工作进度报告

使用深度强化学习算法优化边缘计算任务卸载问题

范也 MG21320002

做了什么

1. 调查20几篇DRL+任务卸载的论文
2. 编写完成基本的实验框架，包含
   1. 边缘计算任务卸载的模拟环境
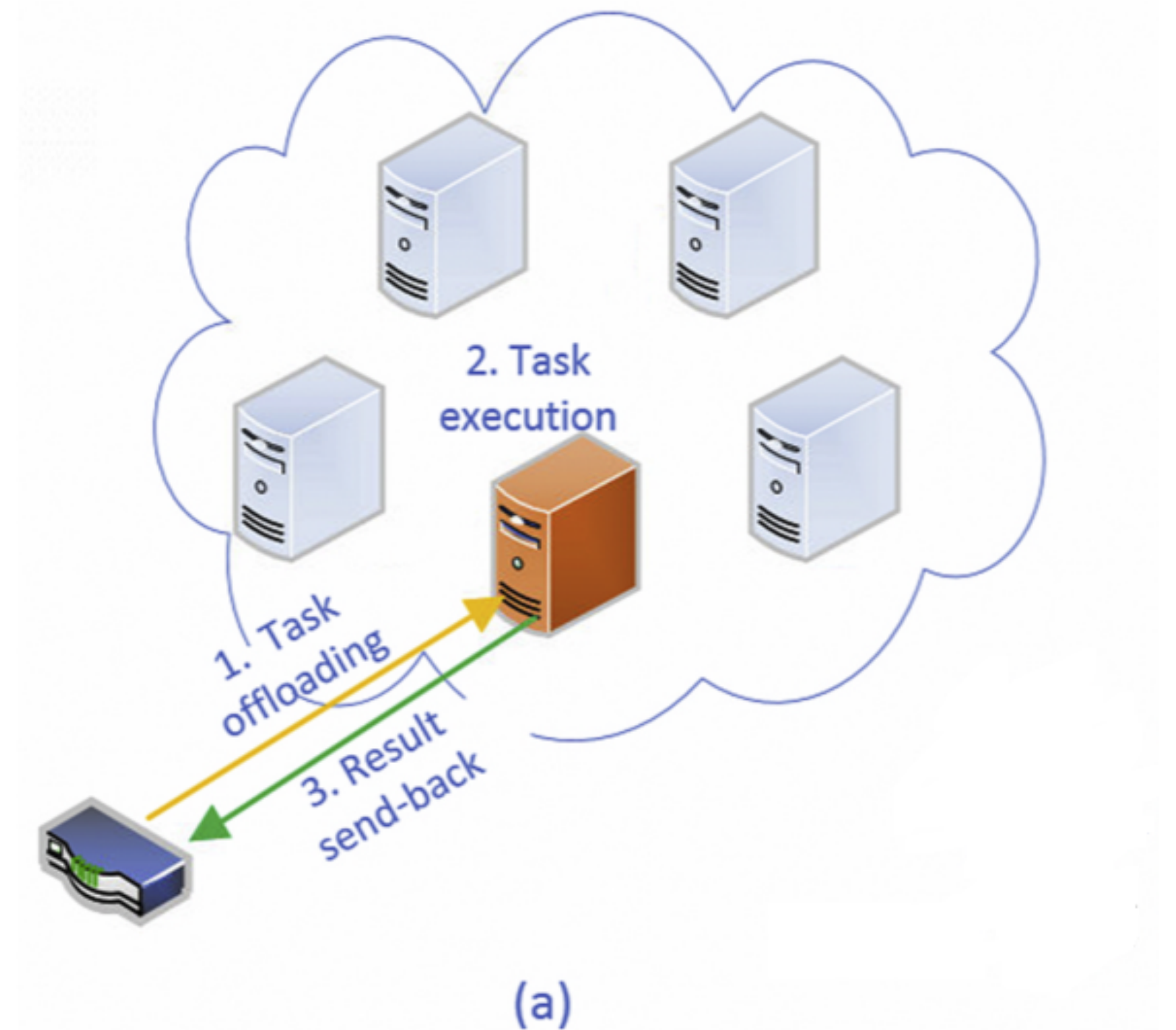   2. 使用较先进的DRL算法：T3C、SAC、PPO进行调度

工作计划

1. 对现有的论文进行较详细的排查，找到比较好的优化目标 2 week
    1. 优化现有的环境和算法
    2. 调参
2. 整合discrete action的DRL算法
3. 实现baseline：贪心、穷举、无调度、线性松弛、DQN
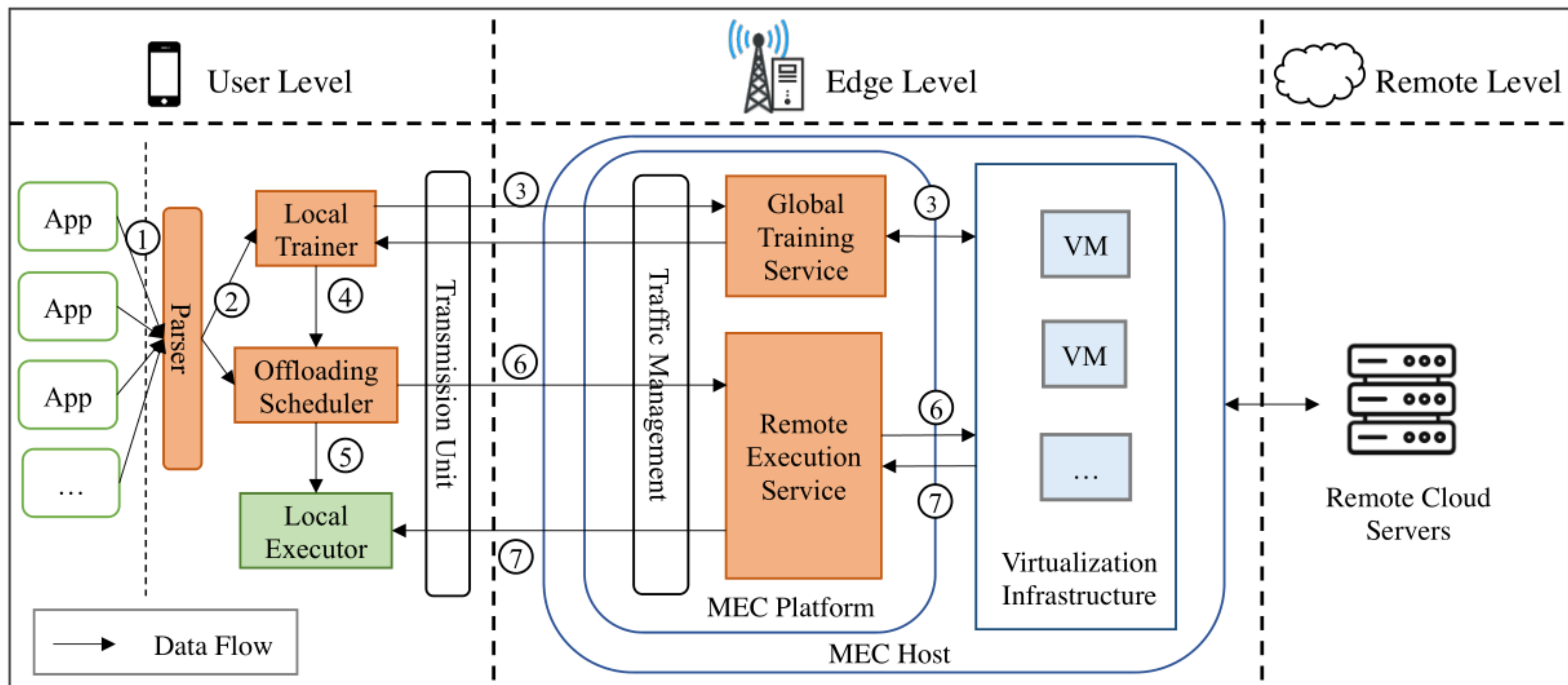
MEC offloading 简介

1. 云计算的不足

2. 工厂、AR游戏…

3. 任务调度：最优化问题

   1. 传统方式（凸优化、非凸优化）、遗传算法

4. game theory 3. 深度强化学习算法：速度快、泛化性强

5.
$$\min_{(\mathbf{a})} \quad Time + Battery,$$
$$\text{s.t.} \quad a_i \in \{0, 1\}, \forall i \in \mathcal{M},$$
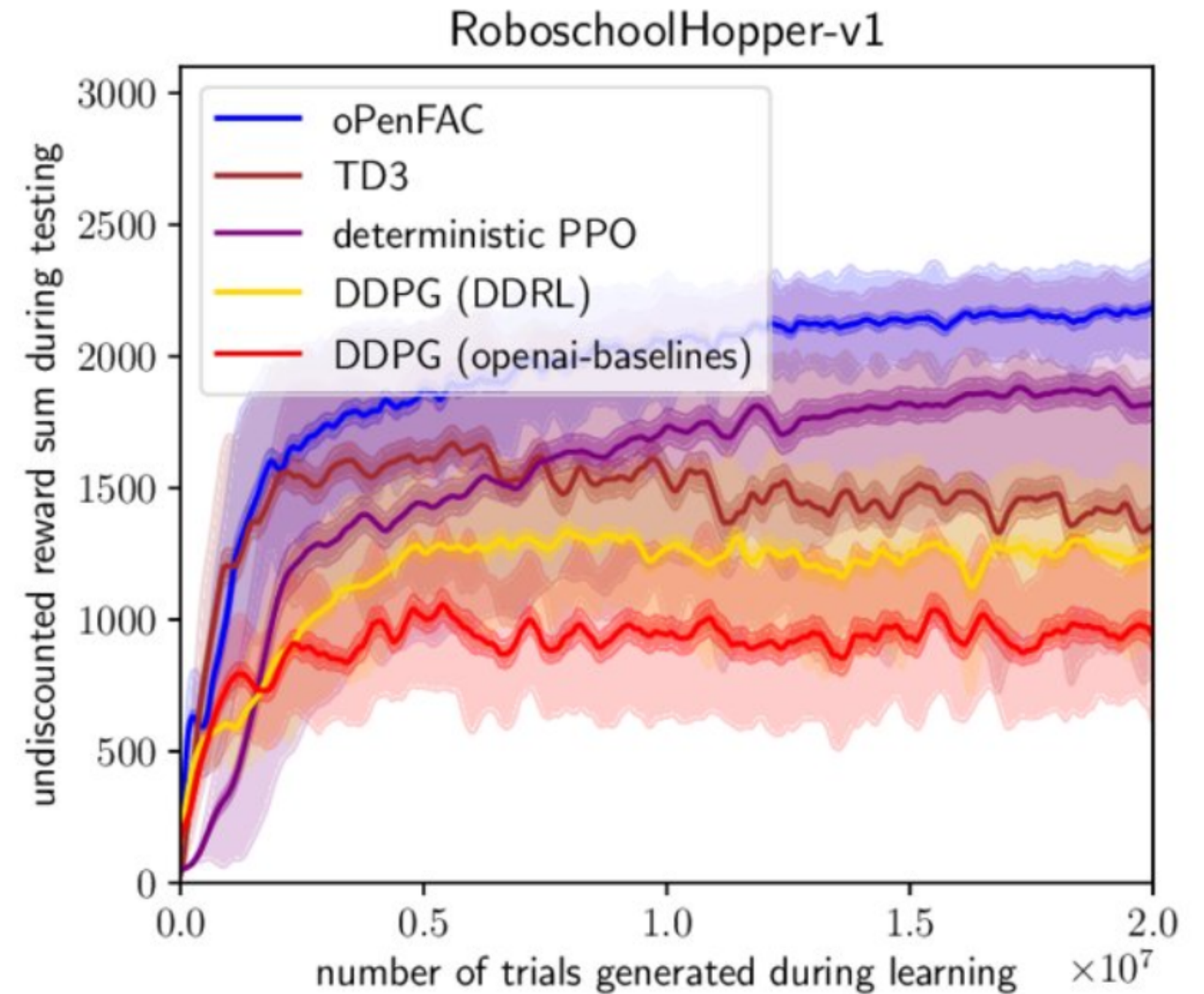

(a)

已读论文

DRL + MEC offloading + Lyapunov + ...

1. 现有论文中使用的DRL算法比较落后：
   REINFORCE、DDPG、A3C、DQN

2. 算法框架基本一样，都是简单的模拟环境+DRL算法

3. 优化目标（公平性、效率、安全性、电量）、考虑的
   环境（多agent、连续、离散、时间分配方式）各有
   不同

4. 通过算法中的微调，lyapunuv 或 np-hard证明增加工
   作量

目前先实现基本的框架，根据具体目标进行微调

RoboschoolHopper-v1

Modeling method	Centralized/ distributed	Scalability	Static/ dynamic	Achieved objectives
(Non)convex optimization	Centralized	Not scalable	Static	Global offloading optimization
MDP	Centralized/ distributed	Not scalable	Dynamic	Local offloading optimization
Game theory	Distributed	Scalable	Dynamic	Nash equilibrium
Lyapunov optimization	Centralized	Relatively scalable	Static	Offloading optimization Lyapunov drift
Machine learning	Centralized	Scalable	Dynamic	Intelligent decision

DNN

Quantization g_k

$\arg\max Q^*(\boldsymbol{h_t}, \boldsymbol{x_k})$

Channel
Gain

Input for the t-th
time frame

$\boldsymbol{h_t}$

$\hat{\boldsymbol{x}_t}$

$\boldsymbol{x}_1$

$\boldsymbol{x}_K$

Compute
$Q^*(\boldsymbol{h_t}, \boldsymbol{x_k})$
by solving
convex
problem (P2)

Selected
Action

$\boldsymbol{x_t}^*$

**Offloading Action
Generation**

Output for the t-th time
frame $(\boldsymbol{x_t}^*, \boldsymbol{a_t}^*, \boldsymbol{\tau_t}^*)$

Train

Replay Memory

Sample random batch

**Offloading Policy
Update**

Training Samples

Channel $\boldsymbol{h_t}$ Action $\boldsymbol{x_t}^*$
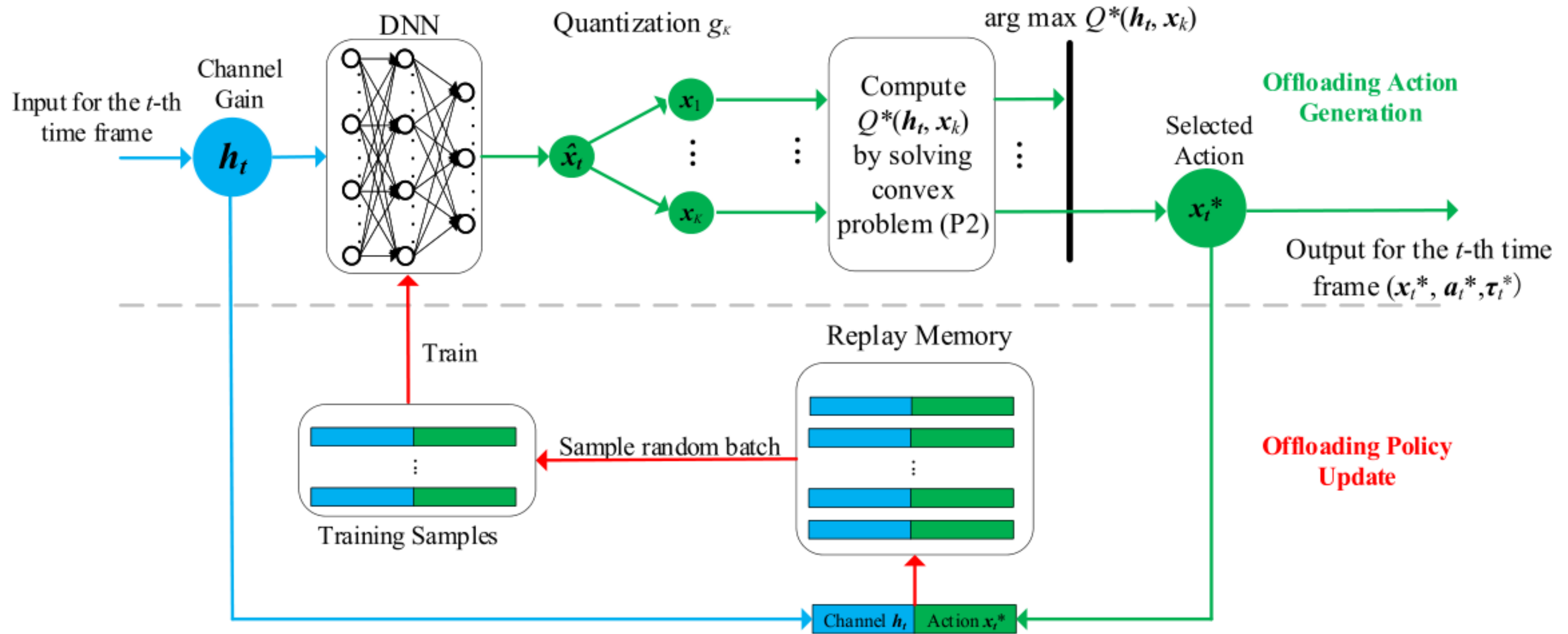
模拟实现
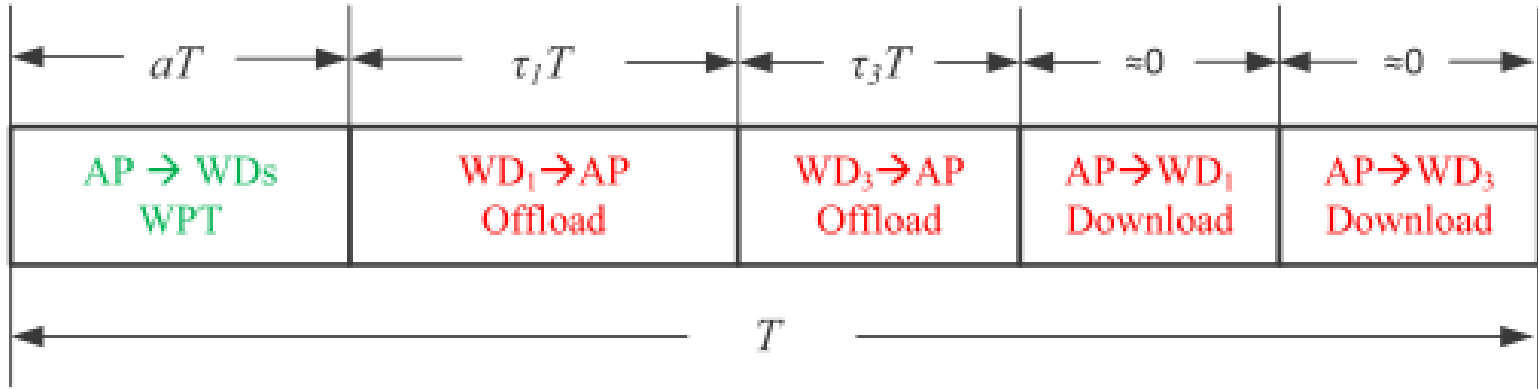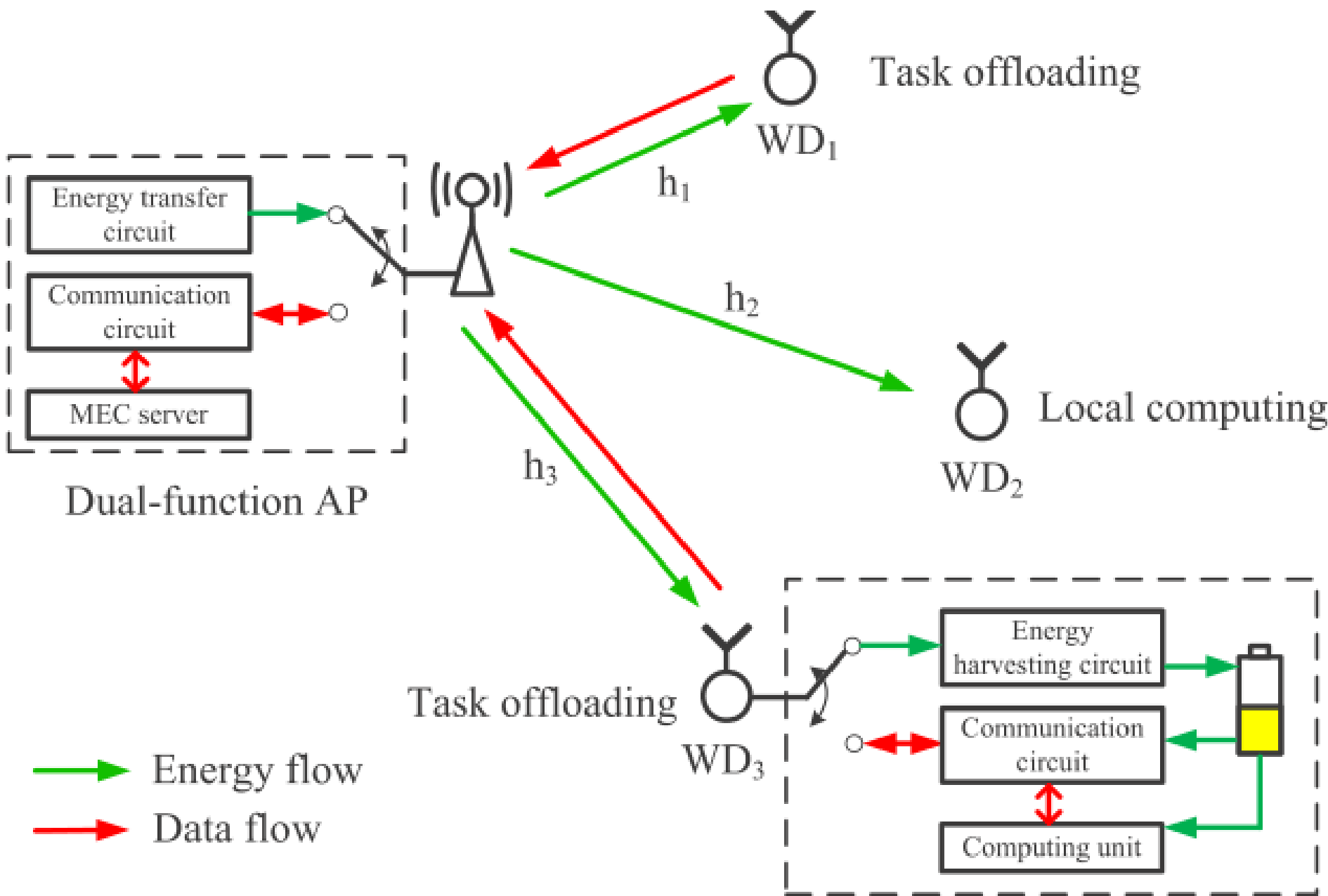
1063 line

已经包含功能特性

1. 电量消耗
2. 处理速度
3. 任务缓冲
4. 数据传输
5. 信道占用

拟添加：

1. 虚拟机配置
2. 位置移动
3. 多server

DRL算法

~1000 line

已包含大部分连续动作空间算法：

- TD3、PPO、SAC、DDPG

拟实现其他算法：

- D3QN（离散）、D4PG(连续)

VPG. Vanilla Policy Gradient. 2000
AC. Actor-Critic Methods. 2000

Stochastic Policy
NPG. Natural Policy Gradient. 2002

Trust Region
TRPO. Trust Region Policy Optimization. 2015

GAE. Generalized Advantage Estimation. 2015

Advantage Function
A2C (Advantage Actor-Critic)
A3C. Asynchronous A2C. 2016

Trust Region (approximated, surrogate)
KL Penalty Coefficient
PPO. Proximal Policy Optimization. 2017
Planning (Model-based RL)
MBPO. Model Based PO. 2019
Auxiliary Task (on-policy, off-policy)
PPG. Proximal Policy Gradient. 2020

Energy-Based Policy
SQL. Soft Q-learning. 2017
Maximum Entropy
Automating Entropy Adjustment
SAC. Soft Actor-Critic. 2018

Q-learning. 1992

Q-table → Q net.
Experience Replay
DQN. Deep Q Network. 2014

Q net. → 2 Q net.
Double DQN. 2016
+Advantage Function
Dueling DQN. 2016

Deterministic Policy
DPG. Deterministic Policy Gradient. 2015

Taming the Noise via soft update. 2015

Greedy-Policy → Policy Net.
DDPG. Deep DPG. 2016
D4PG(Distributed Distributional DDPG). 2017

2 Q net. → Twin Critic
Delay Target Update
Policy Smoothing (SPG in DPG)
TD3. Twin Delayed DDPG. 2018

Distributional Perspective
C51 DQN (Categorical 51 grids). 2017

Quantile Regression
QR-DQN. 2017

Prioritized sweeping. 1993
PER. Prioritized Experience Replay. 2016
HER. Hindsight Experience Replay. 2017

All DQN Variances
Rainbow DQN. 2017

Use Monte Carlo Tree Search 2006
AlphaGo. 2015
Without human knowledge
AlphaZero. 2017
Planning using Dynamics Model
Value Prediction Network. 2017
+Atari game (continuous state space)
MuZero. 2019 (Model-based RL)

Ape-X DQN. 2018
Ape-X DPG. 2018

IMPALA. 2018
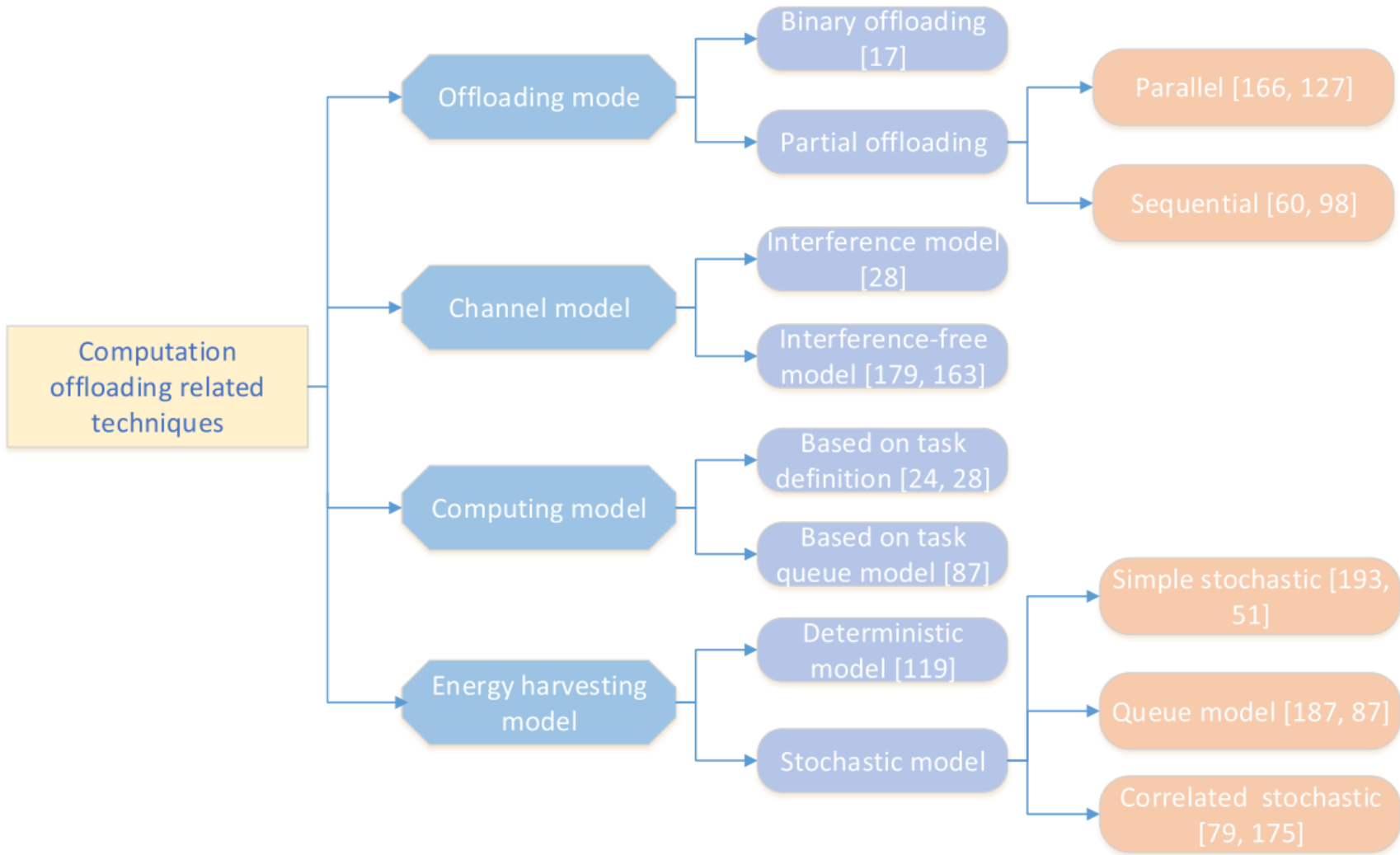**Seed RL. 2020**

后续工作计划

寻找较新颖的优化目标

将现有论文中的

- 优化目标
- 特性（距离变化、连续时间片）
- 卸载模型
- 数学理论
- 环境配置

进行总结，找出比较好的优化目标，对现有算法进行调整

实验

- baseline：贪心、穷举、无调度、线性松弛、DQN
- discrete：D3QN、Rainbow。。。

总结

基本实现实验框架，包含模拟环境和调度算法

后续工作：

1. 选择更新颖的优化目标和环境设置
2. 调参、基线实验