中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Schaferct: Accurate Bandwidth Prediction for Real-Time Media Streaming with Offline Reinforcement Learning

Qingyue Tan, Gerui Lv, Xing Fang, Jiaxing Zhang, Zejun Yang, Yuan Jiang, **Qinghua Wu**

Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences

Grand Challenge

- **Goal**: Developing a deep learning-based policy model (receiver-side bandwidth estimator, π) with **offline RL** techniques to improve **QoE** for **RTC** system users as measured by **objective audio/video quality scores**.

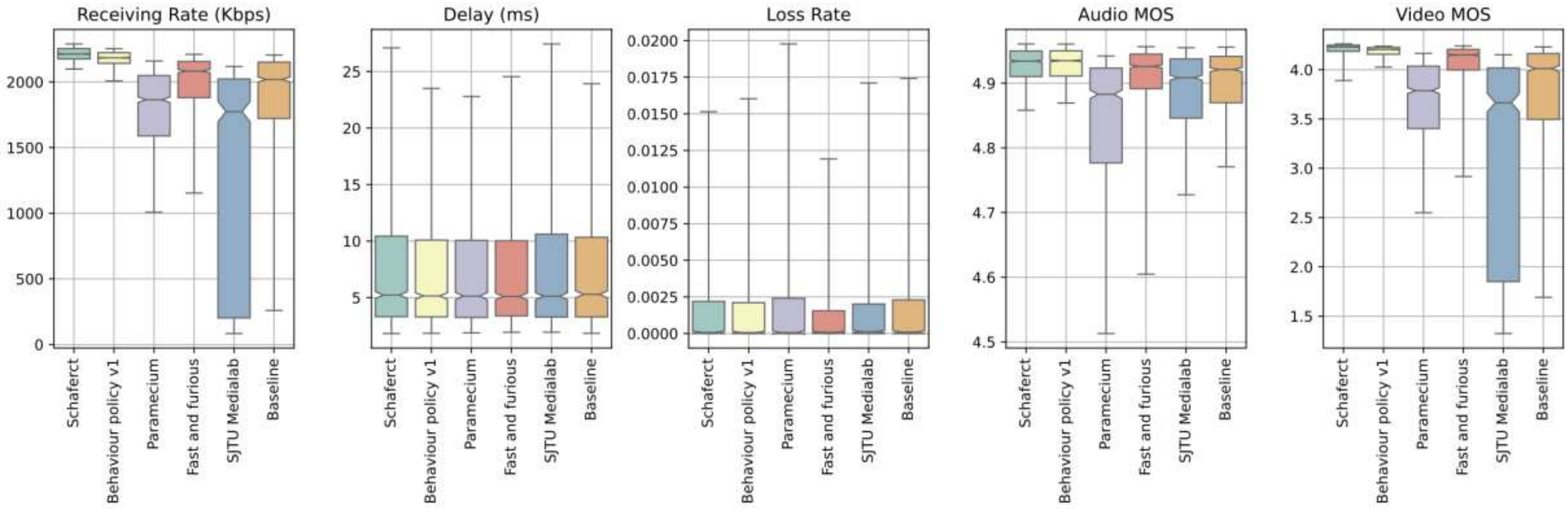
- **Given**: Dataset of trajectories for Microsoft Teams audio/video calls.
  - Training dataset: 18859 calls
  - Evaluation dataset: 9405 calls containing ground truth (bottleneck link bandwidth).

- **Evaluation**: The scores in 2-stage evaluation. The scoring function:

$$\mathbb{E}_{call\ legs}\left[\mathbb{E}_n\left[r_n^{audio} + r_n^{video}\right]\right] \epsilon\ [0, 10]$$

Results: Conducted by Grand Challenge Committee

- Our model, Schaferct, demonstrates comparable performance to the best behavior policy (v1) in the released datasets across all metrics.



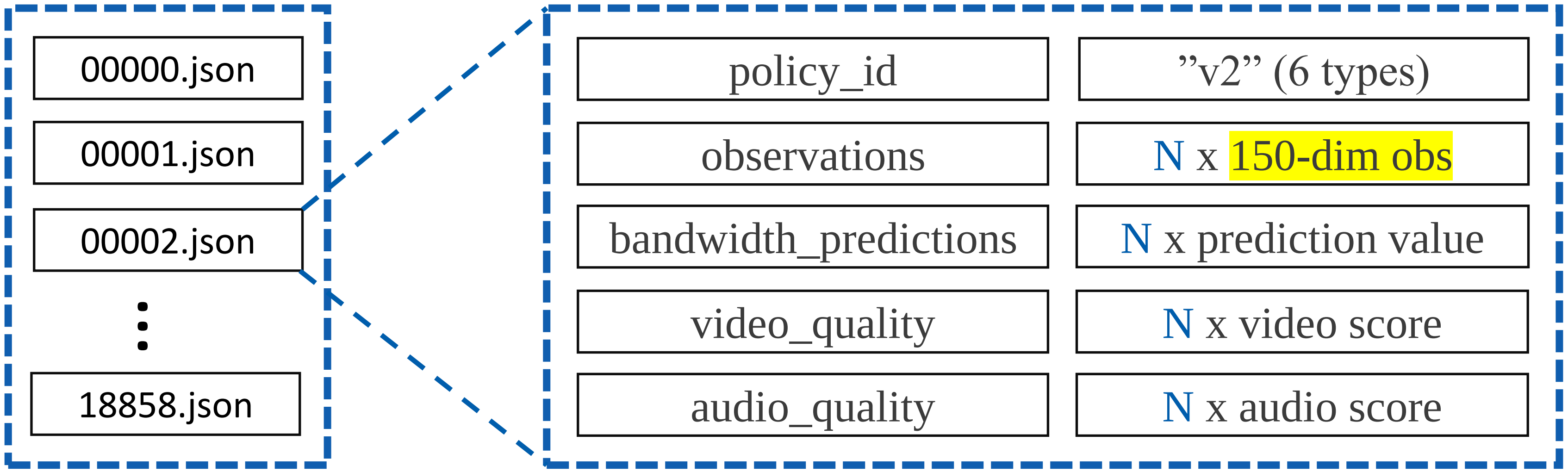
Results: Final Evaluation Stage Rankings

- A real-world test (600 3-minute calls) across diverse network conditions with temporal fluctuations over the internet.

Rank	Model	Score	95% CI
1	**Schaferct**	8.93	[8.88, 8.97]
2	**Fast and furious**	8.70	[8.65, 8.76]
3	Paramecium	8.34	[8.28, 8.39]
4	SJTU Medialab	7.89	[7.82, 7.96]

Dataset

18,859 sessions **N (3.5K+) transitions**

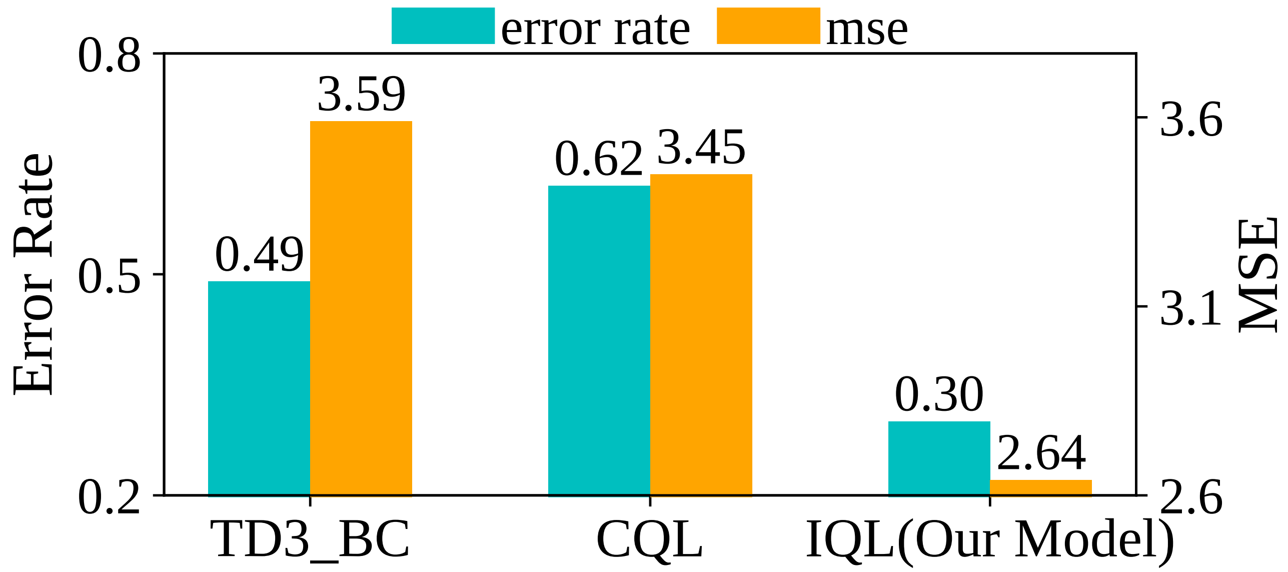


150-dim obs: 15 Features X [5 Long MI (600ms) + 5 Short MI (60ms)]

1	Receiving rate	6	Minimum seen delay	11	Packet loss ratio
2	Number of received packets	7	Delay ratio	12	Average number of lost packets
3	Received bytes	8	Delay average minimum difference	13	Video packets probability
4	Queuing delay	9	Packet interarrival time	14	Audio packets probability
5	Delay	10	Packet jitter	15	Probing packets probability

Design Choice: Offline RL Algorithm

- **Main challenge in offline RL**: trading off policy improvement against distributional shift



- **Implicit Q-Learning (IQL)** solves this by trading off between how much the policy **improves** and how vulnerable it is to **misestimation** due to distributional shift, by never needing to directly query or estimate values for actions that were not seen in the data.

Design Detail of IQL

- In the policy evaluation stage, IQL uses the **expectile regression update** method to approximate the optimal value function $V(s)$:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[L_2^\tau(\, Q_{\hat{\theta}}(s,a) \ - \ V_\psi(s)\,)]$$

$$L_2^\tau(u) = |\tau \ - 1(u < 0)|u^2$$

- The state-action value function $Q_\theta(s,a)$ is updated by minimizing the temporal difference (TD) loss:

$$\mathcal{L}_{\mathrm{Q}}(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}[(r(s,a) + \gamma V_\psi(s') - Q_\theta(s,a))^2]$$

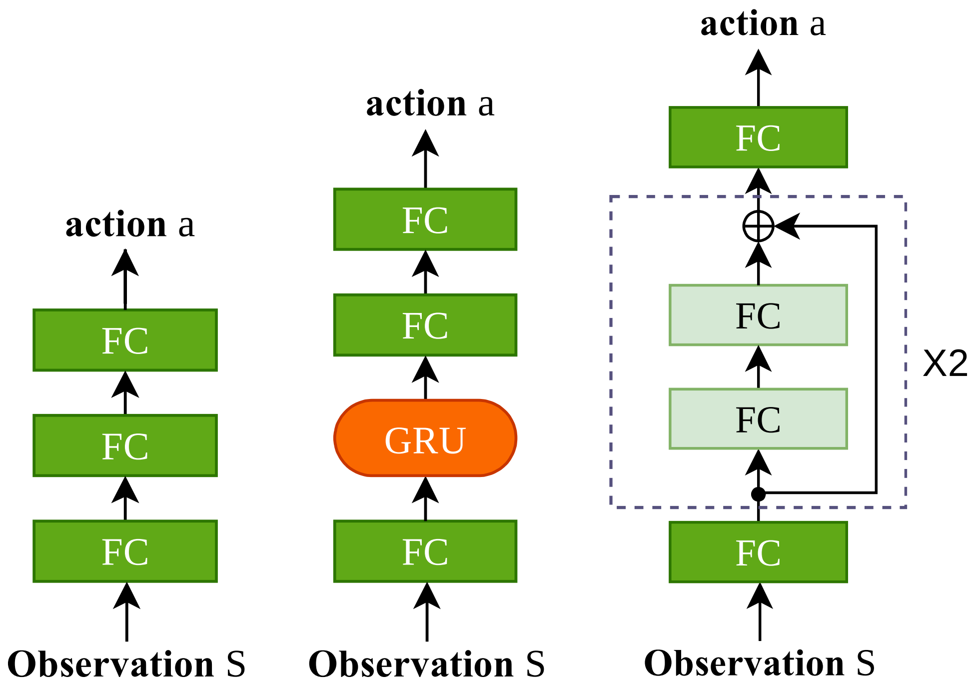
- In the policy extraction stage, IQL minimizes the loss for optimizing the final policy $\pi_\phi(s)$ is:

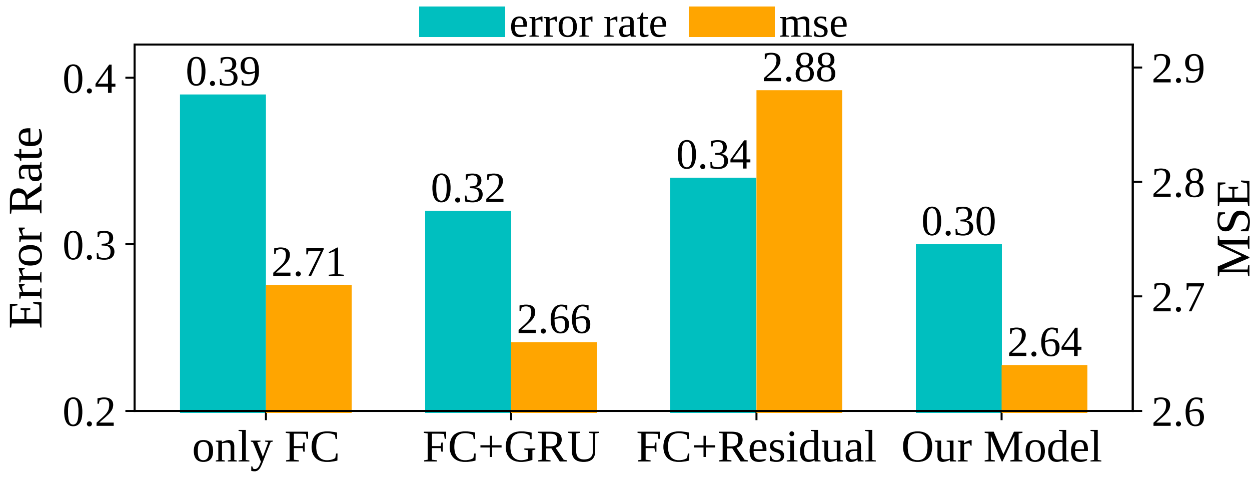
$$\mathcal{L}_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[\exp(\,\beta\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\,\log \pi_\phi\,(a|s)\,)]$$

Design Choice: weight αin Reward Function

- let α = 1.5.

Design Choice: Actor Network Structure

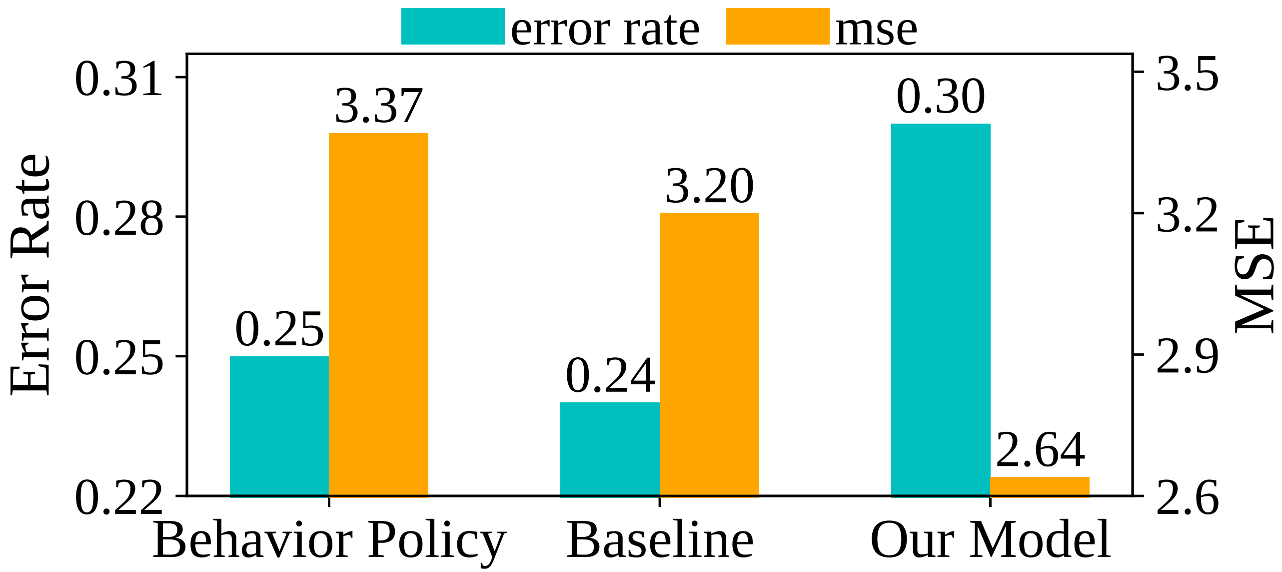




- (i) Only three FC layers.
- (ii) Three FC layers with GRU.
- (iii) Three FC layers with two Residual Blocks.

Evaluation: Prediction Accuracy

- Our model has the lowest MSE and lowest over-estimated rate, yet the highest error rate.

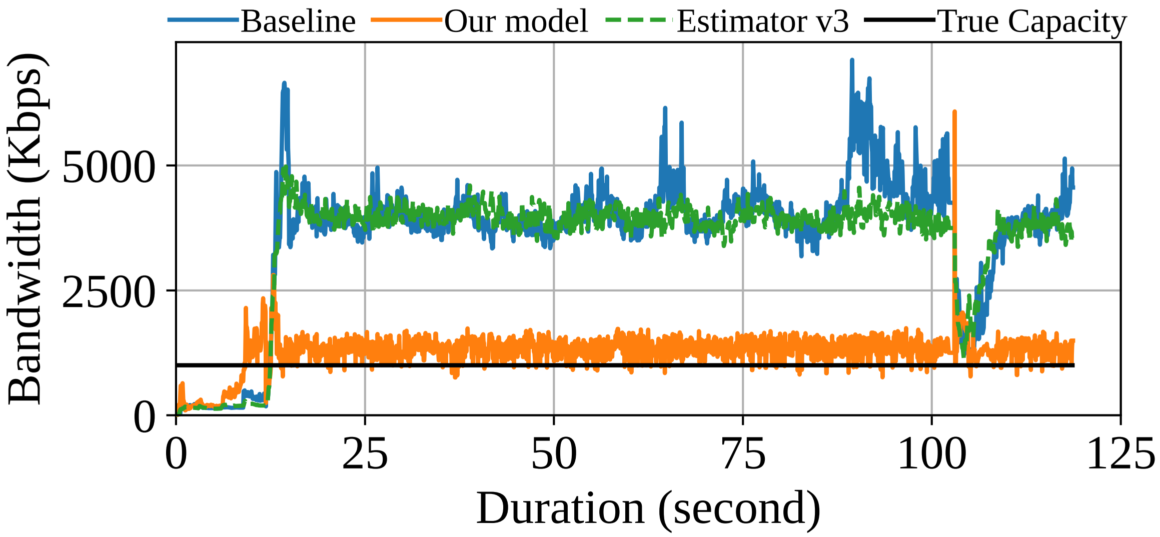


$$error\_rate = \mathbb{E}[\min(1, \frac{|\hat{B} - B|}{B})]$$

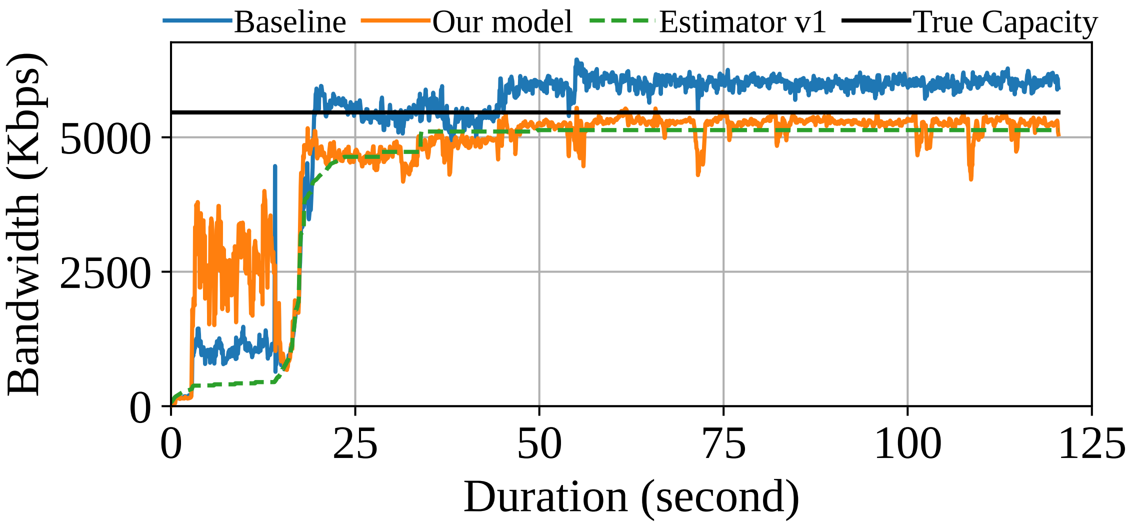
$$MSE = \mathbb{E}[(\hat{B} - B)^2]$$

Evaluation: Case Study

- Case #1:
  The behavior policy: significantly overestimates the link bandwidth.
  The baseline model: follows the behavior policy, end up in overestimation.
  Our model: closely aligns with the true capacity.



- Case #2:
  The baseline: overestimate after the start-up phase.
  Our model: align with the behavior policy with more conservative and accurate predictions.



Limitations

- **Dataset**: Only 1,800 sessions are used for training due to the hardware constraints (e.g., GPU memory size) in our training environment;
- **Selection:** Session selection is random, without considering the distribution of observation-action-reward.
- **Feature engineering**: All metrics are used.

Conclusion

- We proposed an offline-RL-based bandwidth prediction method to predict the bottleneck link bandwidth.
- Based on IQL, we redesign the neural network structure and the reward function.
- Our model reduces 18%-22% MSE compared to both the baseline and six behavior policies, and won **the first prize** of the Bandwidth Estimation Challenge at ACM MMSys 2024.

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Thanks!

Q&A