AReaL: A Large-Scale Asynchronous Reinforcement Learning System for Language Reasoning

Wei Fu (me), Jiaxuan Gao , Xujie Shen , Chen Zhu , Zhiyu Mei , Chuyi He , Shusheng Xu, Guo Wei , Jun Mei , Jiashu Wang, Tongkai Yang, Binhang Yuan, Yi Wu

Contact & Collaboration: fuwth17@gmail.com & jxwuyi@gmail.com

清华大学 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

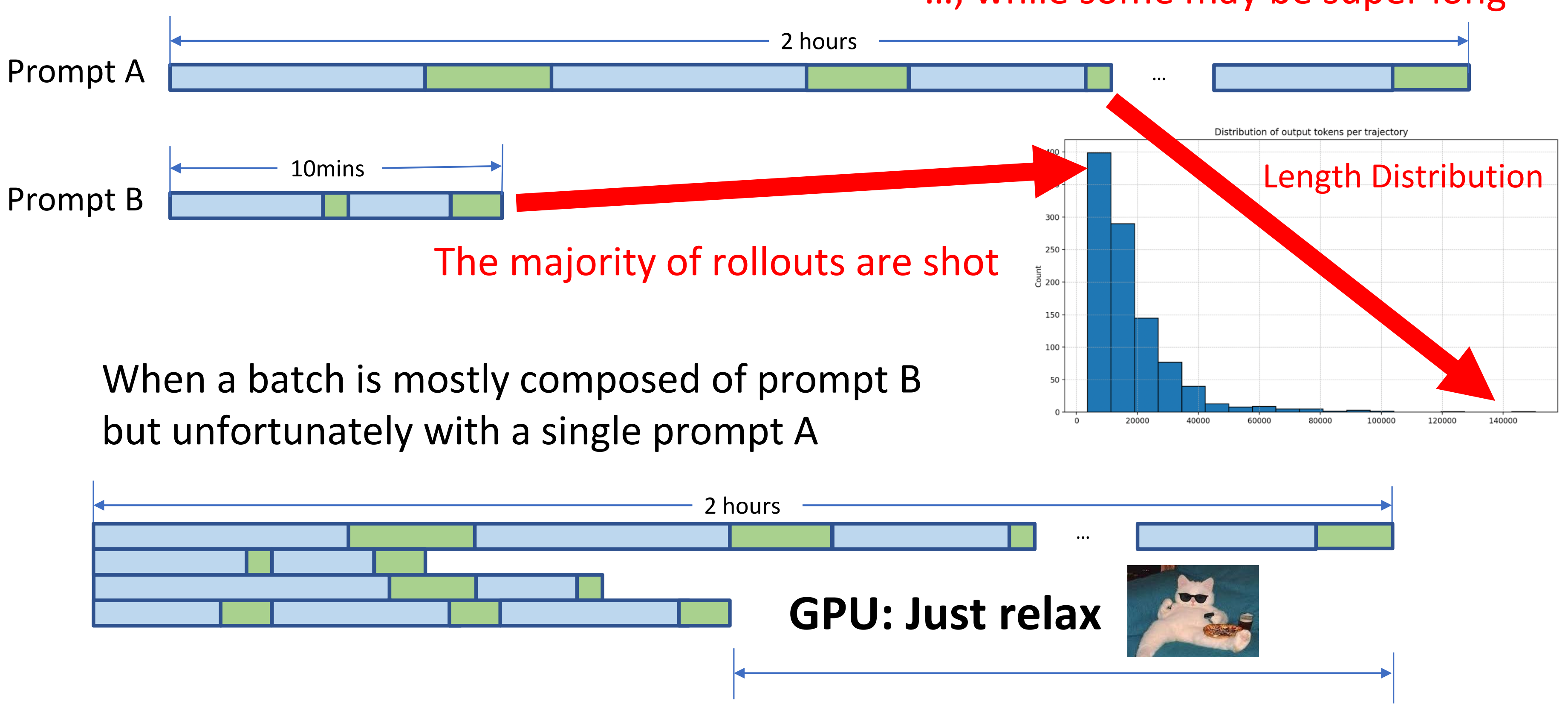
蚂蚁集团
ANT GROUP

NEURAL INFORMATION PROCESSING SYSTEMS

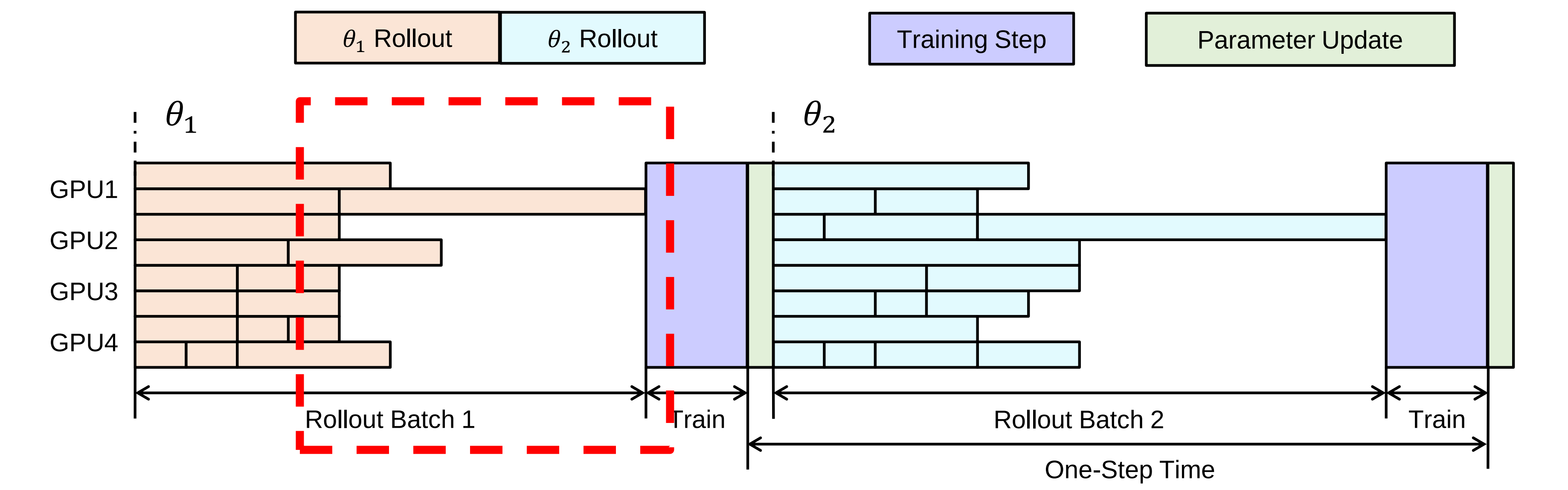
CHALLENGES OF SYNC. RL SYSTEMS

Challenge 1: Significant ***GPU idle times during*** inference/generation when lengths vary

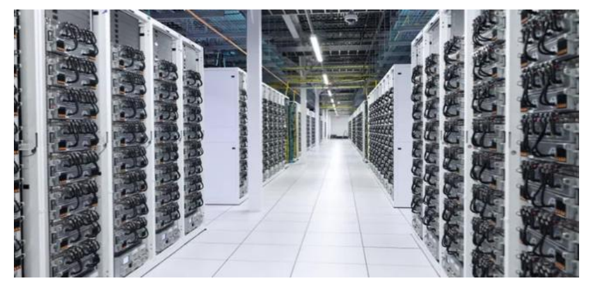
Multi-Turn (>128 turns) Agentic RL with Reasoning



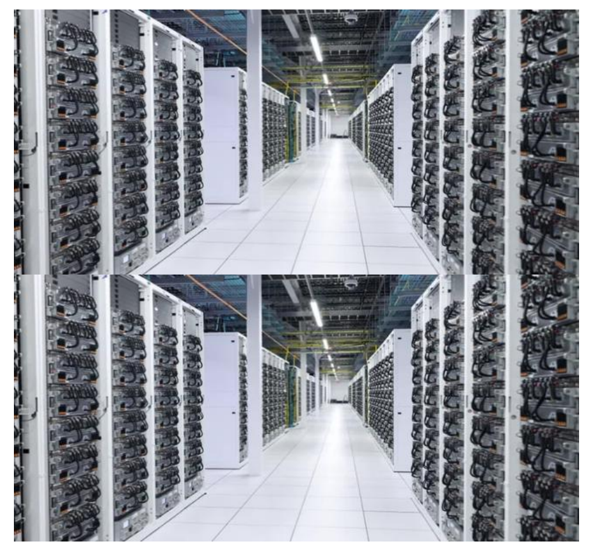
When a batch is mostly composed of prompt B but unfortunately with a single prompt A



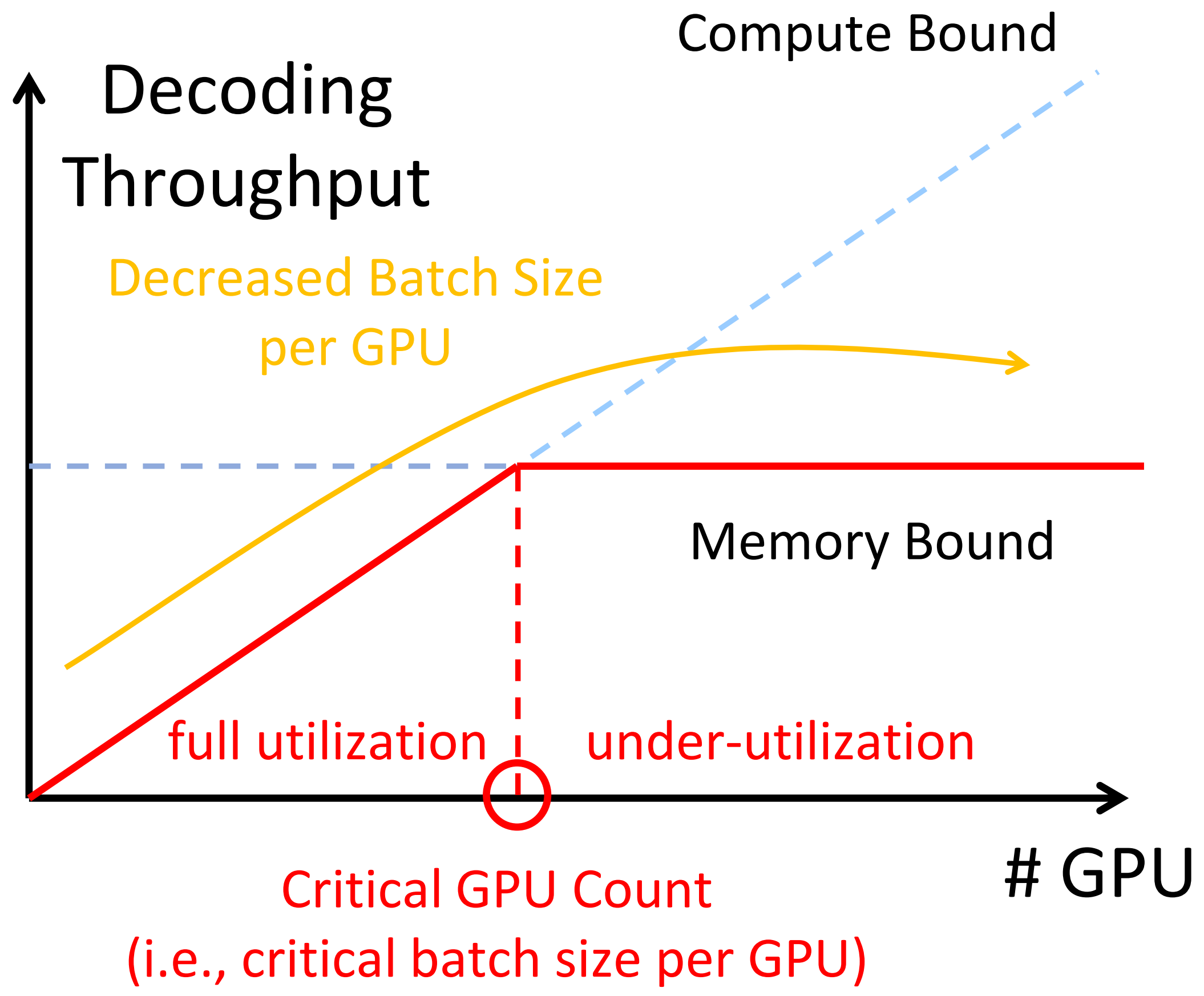
Challenge 2: ***Hard to scale up***



= 100 hours RL training



≠ 50 hours RL training



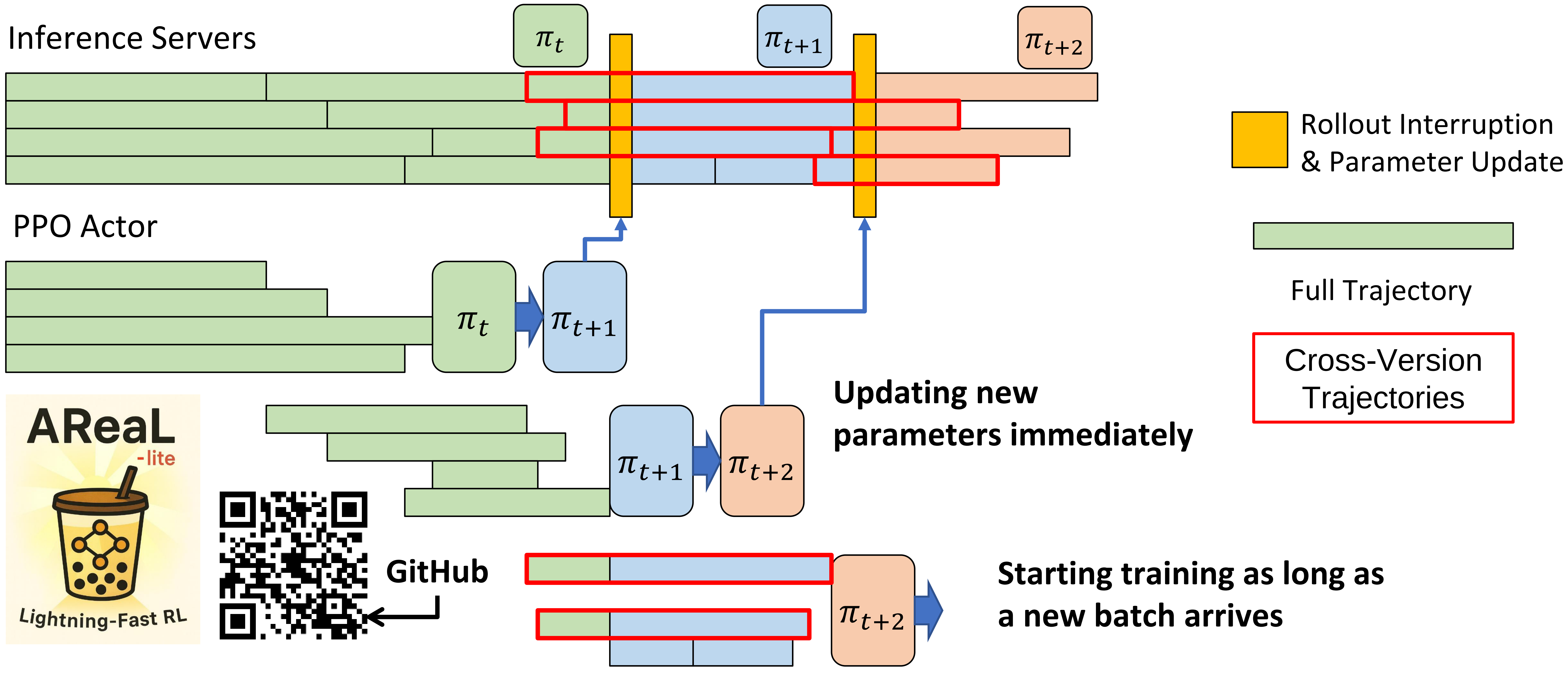
FULLY ASYNCHRONOUS RL IN AREAL

Keys to the challenges:

1. Continuous Rollout Batching with Interruption

2. Disaggregated and Overlapped Training & Inference

Both training and inference achieve full GPU utilization



Updating new parameters immediately

Starting training as long as a new batch arrives

Algorithmic Challenges

•Data Staleness

Data generated from old policies impedes efficient learning.

•Inconsistent Policy Versions

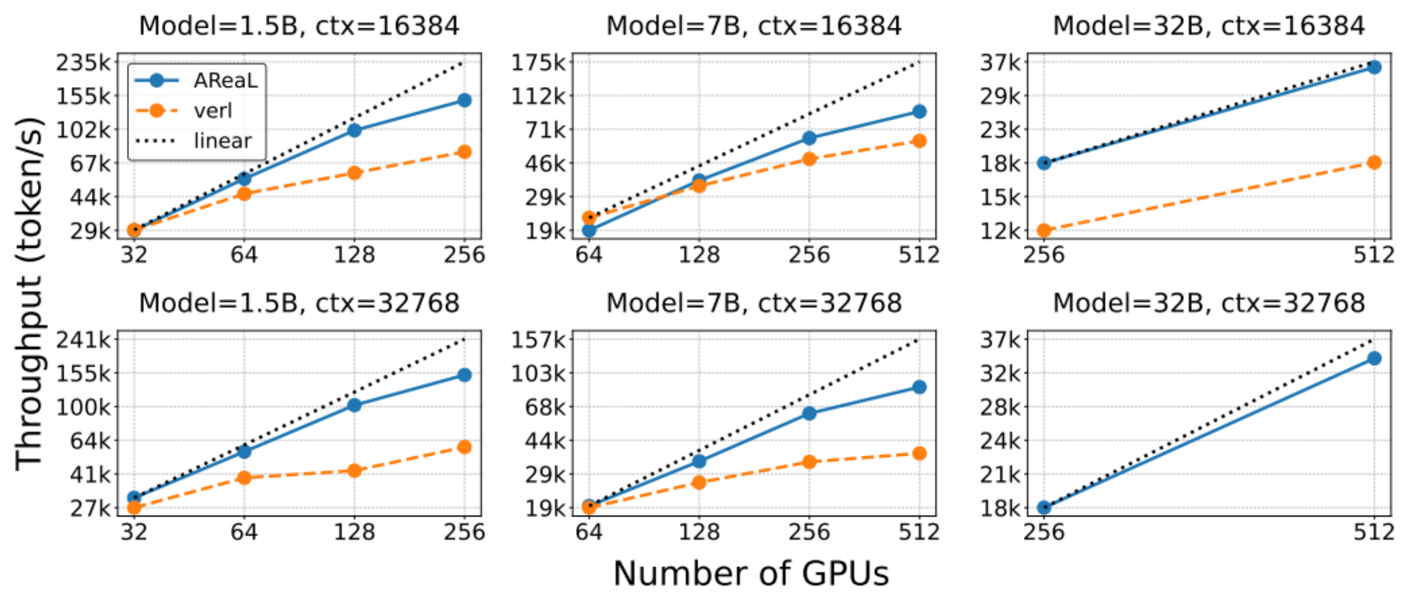
Interrupted trajectories involve segments generated from different policy versions, which violates the PPO objective.

•Staleness-Aware Training

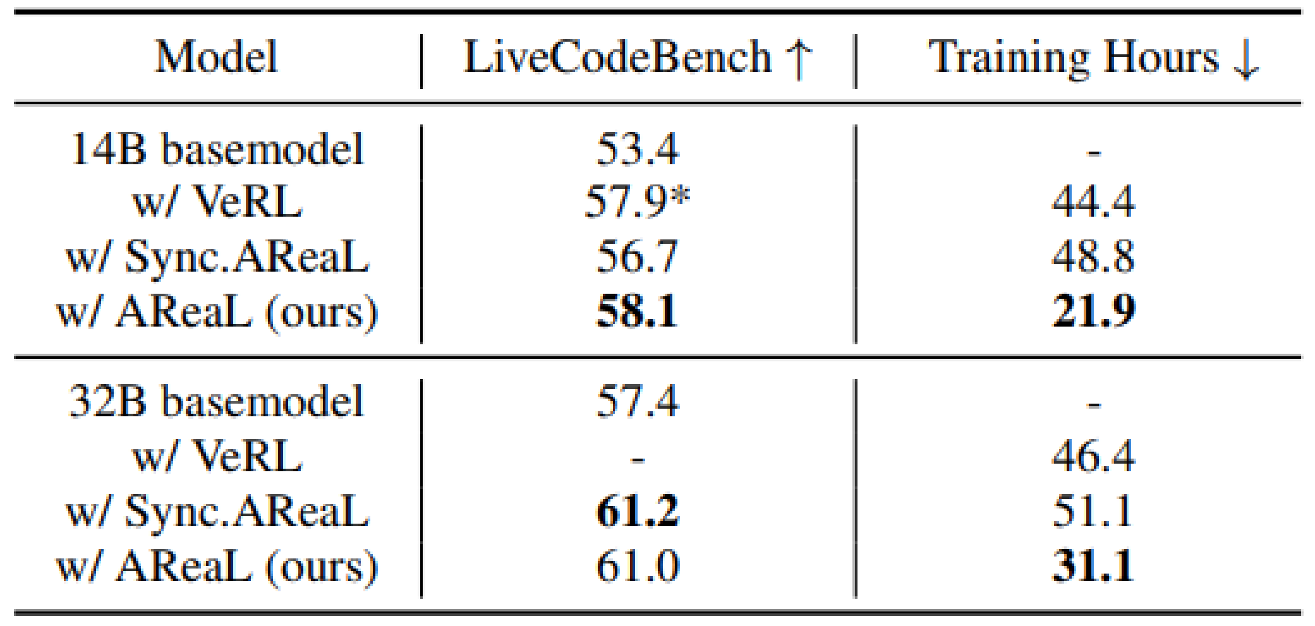
Control the maximum staleness with algorithm-system co-design.

•Decoupled PPO objective

$$J(\theta) = \mathbb{E}_{q\sim\mathcal{D},a_t\sim\pi_{\text{behav}}}\left[\sum_{t=1}^{H}\min\Big(\underbrace{\boxed{\frac{\pi_\theta}{\pi_{\text{behav}}}}}_{\text{Importance Ratio}}\hat{A}_t,\ \overbrace{\frac{\pi_{\text{prox}}}{\pi_{\text{behav}}}\text{clip}\Big(\underbrace{\boxed{\frac{\pi_\theta}{\pi_{\text{prox}}}}}_{\text{Trust Region Center}},1-\epsilon,1+\epsilon\Big)\hat{A}_t}^{\text{Importance Ratio}}\Big)\right]$$

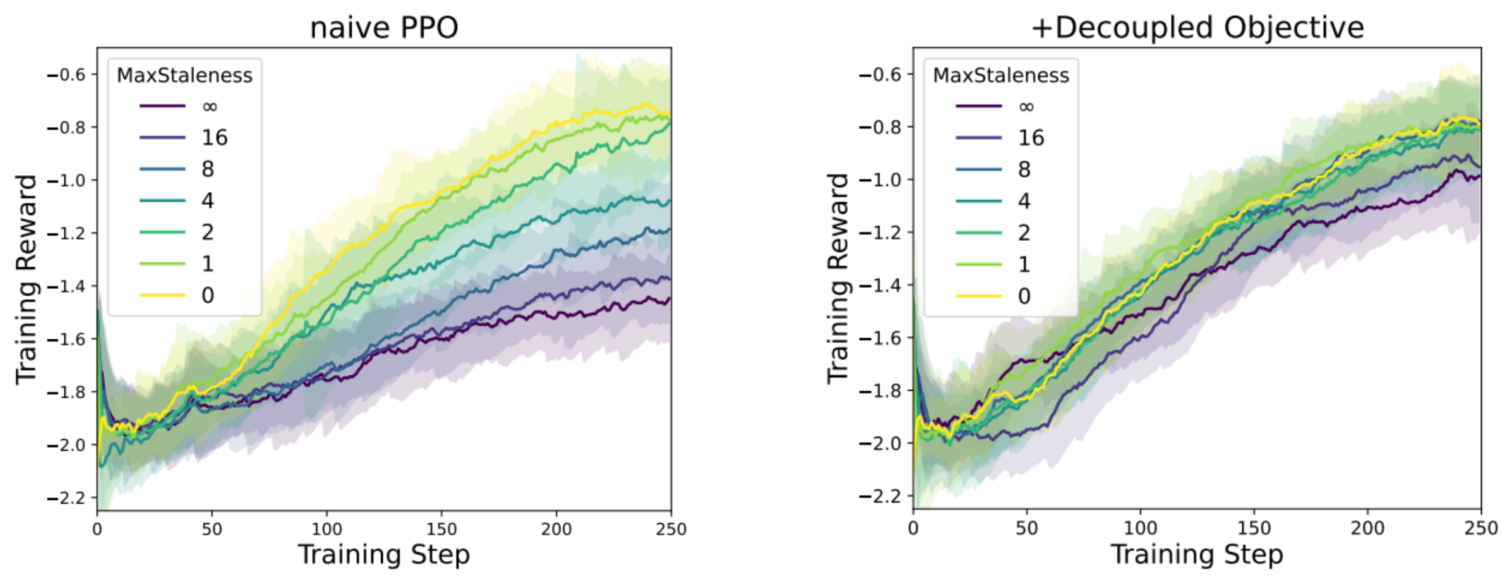


2x Higher **Effective Training Throughput Compared with veRL**



Model	LiveCodeBench ↑	Training Hours ↓
14B basemodel	53.4	-
w/ VeRL	57.9*	44.4
w/ Sync.AReaL	56.7	48.8
w/ AReaL (ours)	**58.1**	**21.9**
32B basemodel	57.4	-
w/ VeRL	-	46.4
w/ Sync.AReaL	**61.2**	51.1
w/ AReaL (ours)	61.0	**31.1**

2x Convergence Speed **in Competitive Programming Tasks**



Both algorithm modifications are imperative **to async. RL performance**