Meta-Critic Reinforcement Learning for IOS-Assisted Multi-User Communications in Dynamic Environments

Qinpei Luo*, Boya Di*, Zhu Han†

* State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics, Peking University

† Electrical and Computer Engineering Department, University of Houston, TX, USA

VTC2023-Spring
FLORENCE
Connecting the Mobile World

Syllabus

北京大學
PEKING UNIVERSITY

- Introduction to IOS
- Related Works and Limitations
- System Model & Problem Formulation
- MC-DDPG: A Faster Method for IOS Configurations in Dynamic Environment
- Simulation Results & Conclusion

Agenda

北京大學
PEKING UNIVERSITY

Introduction to IOS

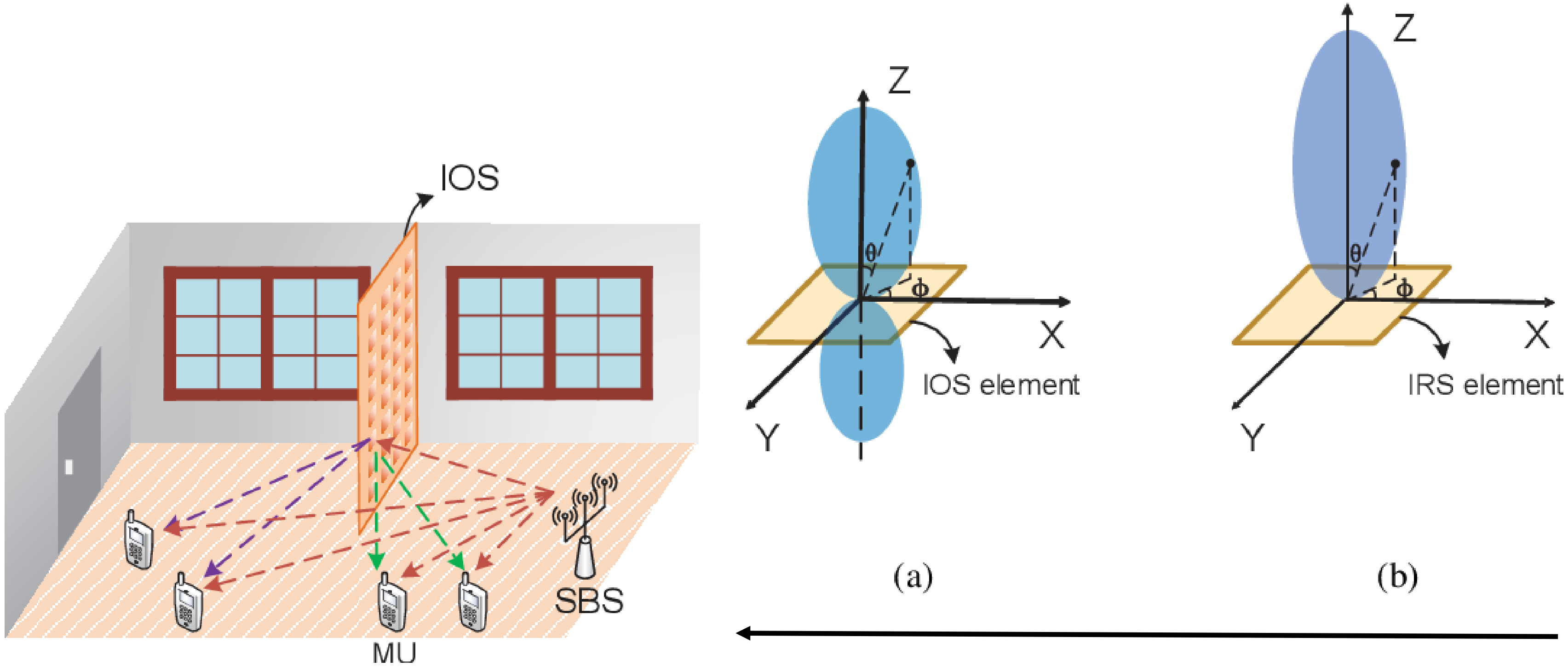
Related Works and Limitations

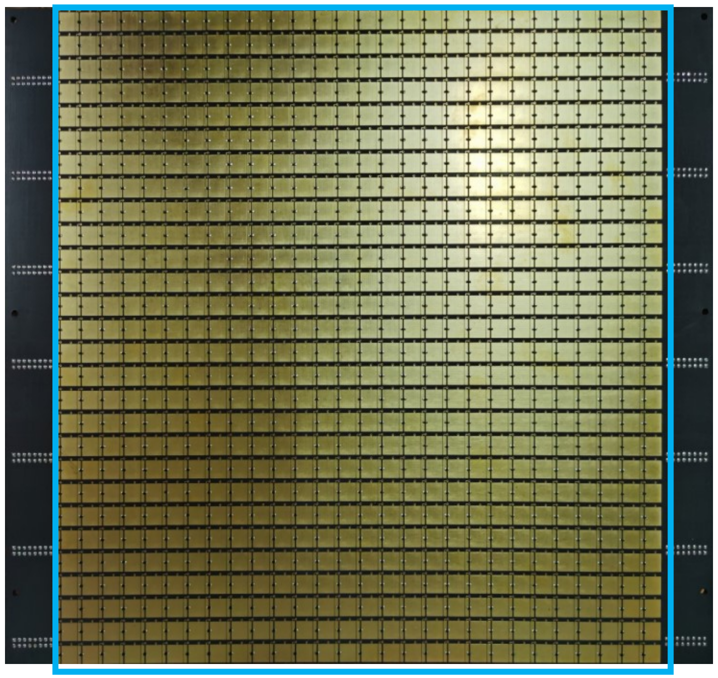
System Model & Problem Formulation

MC-DDPG: A Faster Method for IOS Configurations in Dynamic Environment

Simulation Results & Conclusion

What is IOS?A promising solution to enhance the capacity of wireless networks





Intelligent Omni-Surface (IOS)
Simultaneously Reflection & Refraction

Reflective Intelligent Surface (RIS)
Only Reflection of incident signal

*Source: Zhang, S., Zhang, H., Di, B., Tan, Y., Renzo, M.D., Han, Z., Poor, H.V., & Song, L. (2020). Intelligent Omni-Surface: Ubiquitous Wireless Transmission by Reflective-Transmissive Metasurface. *ArXiv, abs/2011.00765.*

Challenges for Implementation of IOS

PEKING UNIVERSITY 北京大學

- Numerous IOS elements
  - Phase shifts of all of IOS elements need to be configured simultaneously, which brings difficulty in solution searching.
- Dynamic Environment
  - The channel state of environment changes rapidly, which requires real-time updates of IOS configuration.
- The above two things combines together to require an efficient beamforming scheme to tackle numerous IOS elements adapting to the varying channel information, users' positions, etc.

Agenda

北京大學
PEKING UNIVERSITY

- Introduction to IOS
- Related Works and Limitations
- System Model & Problem Formulation
- MC-DDPG: A Faster Method for IOS Configurations in Dynamic Environment
- Simulation Results & Conclusion

Machine-Learning Based Beamforming

北京大學
PEKING UNIVERSITY

- Why ML is widely used?
  - Advanced ability in extracting features from channel state information.
- Reinforcement learning (RL) Method
  - Able to well depict the interaction process between intelligent surface and the environment.
  - HUANG, et al. (2020) develop a Deep RL based method to jointly design the transmit beamforming matrix and phase shifts of RIS.
  - LEE, et al. (2020) also use DRL to solve the problem of energy efficiency optimization.
  - ZHANG, et al. (2022) consider a system with multiple RISs and design a hierarchal policy network to improve the sum rate.

Limitations of Current Methods

Small-sized intelligent surface in consideration	⟷	Large-scale intelligent surface in real world
Static channel information and locations of users in consideration	⟷	Highly dynamic environment in real world

- We aim to develop an efficient beamforming scheme to address practical concerns
  - How to adapt to the dynamic case where the channel information and user positions vary with time?
  - How to deal with the numerous phase shift variables brought by a large-scale IOS in this case?

Agenda

北京大學
PEKING UNIVERSITY

Introduction to IOS

Related Works and Limitations

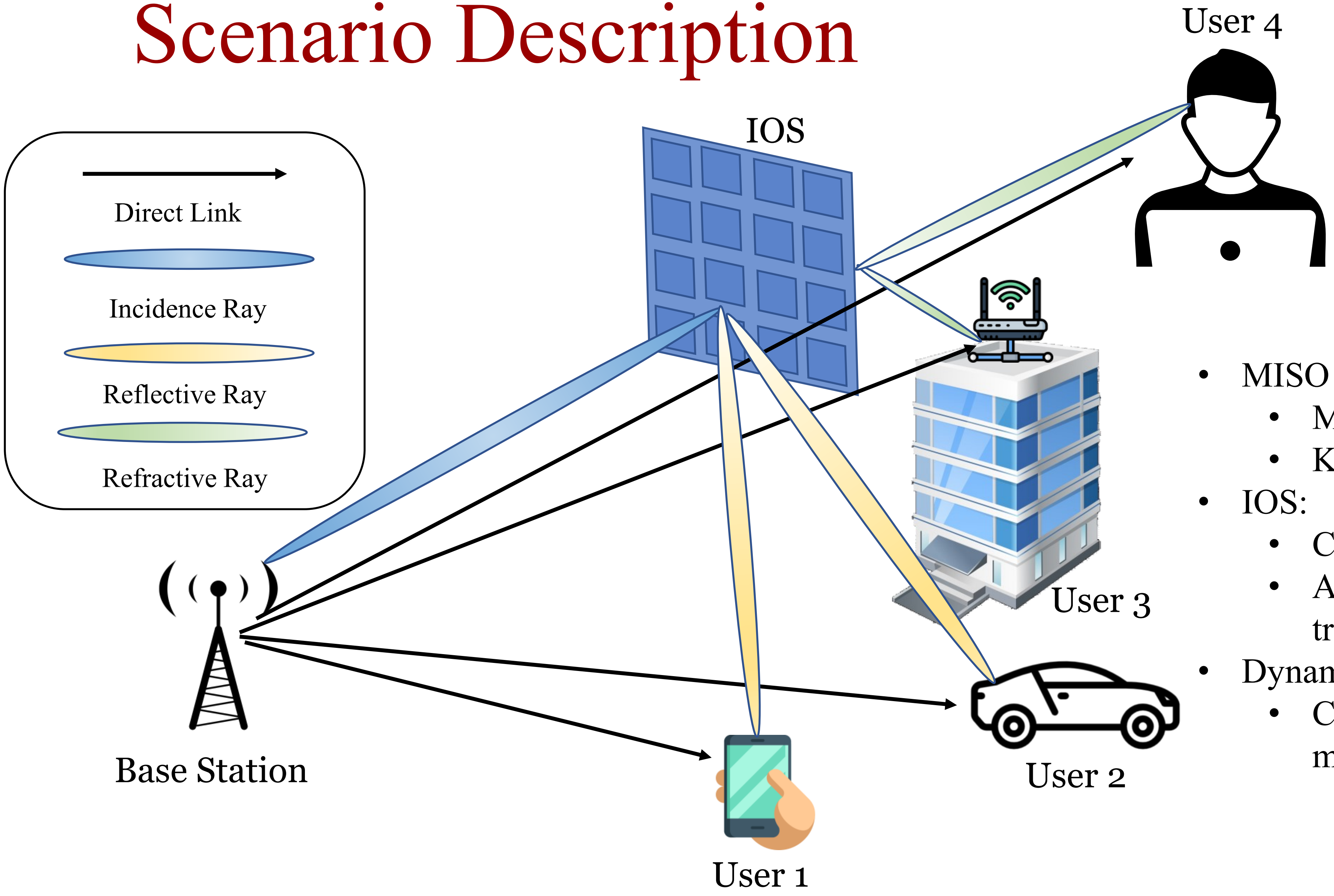
System Model & Problem Formulation

MC-DDPG:A Faster Method for IOS Configurations in Dynamic Environment

Simulation Results & Conclusion

Scenario Description

PEKING UNIVERSITY 北京大學



- MISO system:
  - M antennas at Base Station (BS)
  - K users with single antenna
- IOS:
  - Consisting of N elements
  - Able to reflect and refract the transmit signal
- Dynamic Environment
  - Channel states and locations of users may vary with time

Channel Model

PEKING UNIVERSITY

- For each user k we consider the Light-of-sight channel as a hybrid channel

$$\mathbf{H}_k^{LOS} = \Delta^u \mathbf{H}_{IU,k} \boldsymbol{\Theta} \mathbf{H}_{BI} + \mathbf{H}_{BU,k}$$

Where $u \in \{r, t\}$ refers to the reflective and refractive respectively, while Δ^u represents the energy split for each type of users. $\boldsymbol{\Theta} = \mathrm{diag}\{[\mathrm{e}^{\mathrm{j}\theta_1}, \mathrm{e}^{\mathrm{j}\theta_2}, \cdots, \mathrm{e}^{\mathrm{j}\theta_N}]\}$ stands for the phase shifts of IOS.

- According to Saleh-Valenzuela Model, the channel of IOS-user, BS-IOS and BS-user can be further written into

$$\mathbf{H}_{BI} = \sqrt{S_1} \boldsymbol{A}_I \boldsymbol{\Sigma}_{BI} \boldsymbol{D}_B^H, \mathbf{H}_{IU} = \sqrt{S_{2,k}} \boldsymbol{A}_{IU,k} \boldsymbol{\Sigma}_{IU,k} \boldsymbol{D}_{I,k}^H, \mathbf{H}_{BU} = \sqrt{S_{3,k}} \boldsymbol{A}_{BU,k} \boldsymbol{\Sigma}_{BU,k} \boldsymbol{D}_{B,k}^H$$

In which $\boldsymbol{A}$ and $\boldsymbol{D}$ refers to transmit/receive steering matrices, the i-th column of each matrix is the steering vector and can be expressed by $f(M, \theta) = \frac{1}{\sqrt{M}}\left[1, e^{j\pi\theta}, \cdots, e^{j\pi(M-1)\theta}\right]^H$ where M is the number of antennas and θ is the Angle-of-Arrival (AoA) or Angle-of-Departure (AoD). $\boldsymbol{\Sigma}$ represents the gain of each channel, while S stands for the path loss.

- We assume the equivalent channel of each user follows Rician Distribution, i.e.,

$$\mathbf{H}_k = \sqrt{\frac{K^R}{1+K^R}} \mathbf{H}_k^{LOS} + \sqrt{\frac{1}{1+K^R}} \mathbf{H}_k^{NLOS}$$

K^R is the Rician factor. $\mathbf{H}_\mathrm{k}^{\mathrm{NLOS}}$ has similar expression as $\mathbf{H}_\mathrm{k}^{\mathrm{LOS}}$, but its AoDs or AoAs are randomly generated.

Finite State Markov Channel

- We choose to fix the LOS component and discretize the NLOS component $\mathbf{H}_k^{NLOS}$ into L levels.
- $\mathcal{H} = \{\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_k\}$
- Transition probability matrix: $\boldsymbol{P} = \begin{pmatrix} p_{1,1} & \cdots & p_{1,L} \\ \vdots & \ddots & \vdots \\ p_{L,1} & \cdots & p_{L,L} \end{pmatrix}$
- $p_{l,l'} = Prob[\mathbf{H}_{t+1} = \mathbf{H}_{l'} | \mathbf{H}_t = \mathbf{H}_l], \mathbf{H}_l, \mathbf{H}_{l'} \in \mathcal{H}$
- P is generated randomly, so do the NLOS components.

Sum Rate Maximization Formulation

- We consider the sum rate maximization problem of all users in T time slots.

$$y_{k,t} = \left(\Delta_k \mathbf{H}_{IU,k} \mathbf{\Theta}_t \mathbf{H}_{BI} + \mathbf{H}_{BU,k}\right) \sum_{j=1}^{K} \boldsymbol{V}_{j,t} m_j + n_{k,t}$$

$$\gamma_{k,t} = \frac{\left|\left(\Delta_k \mathbf{H}_{IU,k} \Theta \mathbf{H}_{BI} + \mathbf{H}_{BU,k}\right) \boldsymbol{V}_{k,t} m_k\right|^2}{\left(\Delta_k \mathbf{H}_{IU,k} \Theta \mathbf{H}_{BI} + \mathbf{H}_{BU,k}\right) \sum_{j=1, j\neq k}^{K} \boldsymbol{V}_{j,t} m_j + n_{k,t}}$$

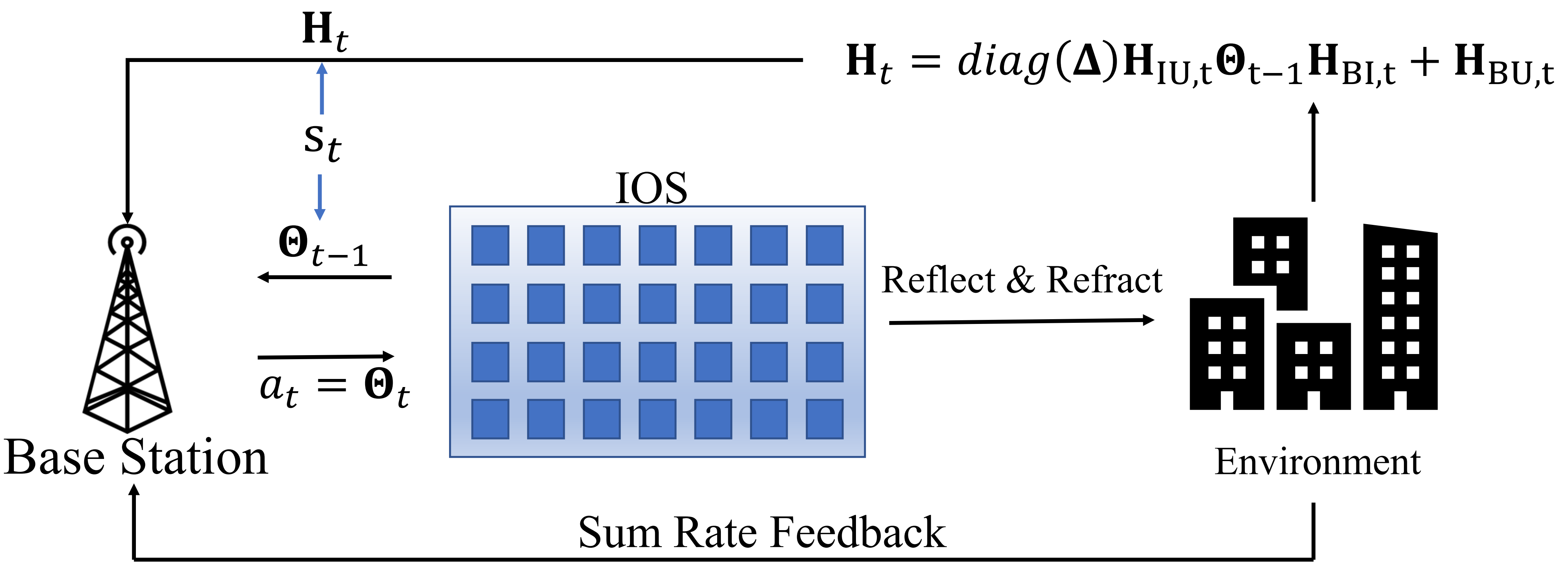
$$R_{k,t} = \log_2(1 + \gamma_{k,t})$$

$$\boldsymbol{P1}: \max_{\{\boldsymbol{V}_t, \mathbf{\Theta}_t\}} \sum_t \sum_k R_{k,t}$$

- Solving $\boldsymbol{V}_t$ with fixed digital beamforming method as water-filling and zero-forcing, we can rewrite the problem as

$$\boldsymbol{P2}: \max_{\mathbf{\Theta}_t} \sum_t \sum_k R_{k,t}$$

Markov Decision Process Reformulation



$$r_t = \eta \sum_{k=1}^{K} R_{k,t}$$

(Next Time Slot)

$$\mathbf{H}_{t+1} = diag(\mathbf{\Delta})\mathbf{H}_{\mathrm{IU,t+1}}\mathbf{\Theta}_{\mathrm{t}}\mathbf{H}_{\mathrm{BI,t+1}} + \mathbf{H}_{\mathrm{BU,t+1}}$$

Agenda

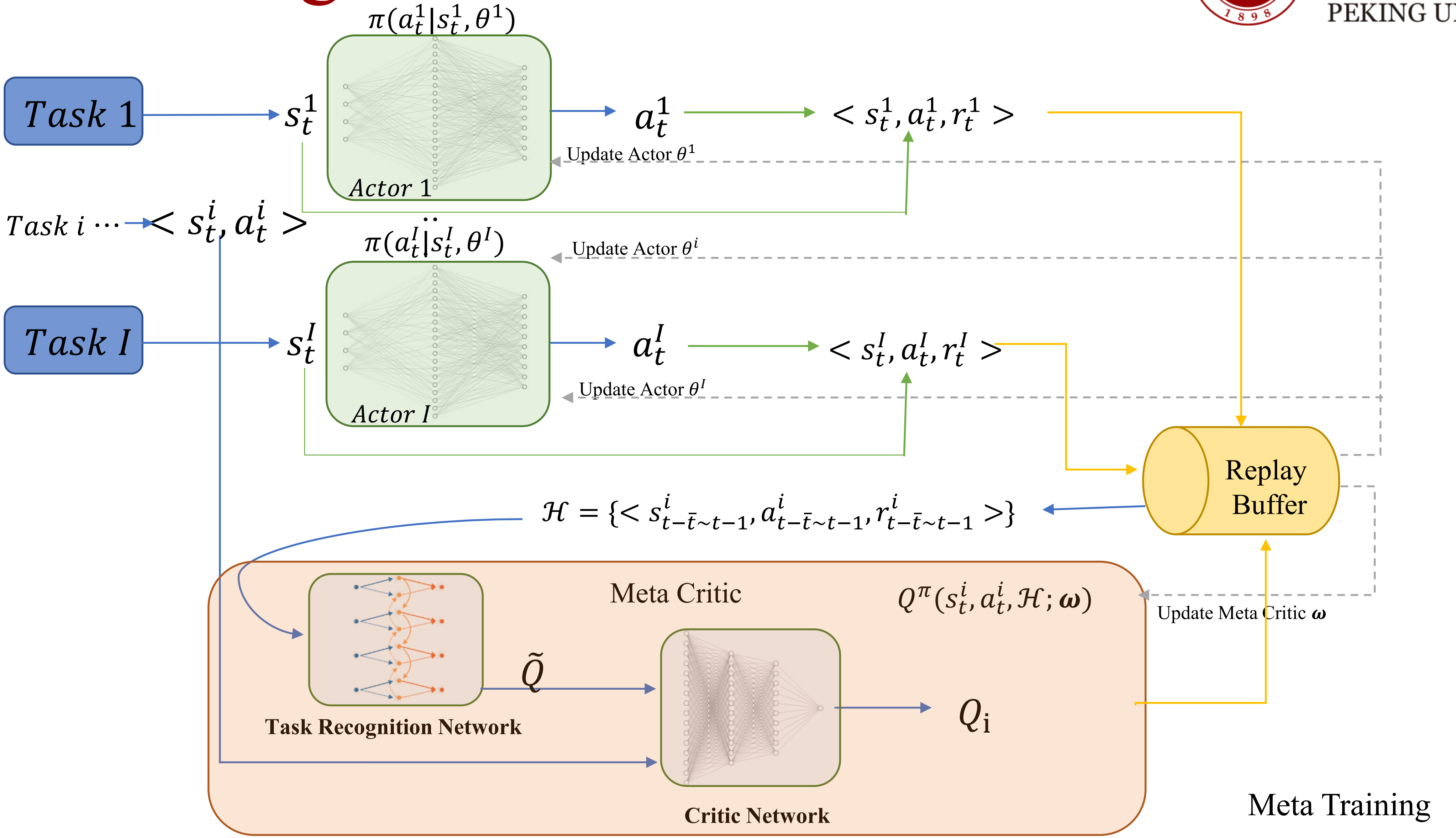
北京大學
PEKING UNIVERSITY

- Introduction to IOS
- Related Works and Limitations
- System Model & Problem Formulation
- MC-DDPG: A Faster Method for IOS Configurations in Dynamic Environment
- Simulation Results & Conclusion

Basic Definition

- **Task:** Denotes a process of the BS maximizing the sum rates of all users in a fixed number of time slots. For different tasks, the parameters of BS and IOS are set as the same, while the channel states and locations of users are various.
- **Actor:** It receives the information of state from the task in each time slot and outputs correspondent action. We adopt a neural network as the policy of the actor.
- **Meta-Critic:** Consists of two parts, a task recognition network and a critic network. The former extracts the history information and generates the task-recognition Q-value, while the critic network outputs a task-specific Q-value to update the actor networks.

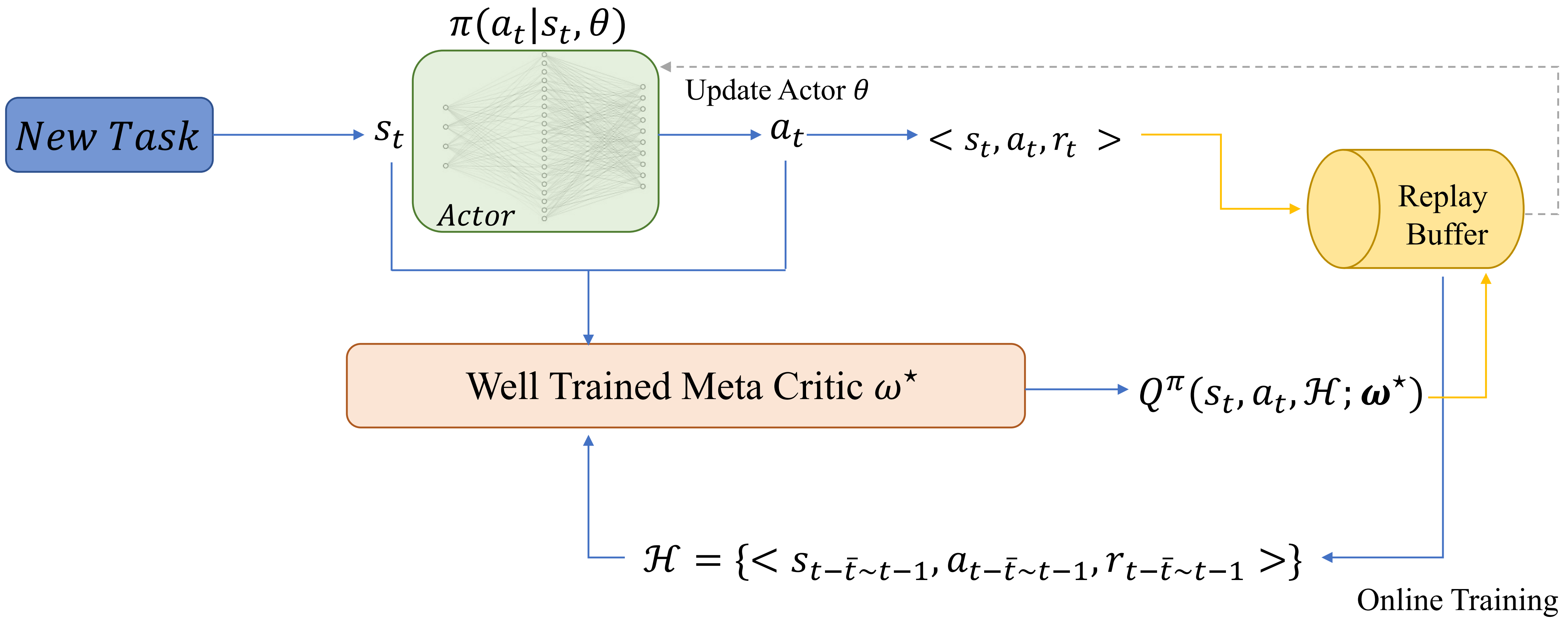
Meta Learning Phase



In the meta-learning phase, both the actors and meta-critic are updated.

Online Learning Phase

北京大學 PEKING UNIVERSITY



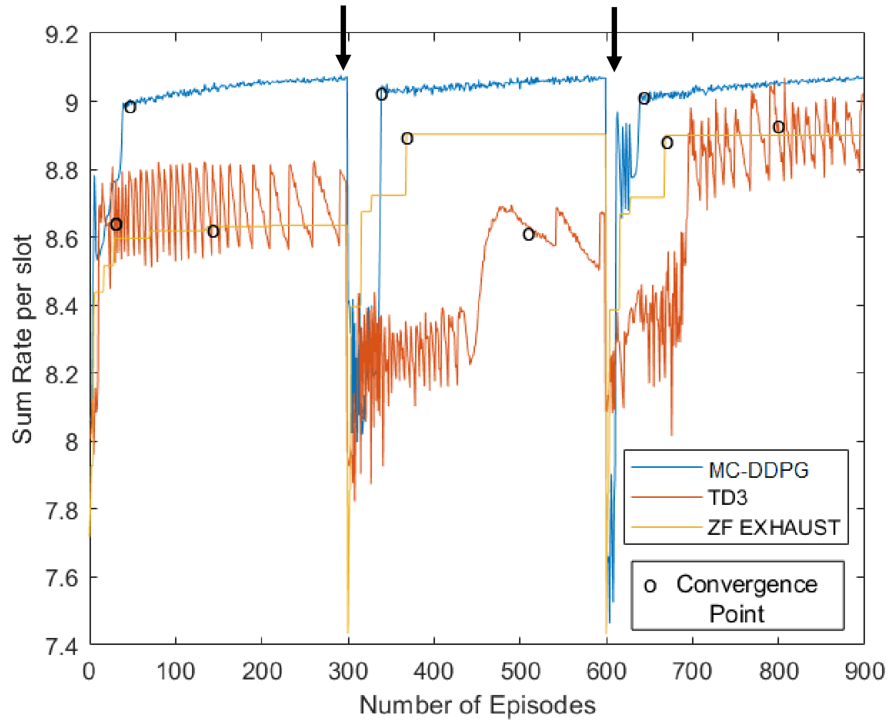
In the online learning phase, only the actors are updated, while the well-trained meta-critic is kept static.

Agenda

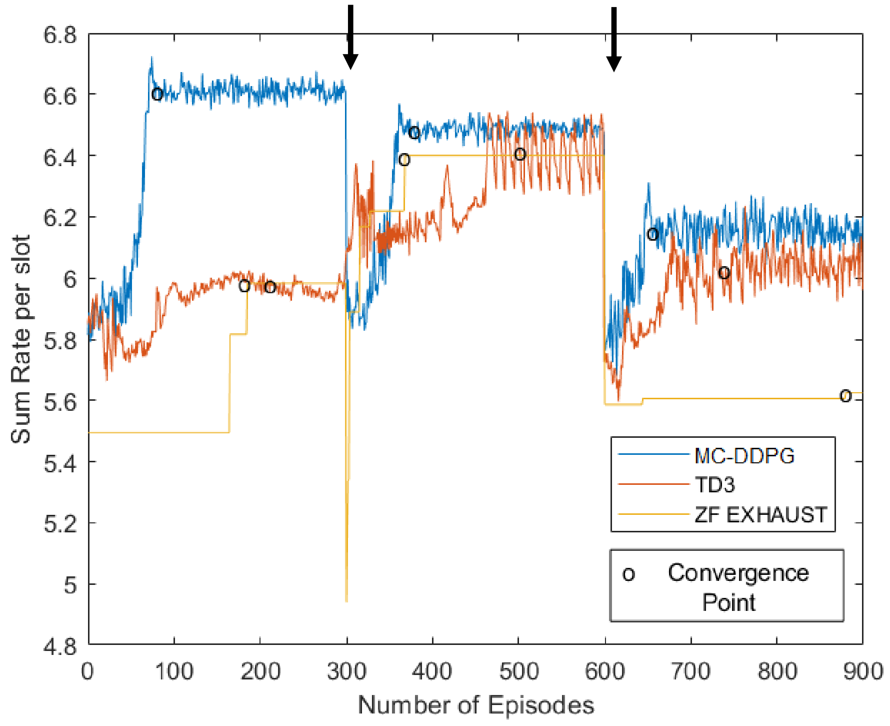
- Introduction to IOS
- Related Works and Limitations
- System Model & Problem Formulation
- MC-DDPG:A Faster Method for IOS Configurations in Dynamic Environment
- **Simulation Results & Conclusion**

Sum Rate Performance in Dynamic Settings

Change point of different tasks



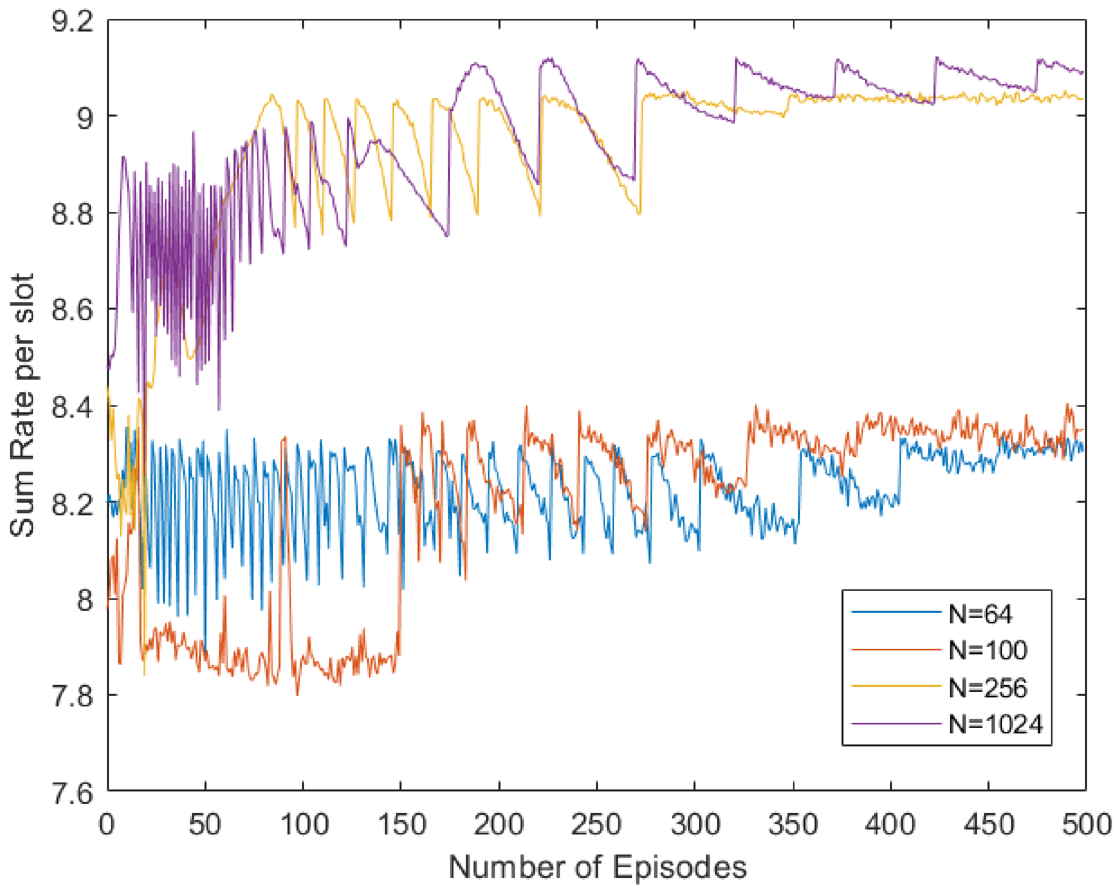
Sum Rate Performance with respect to the varying channel states



Sum Rate Performance with respect to User's locations

Influence of the Number of IOS Elements

北京大學 PEKING UNIVERSITY



Conclusion

北京大學 PEKING UNIVERSITY

- Current works seldom consider the challenges brought by a large number of IOS elements and dynamic environment.
- We proposed MC-DDPG, a meta-critic RL scheme for sum rate maximization given the limited CSI, which is able to:
  - Achieve a faster convergence speed and a higher sum rate compared to the benchmarks.
  - The robustness of MC-DDPG against IOS sizes is verified.
- We can draw two take-away conclusions:
  - The designed meta-critic significantly enhances the robustness of the IOS-assisted multi-user communications against user mobility and the dynamic CSI.
  - There exists a trade-off between the convergence speed and the achievable sum rate of MC-DDPG.

Thank you!
Q&A

Qinpei Luo*, Boya Di*, Zhu Han†

* State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Electronics, Peking University

† Electrical and Computer Engineering Department, University of Houston, TX, USA