Edge Intelligence: the Confluence of Edge Computing and Artificial Intelligence

Hailiang Zhao
hliangzhao@zju.edu.cn

College of Computer Science and Technology, Zhejiang University

November 17, 2019

Outline

Outline

Outline

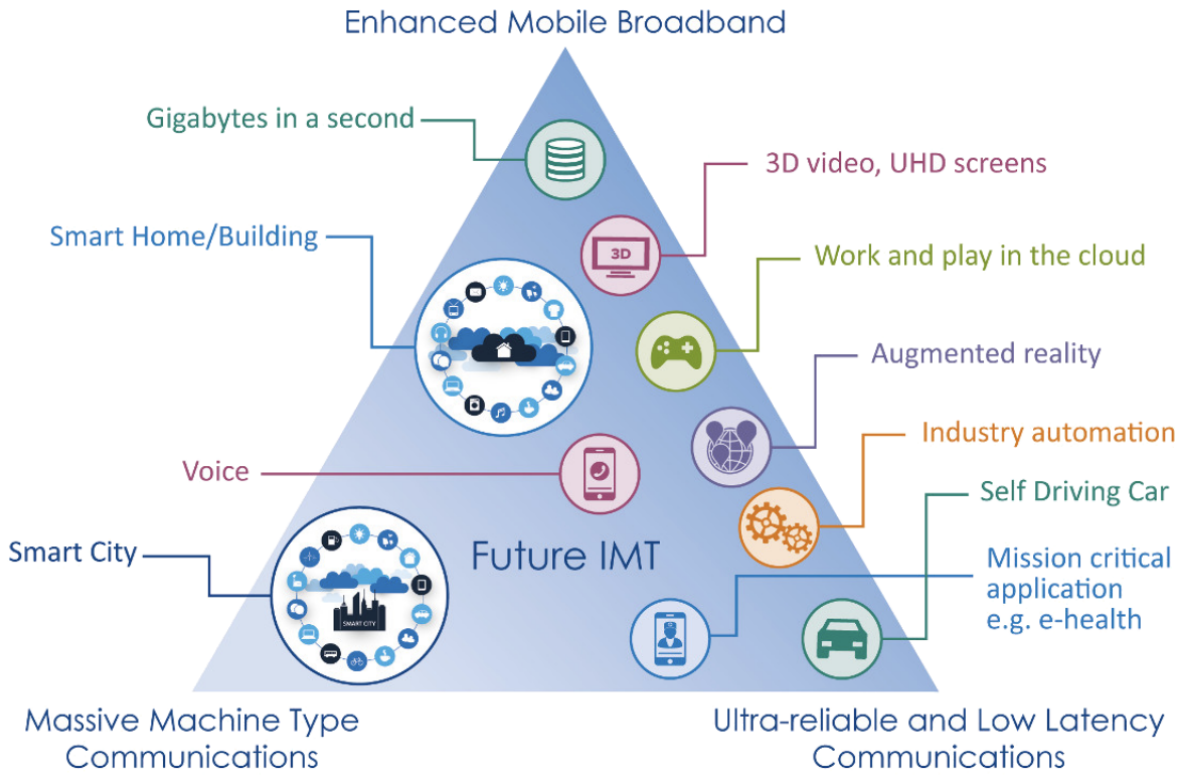
Outline

Outline

5G is coming!

What 5G brings to us

1. enhanced Mobile BroadBand (**eMBB**)
2. Ultra-Reliable Low Latency Communications (**URLLC**)
3. massive Machine Type Communications (**mMTC**)



Processing data nearby[1]

Why **edge**?

1. **explosion** of data generated by mobile and IoT devices
2. **oppressive** network congestion in backbone
3. ...

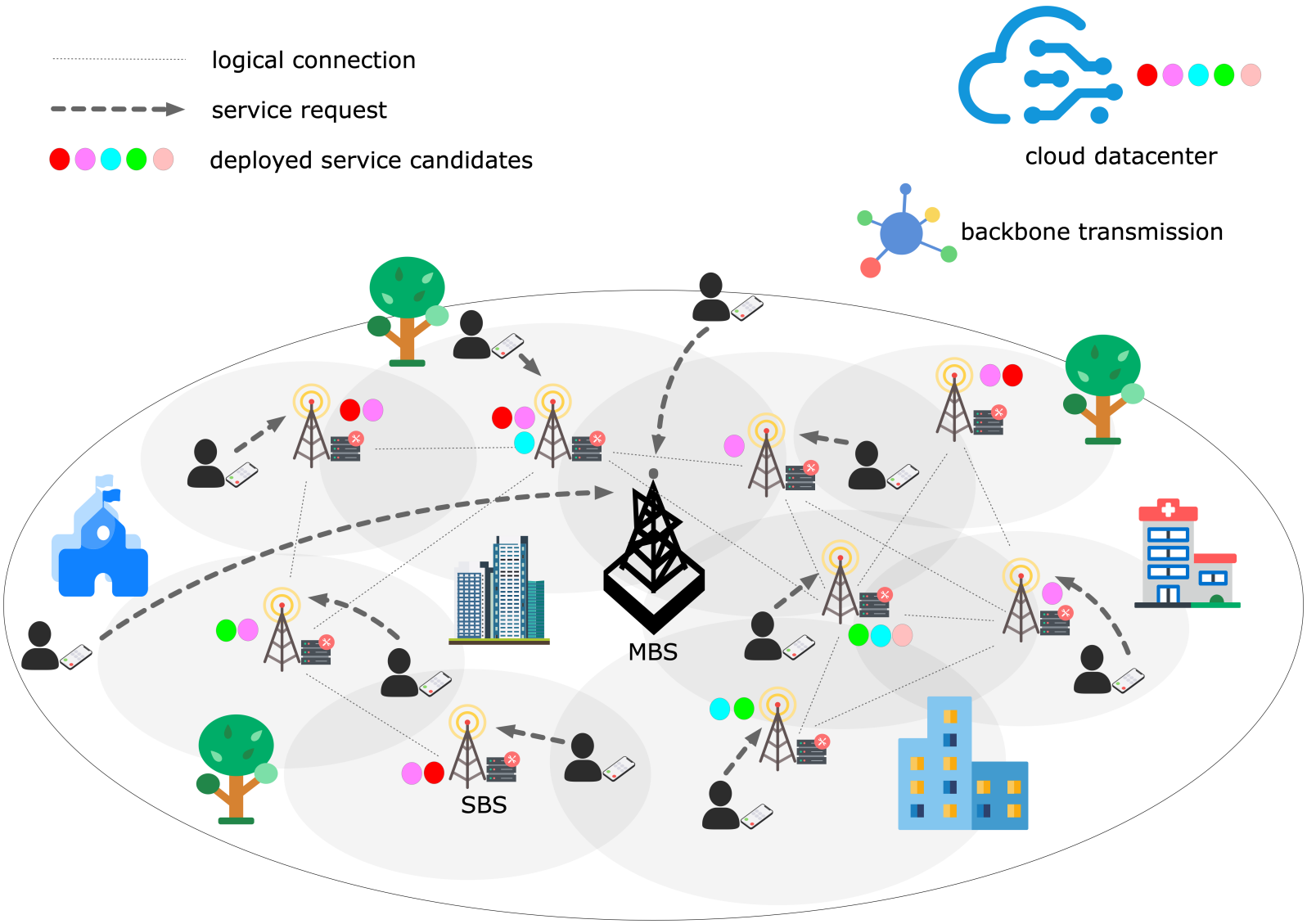
Multi-access Edge Computing (MEC)

1. communication/computation/caching/control at the edge directly
2. provide services
3. perform computations
4. manage resources

MEC avoids unnecessary communication latency and enabling faster responses for end users.

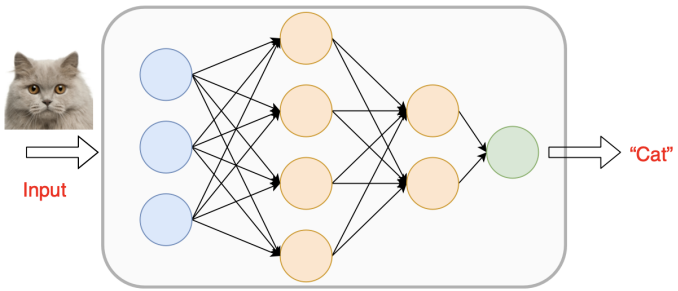
[1] Z. Zhou et al. "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing". In: *Proceedings of the IEEE* 107.8 (2019), pp. 1738–1762.

A typical pre-5G HetNet

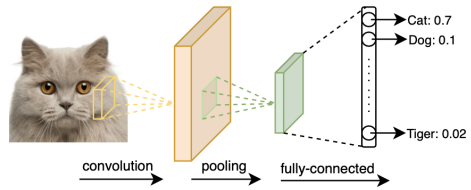


What about Artificial Intelligence?

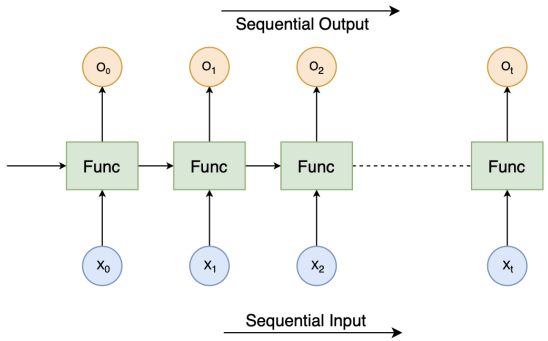
1. powerfull in **big data processing & insights extracting**
2. DNNs: powerfull **knowledge representation**
3. Typical structures of DNNs
   1. Multilayer Perceptrons (MLP)
   2. Convolutional Neural Network (CNN) (AlexNet $\rightarrow$ VGG-16 $\rightarrow$ GoogleNet $\rightarrow$ ResNet)
   3. Recurrent Neural Network (RNN) (RNN $\rightarrow$ LSTM)
4. Popular DNN models
   1. Generative Adversarial Network (GAN)
   2. Deep Reinforcement Learning (DRL)



(a) Multilayer Perceptrons



(b) Convolution Neural Network



(c) Recurrent Neural Network

Can they integrate with each other?

1. AI provides Edge Computing with **methods and technologies**
   1. Complicated resource allocation problems need to solve
   2. Huge volumes of data need to analysis
   3. AI can help in **model formulation** & **optimization**
2. Edge Computing provides AI with **scenarios and platforms**
   1. More and more data is created by widespread and **geographically distributed** mobile and IoT devices
   2. Many more applicaiton **scenarios** (intelligent networked vehicles, autonomous driving, smart hone, smart city, ...)
   3. **Hardware acceleration** on resource-limited IoT devices

Their integration leads to the birth of

Edge Intelligence (a.k.a. Edge AI)

Edge Intelligence: our definition

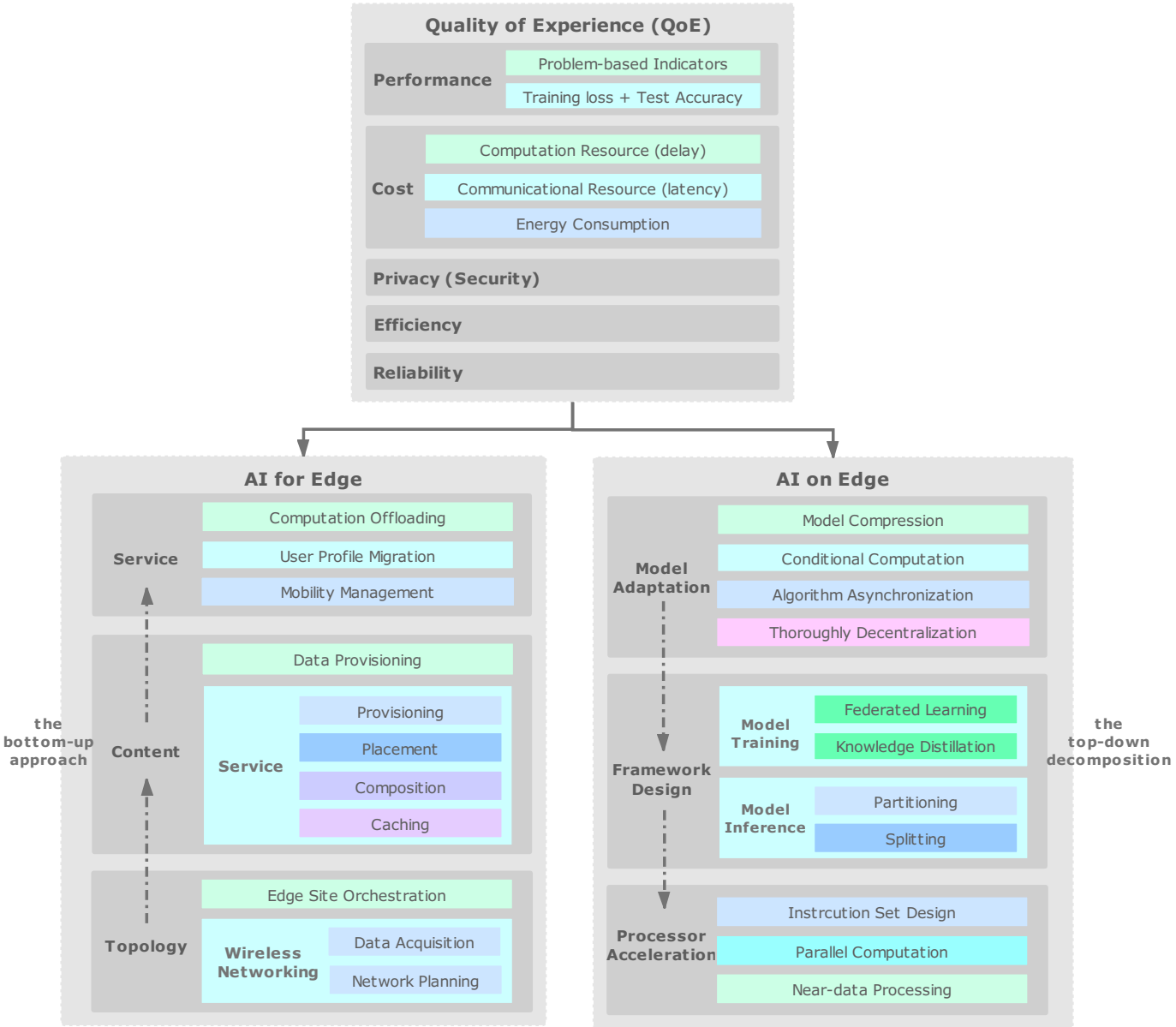
Edge Intelligence

We divide it into **AI for edge** and **AI on edge**.

1. **AI for edge**
   1. provide a better solution to the constrained optimization problems
   2. AI is used for energizing edge with more intelligence and optimality
   3. **Intelligence-enabled Edge Computing (IEC)**
2. **AI on edge**
   1. carry out the entire process of AI models on edge
   2. run model training and inference with device-edge-cloud synergy
   3. **Artificial Intelligence on Edge (AIE)**

Outline

Roadmap Overview



QoE: indicators

1. **performance**
   1. AI for edge: problem-dependent
   2. AI on edge: training loss, inference loss
2. **cost**
   1. computation cost (CPU time, CPU frequency)
   2. communication cost (transmit power, frequency band, access time)
   3. energy consumption (battery capacity)
3. **privacy (security)**
   1. leads to the birth of Federated Learning
4. **efficiency**
   1. excellent performance with low overhead
5. **reliability**
   1. robustness
   2. handle with failure

AI for edge: a recapitulation

1. **Service**
   1. optimize computation offloading via DQN[2][3]
2. **Content**
   1. service placement via MAB[4]
   2. service deployment via DRL[5]
3. **Topology**
   1. optimize UAVs via Multi-agent Learning[6]
   2. learning-driven communication[7]

---

[2] X. Chen et al. "Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning". In: *IEEE Internet of Things Journal* 6.3 (2019), pp. 4005–4018.

[3] M. Min et al. "Learning-Based Computation Offloading for IoT Devices With Energy Harvesting". In: *IEEE Transactions on Vehicular Technology* 68.2 (2019), pp. 1930–1941.

[4] L. Chen et al. "Spatio–Temporal Edge Service Placement: A Bandit Learning Approach". In: *IEEE Transactions on Wireless Communications* 17.12 (2018), pp. 8388–8401.

[5] Y. Chen et al. "Data-Intensive Application Deployment at Edge: A Deep Reinforcement Learning Approach". In: *2019 IEEE International Conference on Web Services (ICWS)*. 2019, pp. 355–359.

[6] J. Xu, Y. Zeng, and R. Zhang. "UAV-Enabled Wireless Power Transfer: Trajectory Design and Energy Optimization". In: *IEEE Transactions on Wireless Communications* 17.8 (2018), pp. 5092–5106.

[7] M. Chen et al. "Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial". In: *IEEE Communications Surveys Tutorials* (2019), pp. 1–33.

AI on edge: a recapitulation

1. **model adaptation** (too many of them)
   1. model compression, conditional computation, algorithm asynchronization, thoroughly decentralization, ...
2. **framework design**
   1. model training: Federated Learning on edge[8], knowledge distillation-based methods[9]
   2. model inference: model splitting/partitioning (Edgent)[10]
3. **processor acceleration**[11]
   1. design special instruction sets
   2. design high parallel computing paradigms
   3. move computation closer to memory

---

[8] Kai Yang et al. "Federated Learning via Over-the-Air Computation". In: *CoRR* abs/1812.11750 (2018). arXiv: 1812.11750.

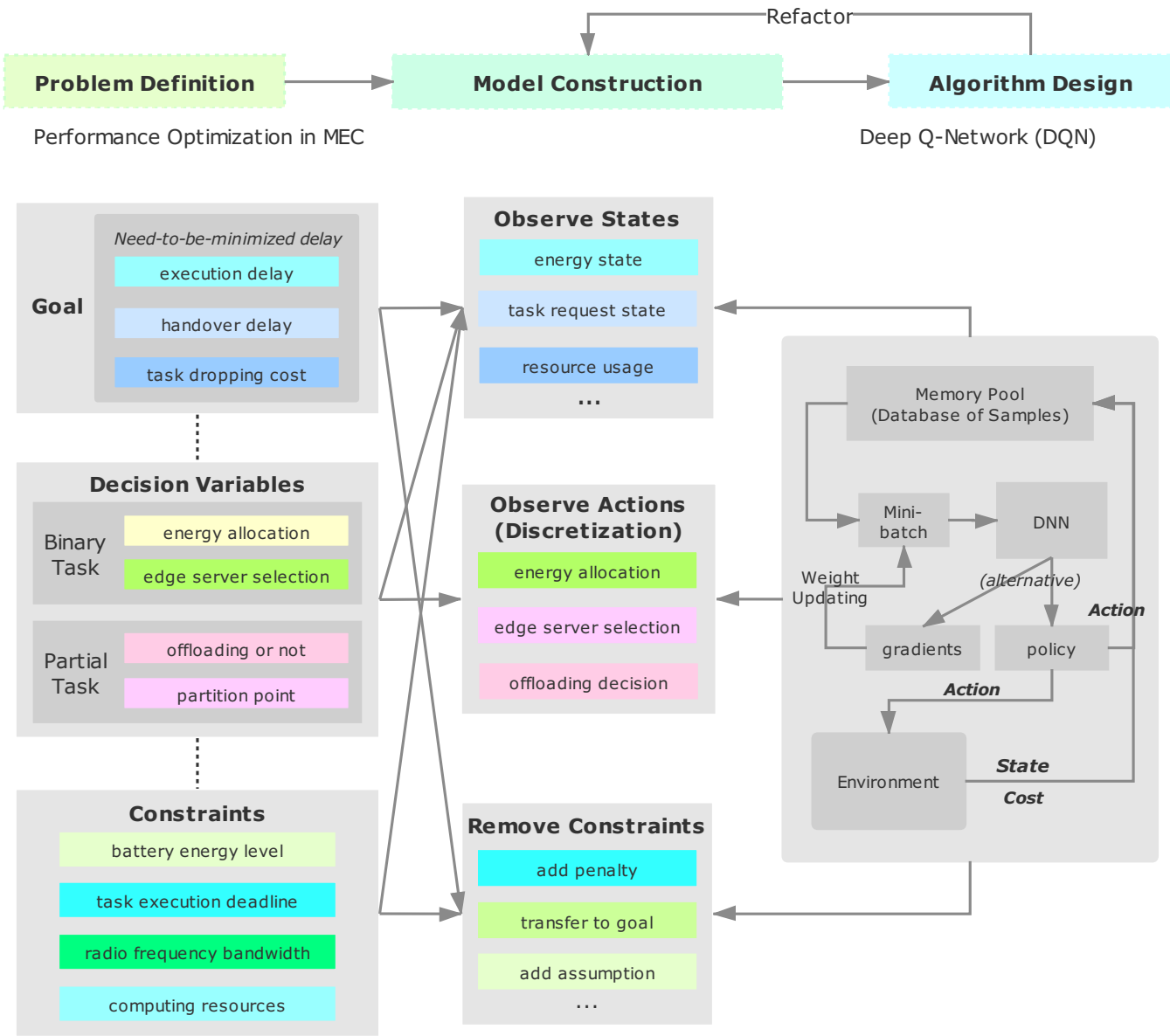
[9] Jin-Hyun Ahn, Osvaldo Simeone, and Joonhyuk Kang. "Wireless Federated Distillation for Distributed Edge Learning with Heterogeneous Data". In: *ArXiv* abs/1907.02745 (2019).

[10] En Li, Zhi Zhou, and Xu Chen. "Edge Intelligence: On-Demand Deep Learning Model Co-Inference with Device-Edge Synergy". In: *Proceedings of the 2018 Workshop on Mobile Edge Communications, MECOMM@SIGCOMM 2018, Budapest, Hungary, August 20, 2018*. 2018, pp. 31–36.

[11] V. Sze et al. "Efficient Processing of Deep Neural Networks: A Tutorial and Survey". In: *Proceedings of the IEEE* 105.12 (2017), pp. 2295–2329.

Outline

Utilize DQN for performance optimization

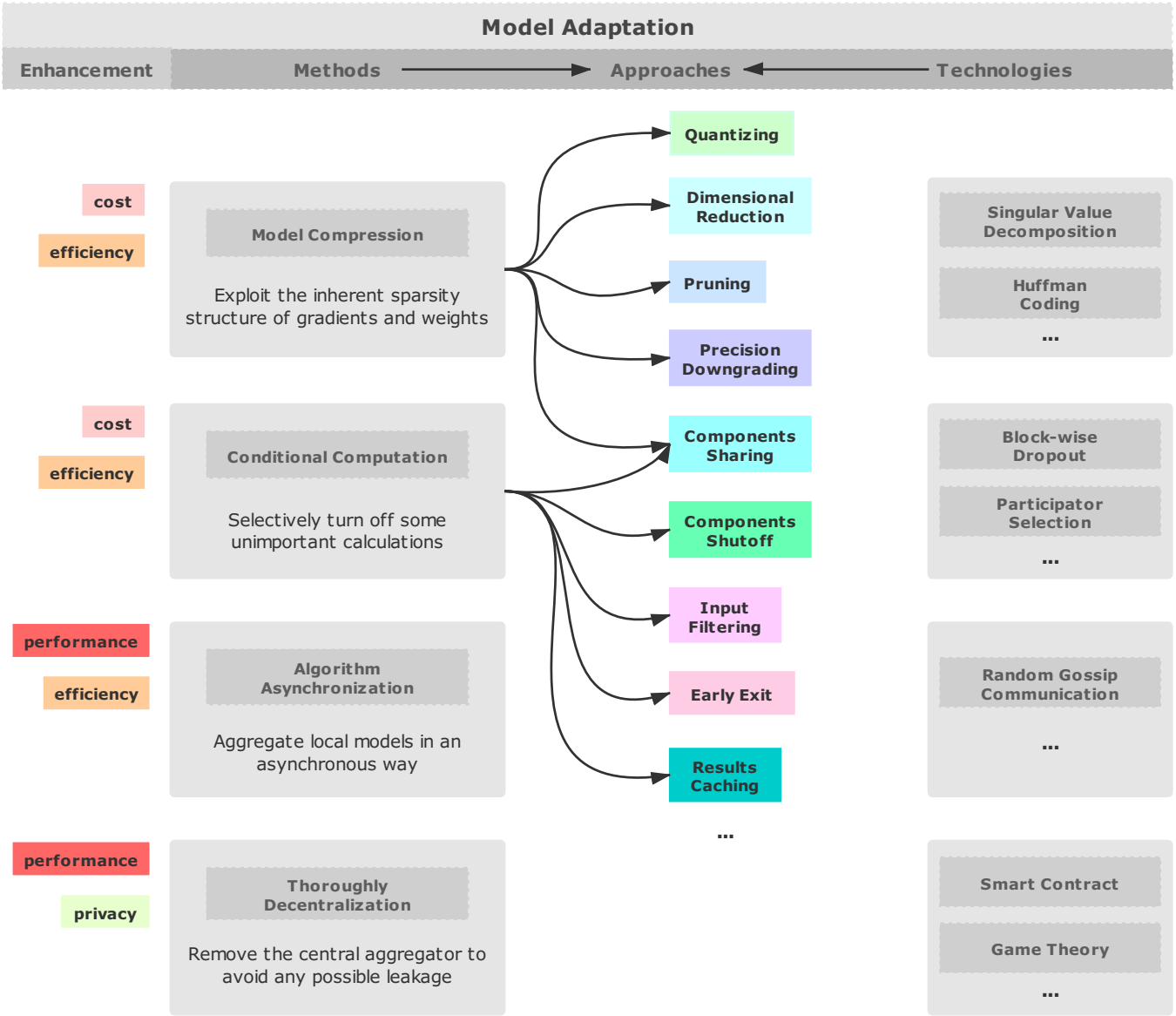


Grand challenges

1. **model establishment**
   1. unrestrained searching space
   2. state/action set cannot be infinite
2. **algorithm deployment**
   1. cannot obtain analytic (approximate) optimal solution
   2. too many iterations $\rightarrow$ hard to deploy in an online manner
   3. who undertake the responsibility?
3. **balance between optimality and efficiency**

Outline

Model Adaptation: a classification



Grand challeneges

1. **data availability**
   1. where to find usable data?
   2. incentive mechnisms
   3. obvious bias from distributed end users (non i.i.d.)
2. **model selection**
   1. select befitting threshold of learning accuracy & scale of models
   2. select probe training frameworks and accelerator architectures
3. **coordination mechanism**
   1. same method achieves different results
   2. compatibility and coordination (cloud-edge-device synergy)
   3. establish a unified API interface?