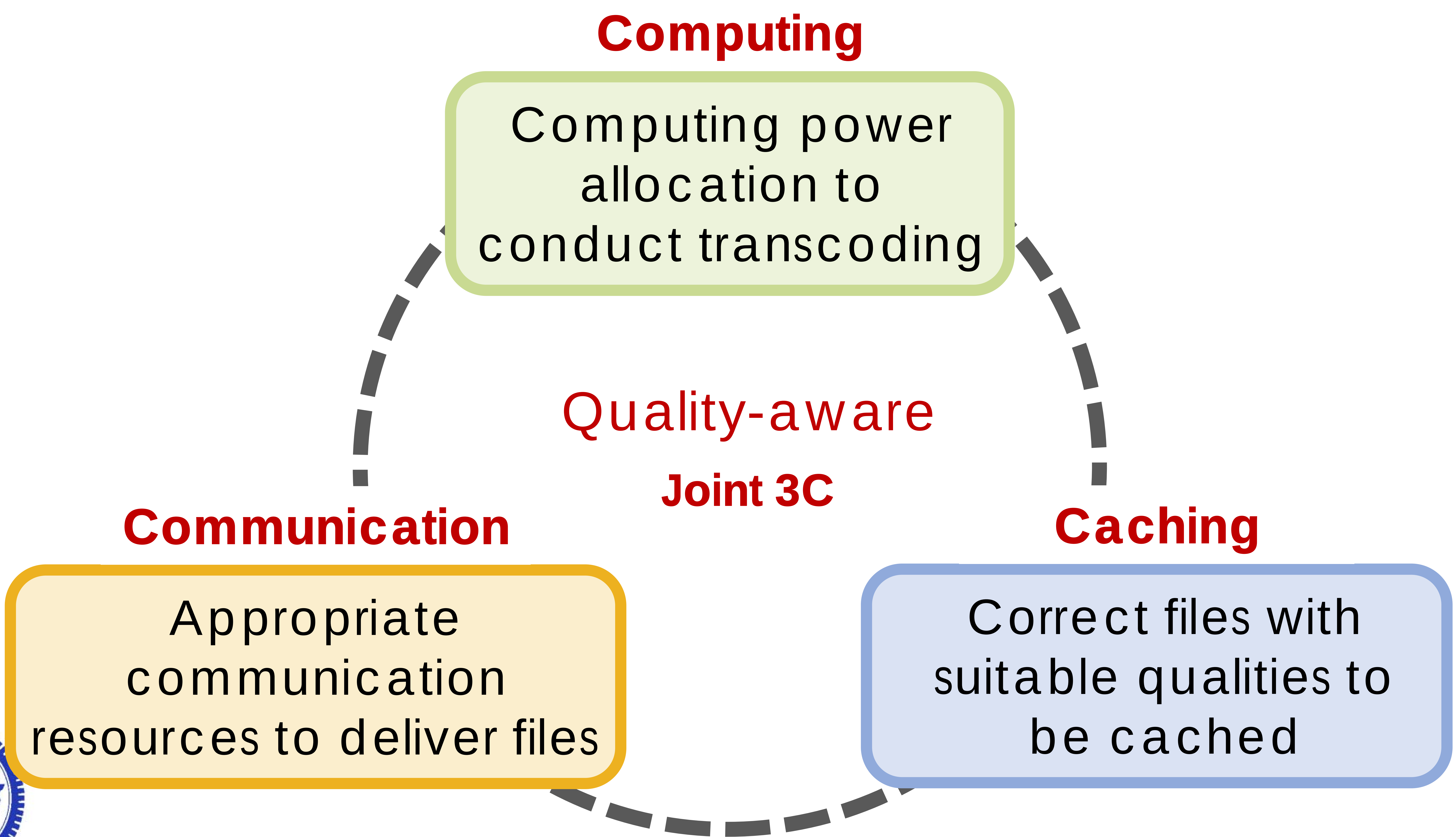
- Socially-Aware Joint Recommendation and Caching Policy Design in Wireless D2D Networks
  - Joint work with Prof. Y.-W. Peter Hong
- Quality-aware Caching, Computing and Communication Design for Video Delivery in Vehicular Networks
  - Joint work with Miss Ting-Yen Kuo and Prof. Ta-Sung Lee
- Knowledge Caching for Federated Learning
  - Joint work with Miss Xin-Ying Zheng and Prof. Y.-W. Peter Hong
- Optimal Delay-Outage Analysis for Noise-Limited Wireless Networks with Caching, Computing, and Communications
  - Joint work with Prof. Andreas Molisch

Quality-Aware 3C

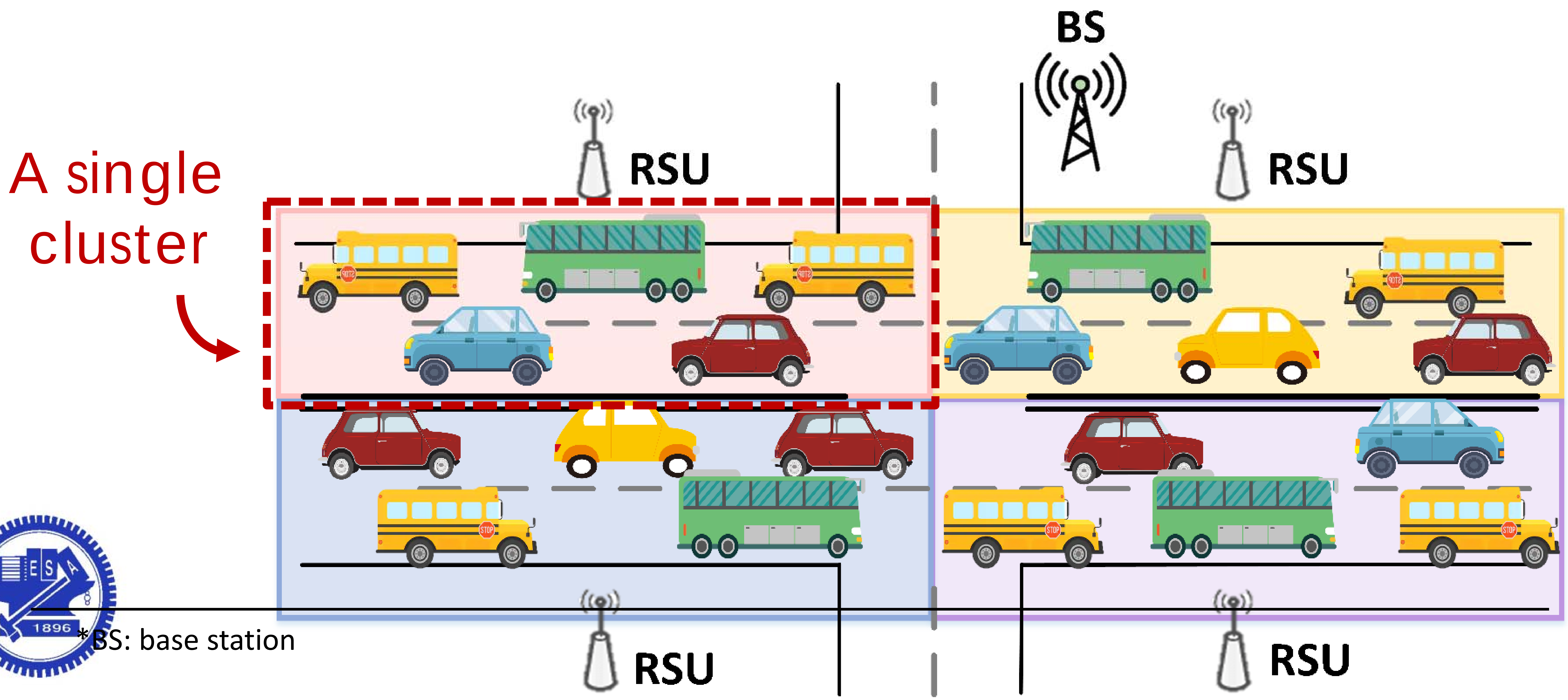
- Being one of the major sources for wireless traffic in vehicular networks, **video services** benefit from both edge-caching and edge-computing
  - Edge-caching chooses videos with suitable qualities to cache
  - Edge-computing adjusts video qualities using different transcoding and offloading schemes
  - Integrating both edge-computing and edge-caching improves video service by transcoding the videos from different files with different qualities cached in the storage

Quality-Aware 3C

- Requirements of the video delivery process
  - Jointly consider quality-aware 3C optimization



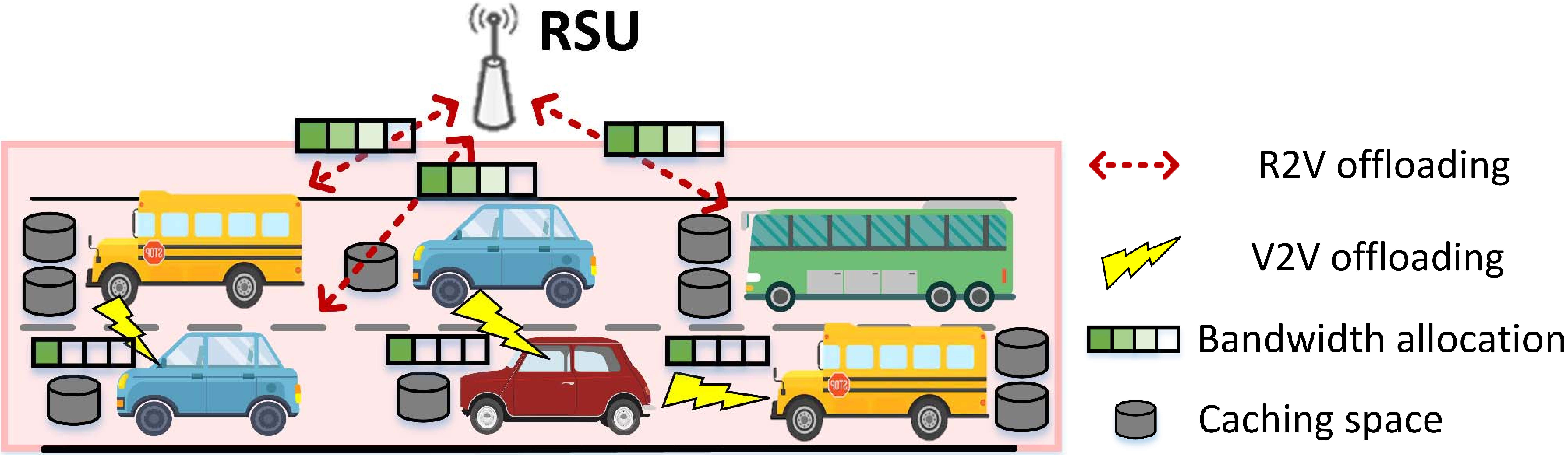
- A vehicular network including **a BS, several RSUs, buses, and cars** is considered
  - Assume that the network is split by the RSUs, and the area covered by a RSU is called a cluster
  - We assume that the **frequency reuse approach** is used, so we can focus on **a single cluster at a time-slot**



*BS: base station

Quality-Aware 3C

- Assumptions in a cluster
  - The RSU, buses, and cars all have 3C capabilities
  - Vehicles can only be served by the RSU of the cluster via **R2V communications** and by the vehicles of the cluster via **V2V communications**
- Outages can be handled via the BS as the backup



Quality-Aware 3C

- To maximize user utility, we determine
  - What quality to deliver to whom
  - Whom to associate to whom
  - Bandwidth allocation
  - Computing power allocation
  - What content to cache with what quality

Quality-Aware 3C

- Considerations
  - Full-duplex transmission with multiple association and multiple requests per vehicle
  - Goal: **maximize total utility** of the vehicular network
  - Optimize bandwidth allocation $\mathbf{B}$, computing power allocation $\mathbf{L}$, transcoding decision $\mathbf{Y}$ and caching decision $\mathbf{X}$
- Objective function

$$\max_{\mathbf{X},\mathbf{Y},\mathbf{B},\mathbf{L}} \sum_{m}\sum_{n}\sum_{c}\sum_{f}\sum_{f'} y_{m,n}^{(c,f,f')} \cdot u^{(f)} \qquad (3\text{-}1a)$$

  - $u^{(f)}$: utility size for the content

Quality-Aware 3C

- Constraints

Limited total transmit power ←

$$0 \le \sum_{n}\sum_{c}\sum_{f}\sum_{f'} y_{m,n}^{(c,f,f')} \cdot P_m \le P_{\text{ind},m}, \forall m \quad (3\text{-}1b)$$

total transmit power

Bandwidth constraints

$$\sum_{m}\sum_{n}\sum_{c}\sum_{f} b_{m,n}^{(c,f)} \le B \quad (3\text{-}1c)$$

$$0 \le b_{m,n}^{(c,f)} \le \sum_{f'} y_{m,n}^{(c,f,f')} \cdot B, \forall f, c, n, m \quad (3\text{-}1d)$$

Computing power constraints

$$\sum_{n}\sum_{c}\sum_{f} l_{m,n}^{(c,f)} \le l_{CPU}^{\max}, \forall m \quad (3\text{-}1e)$$

total computing power

$$0 \le l_{m,n}^{(c,f)} \le \sum_{f'} y_{m,n}^{(c,f,f')} \cdot l_{CPU}^{\max}, \forall f, c, n, m \quad (3\text{-}1f)$$

Limited caching capacity ←

$$\sum_{c}\sum_{f} x_m^{(c,f)} \cdot Q^{(f)} \le Z_m, \forall m \quad (3\text{-}1g)$$

Quality-Aware 3C

- Constraints

No association if there are no requests ←

$$\sum_{f}\sum_{f'} y_{m,n}^{(c,f,f')} \le r_n^c, \forall c,n,m \quad (3\text{-}1h)$$

No transcoding if the content is not cached ←

$$y_{m,n}^{(c,f,f')} \le x_m^{(c,f')}, \forall f, f', c, n, m \quad (3\text{-}1i)$$

Communication and computing delay constraints

$$d_{\mathrm{t},m,n}^{(c,f)} \le \tau, \forall f, c, n, m \quad (3\text{-}1j)$$

$$d_{\mathrm{c},m,n}^{(c,f)} \le \tau, \forall f, c, n, m \quad (3\text{-}1k)$$

Caching and transcoding indicators

$$x_m^{(c,f)} \in \{0,1\}, \forall f, c, m \quad (3\text{-}1l)$$

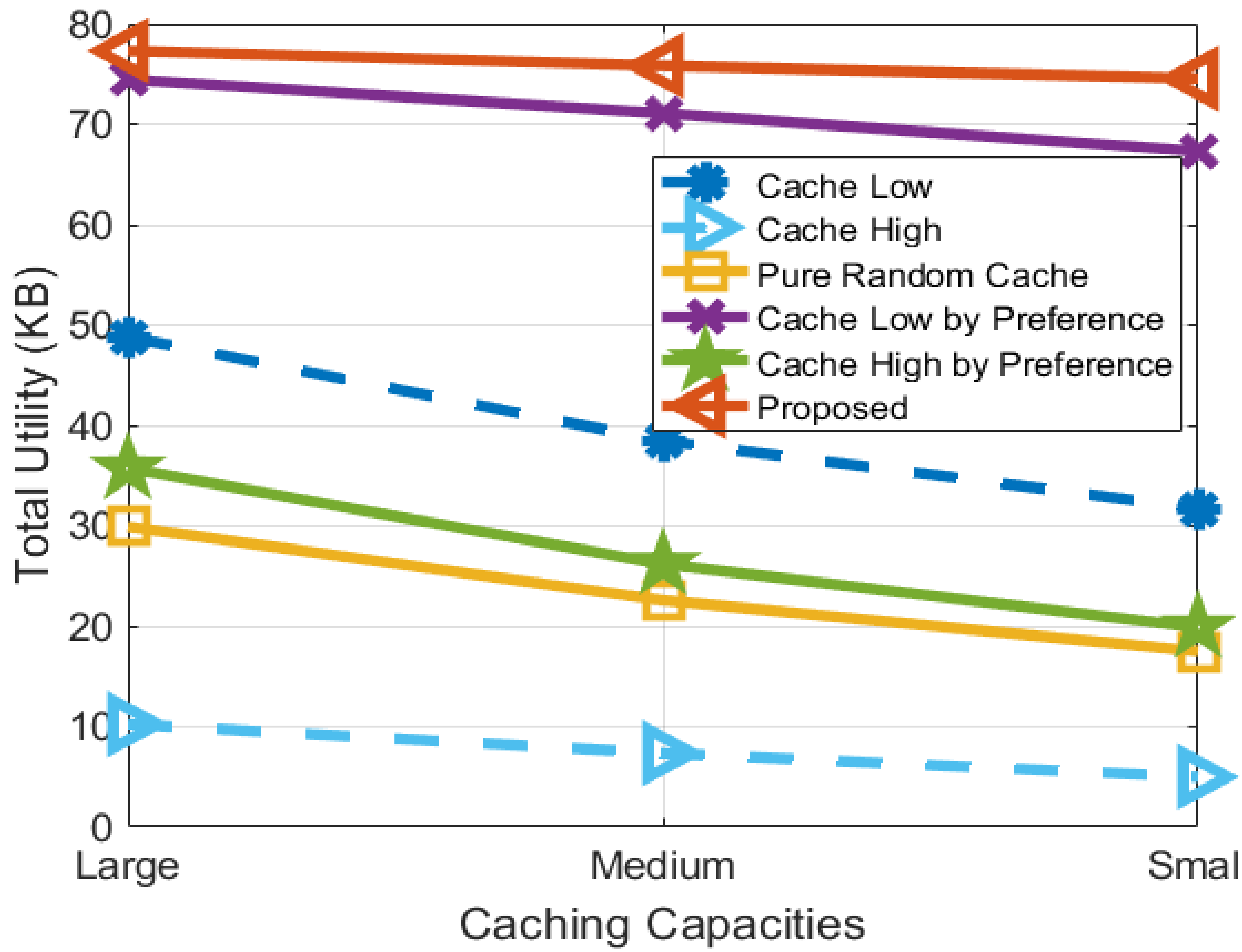
$$y_{m,n}^{(c,f,f')} \in \{0,1\}, \forall f, f', c, \mathrm{n}, m \quad (3\text{-}1m)$$

Quality-Aware 3C

- Investigations under two conditions:
  - Cached contents are pre-determined – happen when what to cache is not centrally controlled
    - Optimize computing and communication resource allocation with cache-awareness
  - Cached contents are jointly determined – happen when what to cache is centrally controlled
    - Jointly optimize 3C
    - Not practical at all, though serving as upper bound and benchmark design: for example, we can design the ideal caching, and then use practical mechanism to approximate
- Solution approach uses the submodularity and knapsack interpretation

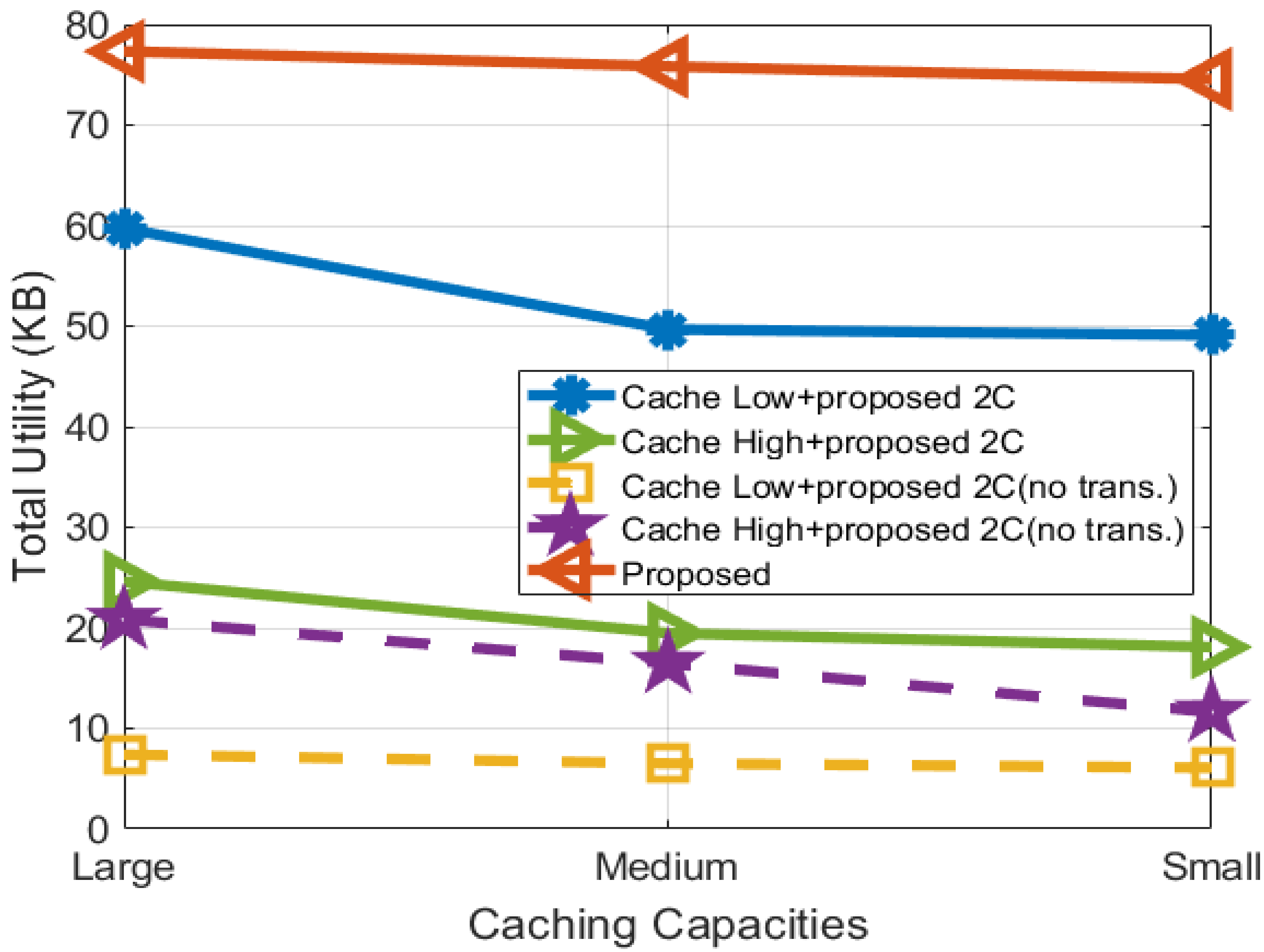
Quality-Aware 3C

- The caches can be updated dynamically
  - Comparison of different caching approaches



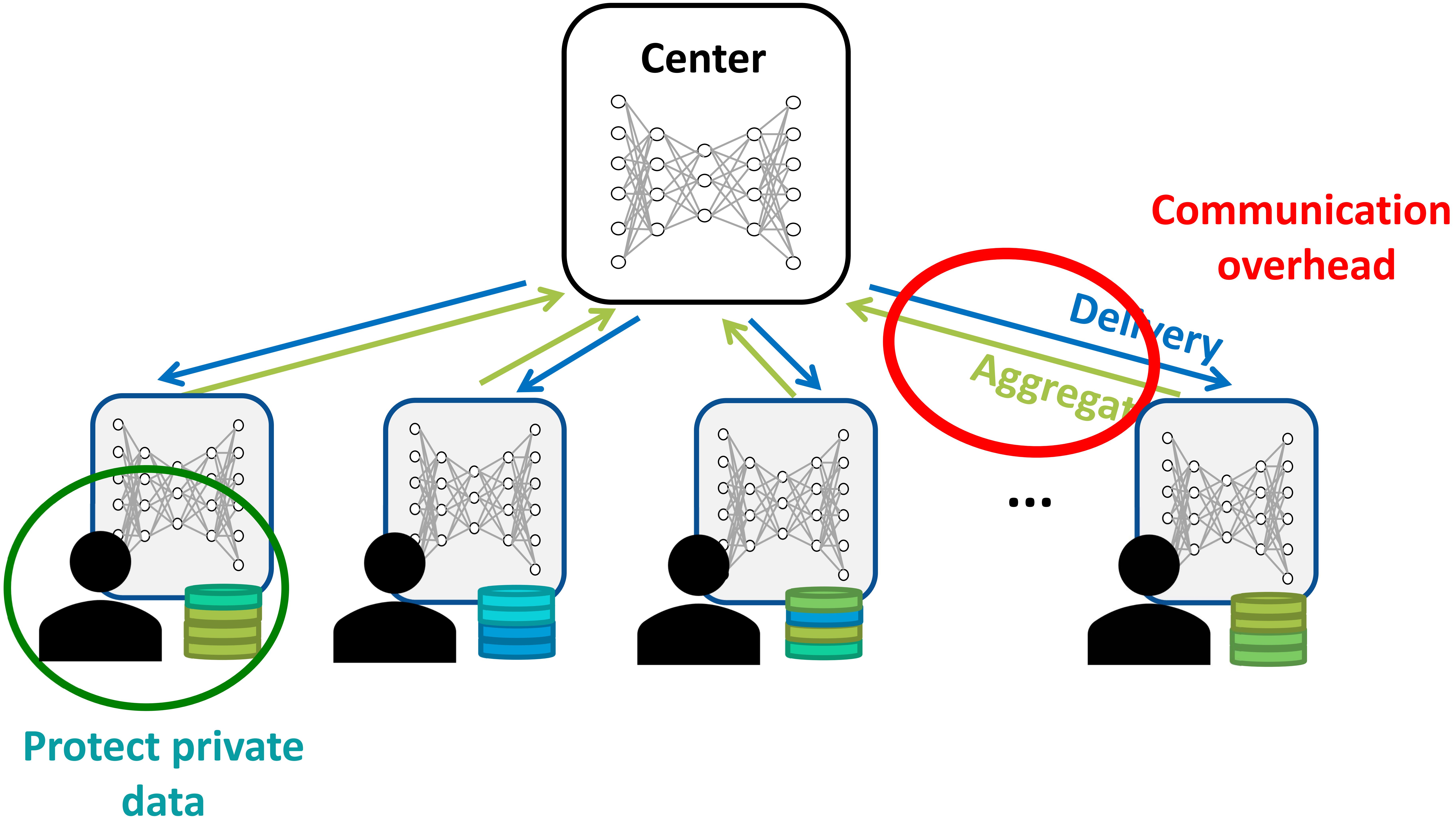
Quality-Aware 3C

- The caches are predetermined at the beginning
  - Let the proposed 3C design to be the upper bound



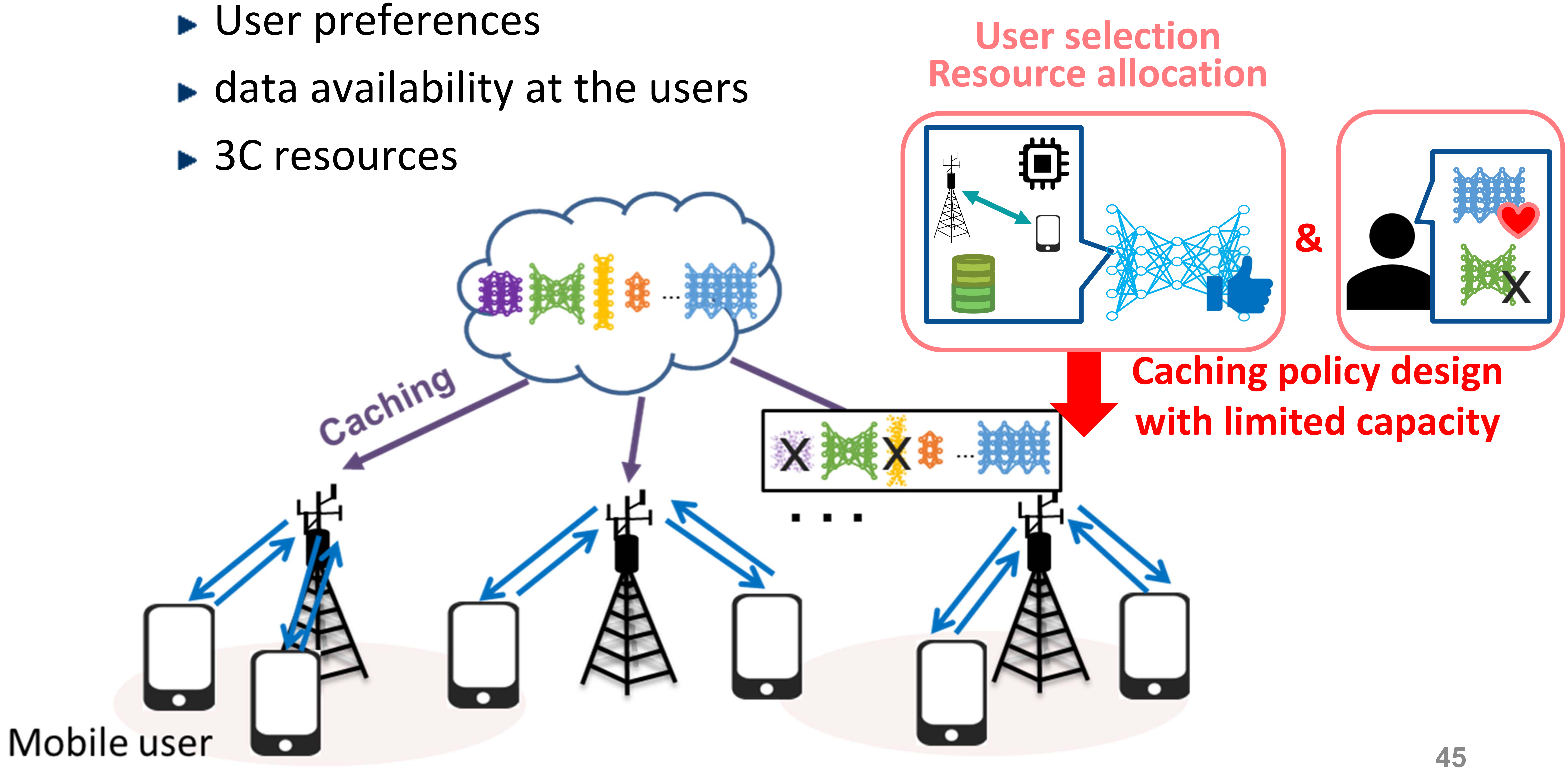
Federated Learning using 3C

- Federated learning is to distributedly train the machine under the privacy consideration
  - No need to convey personalized data to the center

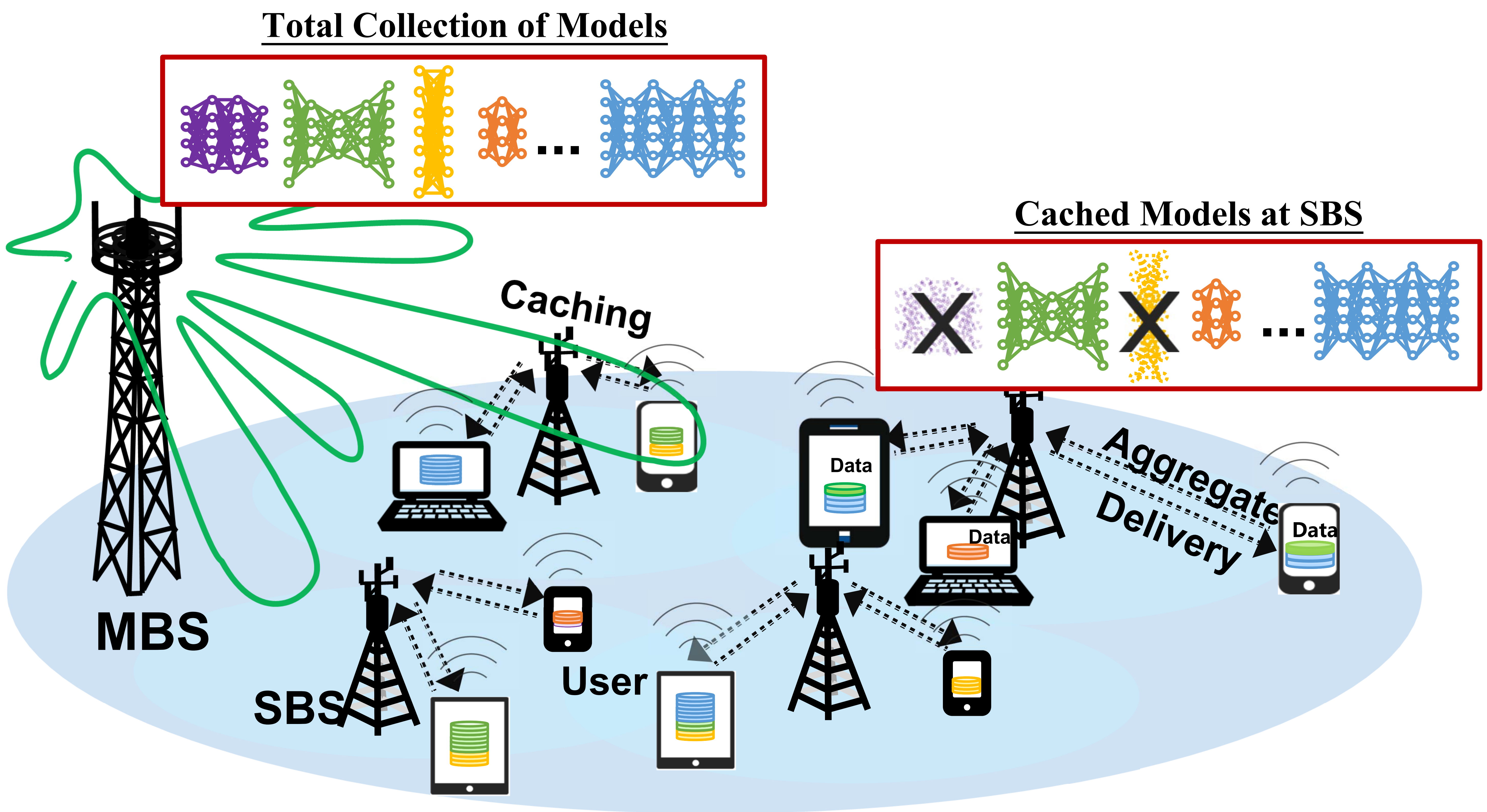


Federated Learning using 3C

- Jointly considering the training and delivery of multiple models
  - When cache storage is limited, not all models can be cached and trained. Influencing factors are:
    - User preferences
    - data availability at the users
    - 3C resources



Federated Learning using 3C



- SBS set $\mathcal{K} = \{1, \ldots, K\}$; user set $\mathcal{U} = \{1, \ldots, U\}$; model set $\mathcal{M} = \{1, \ldots, M\}$ with size $o_1, \ldots, o_M$.
- The caching policy be defined as $\{c_m^{(k)}\}_{m=1}^{M}$ at each SBS k with the constraint $\sum_{m=1}^{M} c_m^{(k)} o_m \leq O_{\max}^{(k)}$

Federated Learning using 3C

- Problem formulation

$$\min_{\substack{c_m^{(k)}, a_u^{(k)}, \\ s_{m,u}^{(k)}, P_{m,u}^{(k)}, r_{m,u}^{(k)}, \\ \forall k,m,u}} \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{u=1}^{U} a_u^{(k)} \delta_{m,u} \left[c_m^{(k)} \frac{1}{D_m} \sum_{u'=1}^{U} D_{m,u'}(1-s_{m,u'}^{(k)}) + (1-c_m^{(k)}) \frac{1}{D_m}(D_m - D_{m,u}) \right]$$

$$+ \sum_{u=1}^{U} \left(1 - \sum_{k=1}^{K} a_u^{(k)}\right) \sum_{m=1}^{M} \delta_{m,u} \frac{1}{D_m}(D_m - D_{m,u})$$

$$\text{subject to} \quad c_m^{(k)},\ a_u^{(k)}, s_{m,u}^{(k)} \in \{0,1\}, \quad r_{m,u}^{(k)} \in [0,1], \quad 0 \le P_{m,u}^{(k)} \le P_{\max},$$

$$\max_{u: s_{m,u}^{(k)}=1} \left(\frac{c_m^{(k)} D_{m,u} \omega_{m,u}}{f_u} + \frac{c_m^{(k)} o_m}{r_{m,u}^{(k)} W^{\mathrm{UL}} \log_2 (1 + \frac{P_{m,u}^{(k)} g_u^{(k)}}{r_{m,u}^{(k)} W^{\mathrm{UL}} N_0})} \right) + \max_{u: s_{m,u}^{(k)}=1} \frac{c_m^{(k)} o_m}{W^{\mathrm{DL}} \log_2 (1 + \frac{P_B g_u^{(k)}}{W^{\mathrm{DL}} N_0})} \le \gamma_T,$$

$$c_m^{(k)} s_{m,u}^{(k)} v \omega_{m,u} f_u^2 D_{m,u} + \frac{c_m^{(k)} s_{m,u}^{(k)} P_{m,u}^{(k)} o_m}{r_{m,u}^{(k)} W^{\mathrm{UL}} \log_2 \left(1 + \frac{P_{m,u}^{(k)} g_u^{(k)}}{r_{m,u}^{(k)} W^{\mathrm{UL}} N_o}\right)} \le \gamma_{\mathrm{E}}, \quad \sum_{u=1}^{U} s_{m,u}^{(k)} r_{m,u}^{(k)} \le 1, \ \forall k, m,$$

$$\sum_{m=1}^{M} c_m^{(k)} o_m \le O_{\max}, \quad \sum_{k=1}^{K} a_u^{(k)} \le 1, \sum_{u=1}^{U} a_u^{(k)} \le U_{\mathrm{load}}^{(k)}, \quad s_{m,u}^{(k)} \le a_u^{(k)}, \ \forall k, m, u.$$

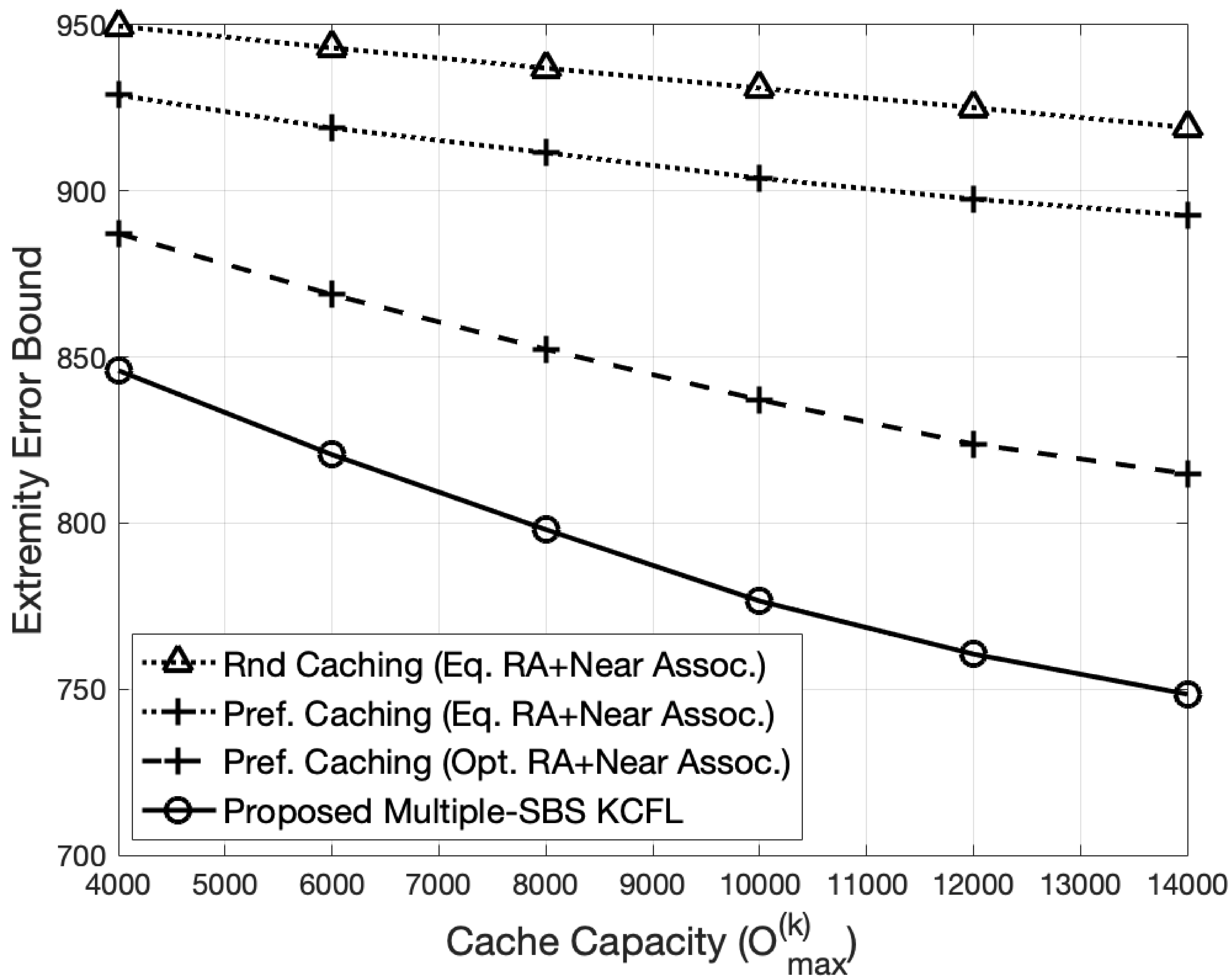
Federated Learning using 3C

- 3C resource optimization for minimizing the expected proportional data loss
  - Alternative measurement of the error bound
  - Energy and latency constraints
  - User association and selection
  - Bandwidth and power allocation
  - Caching policy

- Solution approach uses the block coordinated decent + dual ascent

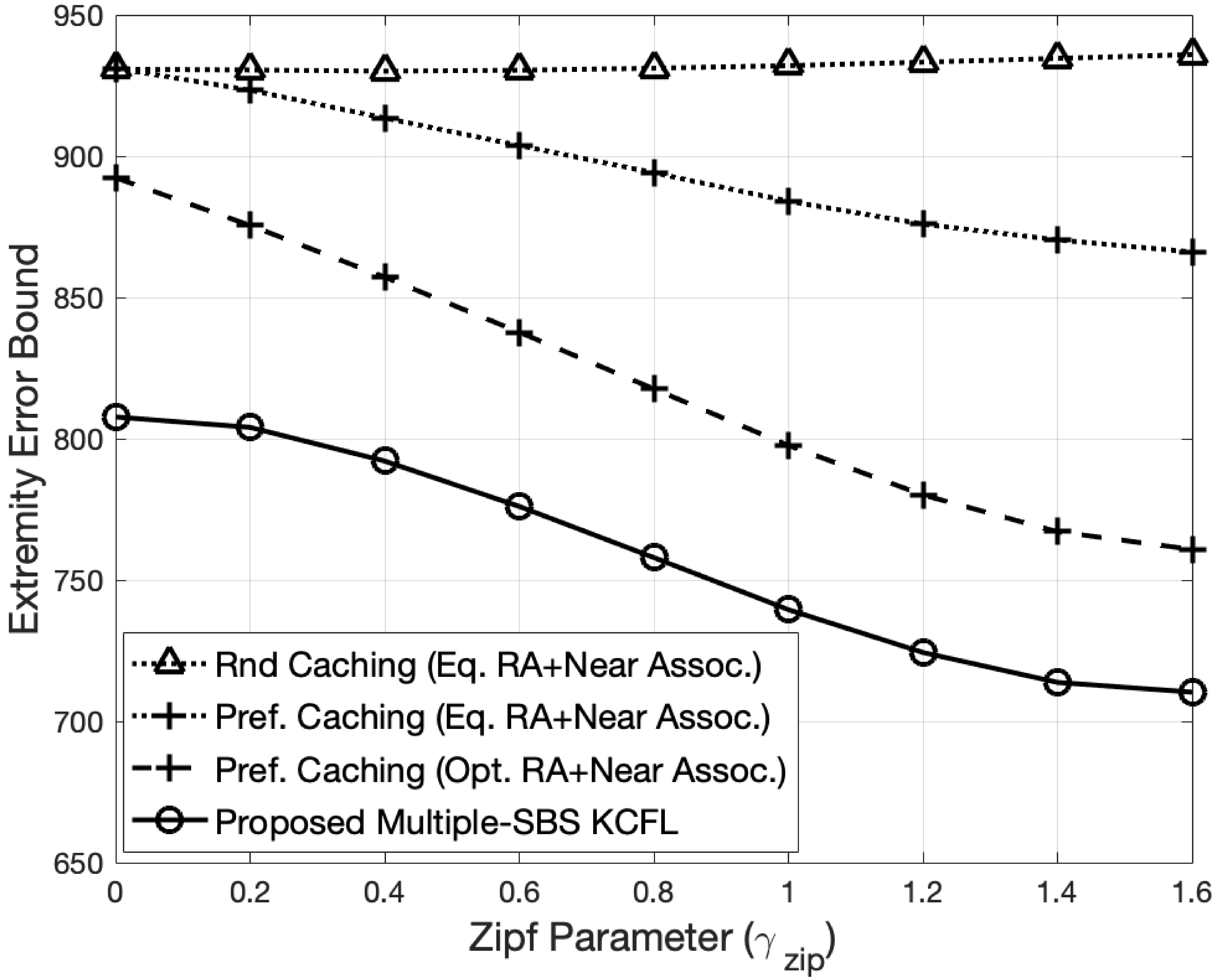
Federated Learning using 3C

- Different Cache Capacity (*K*=2, *U*=40)



Federated Learning using 3C

- Different Zipf parameters (K=2, U=40)



II

Final Remarks

Takeaways

- Edge-caching and edge-computing can significantly improve the network performance
- Emerging applications commonly require the well-collaborated caching, computing, and communication
- Caching, computing, communication joint optimization is critical for realizing edge-AI
- Some examples show that the caching, computing, communication joint optimization can bring benefits

Contact information

Ming-Chun Lee
Assistant Professor
Institute of Communications Engineering
National Chiao Tung University

Email: mingchunlee@nycu.edu.tw
Location: NCTU, ED 833