HETRL: EFFICIENT REINFORCEMENT LEARNING FOR LLMS IN HETEROGENEOUS ENVIRONMENTS

Yongjun He [1 * †] **Shuai Zhang** [2 *] **Jiading Gai** [2] **Xiyuan Zhang** [2] **Boran Han** [2] **Bernie Wang** [2]
Huzefa Rangwala [2] **George Karypis** [2]

ABSTRACT

As large language models (LLMs) continue to scale and new GPUs are released even more frequently, there is an increasing demand for LLM post-training in heterogeneous environments to fully leverage underutilized mid-range or previous-generation GPUs and alleviate the shortage of homogeneous high-end GPUs within a single availability zone. However, achieving high-performance reinforcement learning (RL) training for LLMs on such computing resources remains challenging, as the workflow involves multiple models and tasks with complex computational and data dependencies. In this paper, we present HetRL, a distributed system for efficient RL training in infrastructures with heterogeneous GPUs and networks. HetRL formulates RL training scheduling in heterogeneous environments as a constrained joint optimization problem and provides two complementary approaches for addressing this problem: (1) a hybrid scheduling algorithm that efficiently identifies near-optimal solutions, and (2) an integer linear programming (ILP)-based scheduling algorithm that obtains optimal solutions, enabling flexible trade-offs between solution optimality and efficiency. Our extensive evaluation, consuming 20,000 GPU-hours, shows that HetRL achieves up to 9.17× the throughput of state-of-the-art systems, and 3.17× on average, across a wide range of workloads and settings.

1 INTRODUCTION

Reinforcement learning (RL) has become the predominant technique for improving the reasoning ability of large language models (LLMs) and aligning LLMs with human values (Lambert et al., 2024; Kaufmann et al., 2025; Ahmadian et al., 2024; Casper et al., 2023). Despite the leading performance brought by RL to LLMs, however, comes an explosive growth in computational demand (Qwen Team, 2025; DeepSeek-AI, 2024; Llama Team, 2024; Wu et al., 2025a). In practice, current deployments of RL training relies on individual clusters with a large number of homogeneous GPUs and high-bandwidth networks to meet the computational requirements, as state-of-the-art (SoTA) RL training systems (e.g., verl (Sheng et al., 2025) and OpenRLHF (Hu et al., 2025)) are tailored for homogeneous computing resources. On the other hand, as vendors have released an array of new GPU models in recent years, there are a substantial number of mid-range or previous-generation GPUs remaining underutilized across data centers around the world (Strati et al., 2024; Jiang et al., 2024a; Mei et al., 2025; Wu et al., 2025b; Gao et al., 2024). These geo-distributed heterogeneous GPUs collectively provide substantially more memory and compute resources than individual clusters of homogeneous GPUs. This motivates us to explore an alternative solution by *deploying RL training across a set of heterogeneous GPUs connected via heterogeneous networks*.

Recent studies (Yuan et al., 2022; Mei et al., 2025; Wu et al., 2025b; Jiang et al., 2024b; Um et al., 2024; Strati et al., 2025; 2024) have investigated the deployment of LLM training and serving in heterogeneous environments to improve the utilization of mid-range or previous-generation GPUs, all centered around the problem of allocating GPU resources to a single model or a single task. However, the complexity of the RL workflow presents unique obstacles to high-performance deployment in heterogeneous environments, which cannot be addressed by existing methods. Unlike LLM training and serving, which only involves one model, typical RL workflow (Schulman et al., 2017; Shao et al., 2024; Rafailov et al., 2023; Dai et al., 2024) consists of multiple models and tasks with complex dependencies. For instance, the most widely used RL algorithm, Proximal Policy Optimization (PPO) (Schulman et al., 2017) (Figure 1(b)), incorporates four LLMs: an *actor* model, a *critic* model, a *reward* model and a *reference* model; and six tasks: *actor generation, reference inference, critic inference, reward inference, actor training*, and *critic training*. Given

[*]Equal contribution [†]Work done during internship at AWS [1]ETH Zürich [2]Amazon Web Services. Correspondence to: Yongjun He <yongjun.he@inf.ethz.ch>, Shuai Zhang <shuaizs@amazon.com>.

Proceedings of the 9th MLSys Conference, Bellevue, WA, USA, 2026. Copyright 2026 by the author(s).

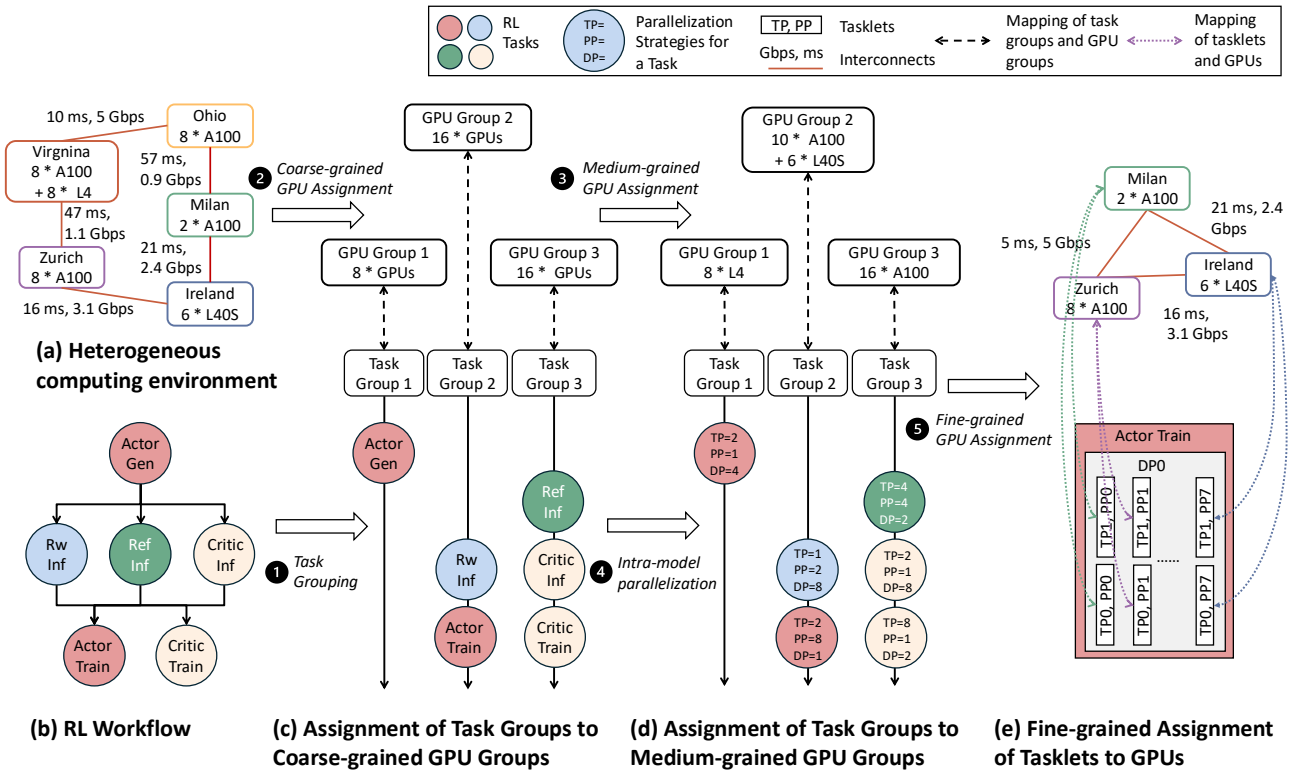

Figure 1. Given a heterogeneous computing environment (a) and an RL workflow (b), HetRL employs a multi-level search framework to progressively construct candidate execution plans through ❶ task grouping, ❷ coarse-grained GPU grouping, ❸ medium-grained assignment of task groups to specific GPUs, ❹ intra-model parallelization, and ❺ fine-grained mapping of tasklets to GPUs. It enables efficient exploration of near-optimal solutions across sub-problems, while jointly optimizing task/model colocation, parallelization strategies, and device placement under heterogeneous compute and network resources.

the heterogeneity of GPU models and interconnects, the dependencies between different models and tasks, and their different computational characteristics, the desired scheduling algorithm required for efficient RL training needs to jointly optimize (1) the colocation of models and parallelism between tasks (Figure 1(c)); (2) the parallelization of computations within each model and each task (Figure 1(d)); and (3) the fine-grained assignment of tasklets to the heterogeneous devices (Figure 1(e)).

Another category of recent studies has investigated efficient deployment of RL training, but their designs are tailored to clusters with homogeneous GPUs and high-bandwidth networks (Sheng et al., 2025; Zhong et al., 2025b; Fu et al.; Wu et al., 2025a; Han et al., 2025; Hu et al., 2025; Yao et al., 2023; Shen et al., 2024), and do not fully account for heterogeneity in the computing environment within their search space. An out-of-the-box solution for scheduling RL training in heterogeneous environments is to apply prior heterogeneity-aware scheduling algorithms for LLM training and serving to existing RL training systems. However, these methods (Yuan et al., 2022; Mei et al., 2025; Jiang et al., 2024b; Um et al., 2024) focus on the scheduling of a single model/task and require hundreds to thousands of seconds to search for the locally near-optimal execution plan for a single model/task (rather than the entire workflow). Considering that RL workflows involve multiple models and tasks with complex computational and data dependencies, directly applying prior methods to existing RL training systems is neither practical nor scalable.

To cope with the challenges described above, we propose HetRL, a distributed system for efficient RL training over a set of GPUs with heterogeneity in compute and memory resources as well as network interconnects. Our contributions are summarized as follows:

- We formulate RL training scheduling in heterogeneous environments as a constrained joint optimization problem and introduce a hybrid scheduling algorithm that efficiently identifies near-optimal solutions by (1) decomposing the complex search space with a multi-level search framework (Figure 1); (2) allocating the search budget using the successive halving algorithm (Jamieson & Tal-

walkar, 2016) (SHA); and (3) generating low-level plans with an evolutionary algorithm (EA). We also provide an integer linear programming (ILP)-based scheduling algorithm to obtain optimal solutions, enabling flexible trade-offs between solution optimality and efficiency.

- We implemented our proposed scheduling algorithms and built HetRL around it on top of verl (Sheng et al., 2025), including a scheduler, a profiler, and an execution engine with extended support for fine-grained resource assignment and load balancing.
- We conduct a comprehensive evaluation to compare the performance of HetRL and SoTA systems across various workloads and heterogeneous environments. The results show that HetRL attains throughput up to $9.17\times$ that of SoTA systems, and $3.17\times$ on average.

2 Background and Motivation

2.1 Reinforcement Learning for LLMs

Typical RL workflows (Schulman et al., 2017; Shao et al., 2024; Rafailov et al., 2023; Dai et al., 2024) involve multiple models and tasks with complex computational and data dependencies. Next, we detail the workflow using PPO (Schulman et al., 2017) (Figure 1(b)) as an example.

RL Models. The *actor* model is the main model to be trained with RL, initialized from pre-trained LLMs. The *reference* model is a frozen copy of the initial actor model. It's used to compute the Kullback-Leibler (KL) penalty that ensures the new actor model close to the style and knowledge of the pre-trained model. The *reward* model is a separate model trained from human preference data. It guides the actor model to generate responses that align with human values and is also frozen during RL training. The *critic* model evaluates the advantage values of the actions taken by the actor model. Its parameters are also updated during RL training. Another recently proposed representative RL algorithm, GRPO (Shao et al., 2024), accelerates RL training by eliminating the need for a separate critic model. The architectures and parameters of the reward and critic models can differ from those of the actor and reference models.

RL Tasks. At the beginning of the RL workflow, *actor generation* uses the prompts of the training dataset as input and generates responses using the actor model. Next, *reference inference* uses the prompts and generated responses as input and calculate the reference log probability of each token using the reference model. Using the same set of prompts and generated responses as input, *reward inference* calculate per-sample scores using the reward model, and *critic inference* calculates per-sample values using the critic model. Finally, *actor training* and *critic training* use the results from the three inference tasks as input to perform forward and backward passes to update the parameters of the actor model and the critic model, respectively.

Recent studies have also investigated asynchronous RL (Noukhovitch et al., 2025; Fu et al.; Han et al., 2025; Zhong et al., 2025a) to improve GPU utilization for RL training. Asynchronous RL training overlaps the time-consuming actor generation of the next few iterations with ongoing training to achieve speedup. However, it leads to decreased accuracy due to data staleness, and increases memory consumption due to the need to maintain two separate copies of the actor model for generation and training.

2.2 Parallelization in Distributed Deep Learning.

To distribute deep learning (DL) workloads over computing devices, three primary parallelization strategies have been proposed. Correctly combining and tuning them (Figure 1(e) bottom) can significantly improve LLM training and serving performance (Zheng et al., 2022; Li et al., 2023).

Data parallelism (DP) (Rajbhandari et al., 2020). Each DP group has a full copy of the model weights and processes a subset of the input dataset, and all DP groups periodically synchronize their model weights using averaged gradients.

Pipeline parallelism (PP) (Huang et al., 2019). The layers of a model are partitioned across PP groups. By splitting the training batch into multiple micro-batches, the forward and backward passes of different micro-batches can be pipelined across all PP group.

Tensor parallelism (TP) (Shoeybi et al., 2019). The weight matrices of each layer are partitioned across TP groups along the row or column dimension, and all TP groups perform all-reduce to aggregate the output of each partitioned GEMM.

2.3 Motivation

The heterogeneous characteristics of today's computing environments and RL workflows reveal opportunities to improve GPU utilization by deploying RL training in heterogeneous environments. Nevertheless, existing systems exhibit limitations that prevent them from realizing the full potential of such deployments.

2.3.1 Opportunities

Heterogeneity in computing environments. Recent studies (Strati et al., 2024; Jiang et al., 2024a; Mei et al., 2025; Wu et al., 2025b; Gao et al., 2024) investigates the availabilities of GPU resources, showing that there are severe shortages of homogeneous high-end GPUs within a single availability zone, while substantial heterogeneous GPUs are available across geographical locations. Therefore, training (Yuan et al., 2022; Wu et al., 2025b; Strati et al., 2025) and serving (Mei et al., 2025; Jiang et al., 2025; 2024b) LLMs in heterogeneous environments have attracted re-

search interests as it allows resource-intensive LLM jobs to leverage all available GPU resources for acceleration.

Heterogeneity in RL workflows. In RL workflows, the actor, critic, reference, and reward models may use LLMs with different model sizes and perform generation, inference, or training during different tasks. Therefore, RL workflows have different requirements for computation, memory, and communication across tasks. For example, the actor generation is memory bound (Zhang, 2024) and needs to maintain key-value cache (KV cache) (Kwon et al., 2023); and the actor/critic training are computation bound (He et al., 2025) and needs to maintain activations, gradients and optimizer states (Rajbhandari et al., 2020). In comparison with LLM training and serving which involves only single model or single task, the heterogeneous characteristics of the RL workflow make it a workload with new potential for utilizing otherwise idle heterogeneous GPU resources.

2.3.2 Limitations

Limited search space in existing RL training systems. To accelerate RL workflows with heterogeneous characteristics, and complex computational and data dependencies, a flurry of RL training systems (Sheng et al., 2025; Hu et al., 2025; Yao et al., 2023; Shen et al., 2024; Zhong et al., 2025b; Wu et al., 2025a; Fu et al.; Han et al., 2025) have been proposed. However, they are all tailored for homogeneous GPUs with high-bandwidth networks, and almost none of them fully account for the heterogeneity of hardware and networks in their search space. StreamRL (Zhong et al., 2025a) is perhaps the most relevant effort to ours, organizing heterogeneous GPUs into two groups: one group for actor generation and a separate group for the remaining tasks. However, it requires that all GPUs within the same group are homogeneous and located in the same data center.

Time-consuming heterogeneity-aware scheduling algorithms. A natural approach to enhance RL training systems is to apply the heterogeneity-aware scheduling algorithms originally designed for LLM training and serving to the scheduling of each model and each task in the RL workflow. As reported by verl (Sheng et al., 2025) and RLHFuse (Zhong et al., 2025b), searching for efficient deployment plans for RL training on homogeneous GPUs connected over a homogeneous network requires examining over millions to billions of plans, which takes hundreds to thousands of seconds. Meanwhile, the heterogeneity-aware scheduling algorithms originally designed for LLM training (Yuan et al., 2022; Um et al., 2024) and serving (Mei et al., 2025; Jiang et al., 2025; 2024b) require $1,000{\sim}10,000\times$ more search time to search for the locally near-optimal plan for a single model/task (rather than the entire workflow), so this naive combination is neither practical nor scalable for real-world deployment.

3 SCHEDULING IN HETRL

We begin by formulating RL training scheduling in heterogeneous environments as a constrained joint optimization problem over partitioning and assignment strategies. We then present our multi-level search framework for decomposing the search space, along with a cost model for efficient execution time estimation. Finally, we introduce two complementary scheduling algorithms: (1) a hybrid approach based on successive halving and evolutionary search that efficiently identifies near-optimal solutions, and (2) an ILP-based approach that provides optimal solutions when sufficient computational resources are available.

3.1 Problem Formulation

Notation. Let $\mathbf{G}^t = (V^t, E^t)$ denote the computational graph of the t-th task in an RL workflow, where $t \in \{1, \ldots, T\}$, V^t denotes the set of computational operators, and E^t denotes the set of tensors shared between operators. The overall computational graph of the RL workflow is then defined as $\mathbf{G} = (\bigcup_{t=1}^{T} V^t, \bigcup_{t=1}^{T} E^t \cup E^{\text{inter}})$, where E^{inter} denotes the set of edges between tasks.

Let $\mathbf{G_D} = (V_D, E_D)$ denote the device topology graph for a heterogeneous environment, where $V_D = \{d_1, \ldots, d_N\}$ are a set of N devices and $E_D \subseteq V_D \times V_D$ are communication channels between devices. Each device d is labeled with computation capability, memory capacity, and HBM bandwidth. Each edge between d and d' is labeled with the latency and bandwidth.

A partitioning strategy ρ transforms the given $\mathbf{G}$ into a new tasklet graph $\mathbf{G_L} = (V_L, E_L)$ by first partitioning the operators with intra-model parallelization and then merging them into tasklets. Each new node $l^t_{i,j,k} \in V_L$ is a tasklet, E_L denotes the set of tensors transferred between tasklets, and t, i, j, k represent the indices of a tasklet in RL tasks, data parallelism, pipeline parallelism, and tensor parallelism, respectively. An assignment strategy $\sigma : V_L \to V_D$ assigns, for each tasklet $l \in V_L$, a device $d \in V_D$. Let C denote a cost model that estimates the execution time per iteration of a given workflow, conditioned on the given resource, partitioning strategy, and assignment strategy. The functions $M_{\text{working}}(l)$ and $M_{\text{model}}(l)$ represent the working and model memory consumption of tasklet l, respectively, while $M_{\text{gpu}}(d)$ denotes the GPU memory capacity of device d.

We defer the notation for device and network attributes to Appendix B, notation for the PPO definition to Section 3.3, and notation for the scheduling algorithm to Algorithm 1.

Definition 1 (Heterogeneity-Aware RL Training Scheduling Problem)**.** Given a computational graph $\mathbf{G}$ for an RL workflow and a device topology graph $\mathbf{G_D}$ for a heterogeneous environment, the *heterogeneity-aware RL training scheduling problem* is to determine an optimal scheduling strategy,

which consists of a partitioning strategy ρ and an assignment strategy σ, such that the execution time of the RL workflow is minimized and resource constraints are satisfied:

$$\min_{\rho,\sigma} \quad C(\rho, \sigma; \mathbf{G}, \mathbf{G_D})$$

$$\text{s.t.} \quad |\{(i,j,k) : l^t_{i,j,k} \in V_L\}| \leq |V_D|, \forall t \in \{1, \ldots, T\} \quad \text{(C1)}$$

$$\bigcup_{d \in V_D} \sigma^{-1}(d) = V_L \quad \text{(C2)}$$

$$\max_{l \in \sigma^{-1}(d)} M_{\text{working}}(l) + \sum_{l \in \sigma^{-1}(d)} M_{\text{model}}(l) \leq M_{\text{gpu}}(d), \forall d \in V_D \quad \text{(C3)}$$

Proposition 1. *The abstract heterogeneity-aware RL training scheduling problem defined in Definition 1 is NP-hard.*

The proof is deferred to Appendix A.

3.2 Multi-Level Search Framework

Design. Given that the heterogeneity-aware RL training scheduling problem is NP-hard, exhaustively searching over all feasible execution plans is computationally intractable. We propose a multi-level search framework that decomposes the search space into structured subspaces, enabling termination of poor-performing high-level decisions and efficient exploration of lower-level decisions. As illustrated in Figure 1, the framework consists of the following levels:

- Level 1 (Task grouping): Given an RL training pipeline, we first partition tasks into disjoint task groups. Tasks within the same task group are executed on a shared set of GPUs, with their associated models co-located.
- Level 2 (Coarse-grained GPU assignment): Given the number of task groups, we partition the GPUs into disjoint GPU groups and assign each task group to one GPU group. At this step, we determine only the number of GPUs in each group, rather than the specific GPU assignments.
- Level 3 (Medium-grained GPU assignment): At this step, we generate candidate assignments that map task groups to specific GPUs.
- Level 4 (Intra-model parallelization): Given a set of candidate medium-grained GPU assignments, we determine feasible parallelization strategies for individual tasks within each assignment. Applying these strategies further decomposes tasks into finer-grained tasklets.
- Level 5 (Fine-grained GPU assignment): Finally, we generate candidate assignments that map tasklets to specific GPUs, yielding complete execution plans.

Existing RL training frameworks already expose several configurable dimensions that naturally correspond to different levels of the search space. In particular, task and model colocation (Level 1), placement groups and resource pools (Level 2), and parallelization strategies (Level 4) are commonly supported configuration knobs. To fully account for heterogeneity in the computing environment within the search space, we extend these dimensions with Level 2 and Level 5, which enables fine-grained tasklet-to-device mapping beyond what is typically supported in current systems.

Interpretation. Our framework can be viewed as a coarse-to-fine constructive approach that operationalizes the joint optimization over (ρ, σ) defined in Section 3.1. Concretely, Levels 1 and 4 instantiate the partitioning strategy ρ, performing task grouping and intra-model parallelization to produce the tasklet graph G_L, while Levels 2, 3, and 5 instantiate the assignment strategy σ, generating coarse-to-fine GPU assignments that map computational subgraphs (or nodes) to devices.

Search space analysis. The search space induced by the multi-level search framework can be characterized as a hierarchical composition of combinatorial subspaces across five levels. We analyze the search space in a level-wise manner:

- Level 1: The search space of this level for T tasks has size B_T, corresponding to the number of set partitions, where B_T denotes the T-th Bell number (Rota, 1964).
- Level 2: For simplicity, we consider the case where each task forms its own GPU group, which yields a worst-case upper bound. The search space of this level for N GPUs has size $\binom{N-1}{T-1}$, corresponding to the number of integer partitions of N into exactly T positive parts.
- Level 3: Given a coarse-grained GPU assignment with GPU group sizes $\{n_t\}_{t=1}^T$, where $\sum_{t=1}^T n_t = N$, the search space of this level has size $\frac{N!}{\prod_{t=1}^T n_t!}$, corresponding to the multinomial coefficient.
- Level 4: For simplicity, we assume that each of data, pipeline, and tensor parallelism has a uniform degree. Therefore, the search space of this level is upper bounded by $\prod_{t=1}^T |\{(i,j,k) \in \mathbb{N}_+^3 : i \cdot j \cdot k \leq n_t\}|$.
- Level 5: Given the parallelization strategies, each task is further decomposed into tasklets. Let $|V_L^t|$ denote the number of tasklets assigned to task group t. The search space of this level is upper bounded by $\prod_{t=1}^T n_t^{|V_L^t|}$.

Note that the search space at each level listed above is conditional on the decisions made in preceding levels, rather than independent. Combining the above levels, the overall search space is upper bounded by

$$B_T \cdot \binom{N-1}{T-1} \cdot \frac{N!}{\prod_{t=1}^T n_t!} \cdot \prod_{t=1}^T |\{(i,j,k) \in \mathbb{N}_+^3 : i \cdot j \cdot k \leq n_t\}| \cdot \prod_{t=1}^T n_t^{|V_L^t|}.$$

3.3 Cost Model

Because the search space is large and running RL training in practice can take tens of minutes per step, we develop a cost model within our framework to quickly estimate execution

time. Cost models differ across RL algorithms and between synchronous and asynchronous modes. To instantiate the cost model, we formalize a synchronous variant of PPO and connect it to the computational graph defined in Section 3.1.

PPO. We consider a PPO setting with a policy (actor) model π_θ, a value function (critic model) v_ϕ, a reward model r_ψ, and a fixed reference policy (model) π_{ref}. Given queries x sampled from the training dataset $\mathcal{D}$ and responses y generated by the old policy model $\pi_{\theta_{\text{old}}}$, the reward at the τ-th generation step is defined as:

$$r_\tau = r_\psi(x, y_{\leq\tau}) - \beta \log \frac{\pi_\theta(y_\tau \mid x, y_{<\tau})}{\pi_{\text{ref}}(y_\tau \mid x, y_{<\tau})},$$

where θ, ϕ, and ψ denote the parameters of the actor, critic, and reward models, respectively, and β is the coefficient of the KL penalty. Based on the rewards and the learned value function, the estimated advantage $\hat{A}_\tau$ is then computed using Generalized Advantage Estimation (GAE) (Schulman et al., 2016). Finally, PPO updates the policy model via a clipped surrogate objective:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \Big[\frac{1}{|y|} \sum_{1}^{|y|} \min \Big(\frac{\pi_\theta(y_\tau \mid x, y_{<\tau}))}{\pi_{\theta_{\text{old}}}(y_\tau \mid x, y_{<\tau}))} \hat{A}_\tau, \text{clip}\Big(\frac{\pi_\theta(y_\tau \mid x, y_{<\tau}))}{\pi_{\theta_{\text{old}}}(y_\tau \mid x, y_{<\tau}))}, 1-\epsilon, 1+\epsilon \Big) \hat{A}_\tau \Big) \Big],$$

where ϵ controls the clipping range, and updates the value function via a mean-squared error loss.

From the above formulation of PPO, we observe that the training process can be decomposed into multiple computation stages with well-defined dependencies. Additionally, a synchronous variant of PPO enforces an iteration-level barrier: the actor generation of the next iteration cannot begin until the training of the current iteration completes. We therefore model synchronous PPO as a computational graph $\mathbf{G}$ formed by a set of task-level computational graphs $\{\mathbf{G}^t\}_{t=1}^6$, where actor generation ($\mathbf{G}^1$) is followed by reward, reference, and critic inference ($\mathbf{G}^2$–$\mathbf{G}^4$) in parallel, before actor and critic training ($\mathbf{G}^5$ and $\mathbf{G}^6$).

Cost model for synchronous PPO. We adopt a compact convention and overload the notation: C denotes both the cost model and its evaluated cost under the current inputs. We now present a concrete instantiation of C for PPO:

$$C_{\text{SyncPPO}} = C^1 + \Phi(\{C^2, C^3, C^4\}) + \Phi(\{C^5, C^6\}),$$

where C^t denotes the cost associated with the t-th task ($\mathbf{G}^t$) for $t \in \{1, \ldots, 6\}$, corresponding to actor generation, reward inference, reference inference, critic inference, actor training, and critic training, respectively. Φ aggregates the costs of tasks without dependencies and is defined as:

$$\Phi(\{C^t\}) = \eta \max_t C^t + (1-\eta) \sum_t C^t, \eta \in [0, 1],$$

where the coefficient η parameterizes the level of task parallelism (0: sequential; 1: fully parallel; else: partial). The cost model of the t-th task is given by:

$$C^t = \begin{cases} \Psi^{\text{gen}}(C^t_{\text{comp}}, C^t_{\text{hbm}}, C^t_{\text{tp}}, C^t_{\text{pp}}), & t = 1, \\ \Psi^{\text{inf}}(C^t_{\text{comp}}, C^t_{\text{tp}}, C^t_{\text{pp}}), & t \in \{2, 3, 4\}, \\ \Psi^{\text{train}}(C^t_{\text{comp}}, C^t_{\text{tp}}, C^t_{\text{pp}}, C^t_{\text{dp}}, C^t_{\text{bubble}}), & t \in \{5, 6\}, \end{cases}$$

where C^t_{comp} estimates a set of per-tasklet or per-subgraph computation costs, C^t_{hbm} estimates a set of per-tasklet or per-subgraph HBM costs (i.e., loading LLMs from HBM to SRAM during decoding), C^t_{bubble} estimate a set of per-tasklet or per-subgraph pipeline-bubble costs, and C^t_{tp}, C^t_{pp}, and C^t_{dp} estimate a set of per-tasklet or per-subgraph TP, PP, and DP computation costs. Ψ^{gen}, Ψ^{inf}, and Ψ^{train} estimate task-level costs based on the given sets of tasklet- or subgraph-level costs. For brevity, we have omitted some inputs to the above equations, including device and network attributes, and also omitted the resharding cost in the end-to-end cost. Details can be found in Appendix B.

3.4 Hybrid Scheduling Algorithm

SHA for search budget allocation. Although theoretically identifying the optimal high-level decision is intractable due to the NP-hardness of the problem, we aim to prioritize promising high-level decisions in a principled manner while keeping the method scalable and simple from an engineering perspective. To this end, we adopt successive halving algorithm (Jamieson & Talwalkar, 2016) (SHA) for search budget allocation, a robust and straightforward bandit-based algorithm designed to the non-stochastic best arm identification problem within a fixed budget. Next, we determine at which levels to apply SHA. In the bandit formulation, each high-level decision corresponds to an arm. Applying SHA only to Level 1 results in too few arms, yielding limited speedup, as the number of elimination rounds is bounded by $\lceil \log_2 B_T \rceil$. In contrast, applying SHA to Levels 1–3 introduces an excessive number of arms due to the combinatorial explosion of the search space, making it infeasible to evaluate them within a fixed budget. Therefore, we apply SHA to Levels 1 and 2, which provide a balanced number of arms for effective speedups while keeping the budget allocation tractable.

The pseudocode in Algorithm 1 shows how HetRL extends SHA to a nested form and applies it to our problem. HetRL treats task groupings as arms at Level 1 and GPU groupings as arms at Level 2 in the multi-armed bandits problem; the execution time estimated by the cost model serves as each arm's loss. Concretely, the inputs are the user-defined budget B, workflow information and device information. In our problem, the search budget corresponds to the wall-clock time allocated for the scheduling procedure. At Level 1 (lines 14-16 and lines 29-33), it first assigns a starting

Algorithm 1 HetRL Hybrid Scheduling Algorithm

Input: search budget B, a computational graph for an RL workflow $\mathbf{G}$, a device topology graph $\mathbf{G_D}$, number of GPUs N
Output: Execution plan with the lowest estimated cost

1: let $TG = \{tg_{\xi_i}\}$ be the set of all feasible task groupings for a given RL workflow, where tg_{ξ_i} denotes a feasible grouping of RL tasks
2: let $GG = \{tg_{\xi_i} \mapsto GG_{\xi_i}\}$ be the set of all feasible GPU groupings, where $GG_{\xi_i} = \{gg_{\xi_{i,j}}\}$ denotes the set of all feasible GPU groupings for a given tg_{ξ_i} and $gg_{\xi_{i,j}}$ denotes a feasible grouping of GPUs
3: let $C_{\text{plans}} = \{tg_{\xi_i}, gg_{\xi_{i,j}} \mapsto \{c_{\xi_{i,j,k}}\}\}$ be the cost of candidate plans, where $c_{\xi_{i,j,k}}$ denote the cost of the k-th candidate plans under tg_{ξ_i} and $gg_{\xi_{i,j}}$
4: initialize $TG \leftarrow \emptyset, GG \leftarrow \emptyset, C \leftarrow \emptyset$
5: $TG \leftarrow \text{TaskGrouping}(\mathbf{G})$
6: **for all** $tg_{\xi_i} \in TG$ **do**
7: $\quad GG_{\xi_i} \leftarrow \text{GPUGrouping}(N, |tg_{\xi_i}|)$
8: $\quad GG \leftarrow GG \cup \{tg_{\xi_i} \mapsto GG_{\xi_i}\}$
9: $\quad$ **for all** $gg_{\xi_{i,j}} \in GG_{\xi_i}$ **do**
10: $\quad\quad C_{\text{plans}} \leftarrow C_{\text{plans}} \cup \{tg_{\xi_i}, gg_{\xi_{i,j}} \mapsto \{\infty\}\}$
11: $\quad$ **end for**
12: **end for**
13: $(TG_0, GG_0) \leftarrow (TG, GG)$
14: **for** $m = 0, 1, \ldots, \lceil \log_2(|TG|) \rceil - 1$ **do**
15: $\quad$ **for all** $tg_{\xi_i} \in TG_m$ **do**
16: $\quad\quad$ let $b_m = \lfloor \frac{B}{|TG_m| \lceil \log_2(|TG|) \rceil} \rfloor$ be search budget for tg_{ξ_i}
17: $\quad\quad GG_{\xi_i,0} \leftarrow GG_m(tg_{\xi_i})$
18: $\quad\quad$ **for** $n = 0, 1, \ldots, \lceil \log_2(|GG_m(tg_{\xi_i})|) \rceil - 1$ **do**
19: $\quad\quad\quad$ **for all** $gg_{\xi_{i,j}} \in GG_{\xi_i,n}$ **do**
20: $\quad\quad\quad\quad$ let $b_{m,n} = \lfloor \frac{b_m}{|GG_{\xi_i,n}| \lceil \log_2(|GG_m(tg_{\xi_i})|) \rceil} \rfloor$ be search budget for $gg_{\xi_{i,j}}$
21: $\quad\quad\quad\quad$ **for** $k = 0, 1, \ldots, b_{m,n} - 1$ **do**
22: $\quad\quad\quad\quad\quad$ let $p_k \leftarrow \text{EA}(tg_{\xi_i}, gg_{\xi_{i,j}}, \mathbf{G_D}, \ldots)$ be a new low-level plan
23: $\quad\quad\quad\quad\quad c_{\xi_{i,j,k}} = C(p_k, \mathbf{G_D}, \ldots)$
24: $\quad\quad\quad\quad\quad C_{\text{plans}}(tg_{\xi_i}, gg_{\xi_{i,j}}) \leftarrow C_{\text{plans}}(tg_{\xi_i}, gg_{\xi_{i,j}}) \cup c_{\xi_{i,j,k}}$
25: $\quad\quad\quad\quad$ **end for**
26: $\quad\quad\quad$ **end for**
27: $\quad\quad\quad GG_{\xi_i,n+1} \leftarrow \text{BestHalf}(GG_{\xi_i,n}, C_{\text{plans}})$
28: $\quad\quad$ **end for**
29: $\quad\quad GG_{m+1}(tg_{\xi_i}) \leftarrow \text{BestHalf}(GG_m(tg_{\xi_i}), C_{\text{plans}})$
30: $\quad$ **end for**
31: $\quad TG_{m+1} \leftarrow \text{BestHalf}(TG_m, C_{\text{plans}})$
32: **end for**

budget b_m to each task grouping, which is shared by all the GPU groupings corresponding to this task grouping. At Level 2 (lines 17-20 and lines 27-29), it also first assigns a starting budget $b_{m,n}$ to each pair of tasking grouping and GPU grouping. It then uses an EA at lower levels and a cost model to generate and evaluate $b_{m,n}$ candidate plans (lines 21–25). Finally, it discards the worst half GPU groupings and continues the procedure with the better half with a doubled budget until the assigned budget b_m is exhausted. This procedure is also repeated at Level 1 by discarding the worse half tasking groupings and doubling the budget for the next round until the global budget B is exhausted. At each new Level 1 round, we retain the best half of GPU groupings for each task grouping. With nested SHA, we allocate more search budget to sets of candidates associated with more promising high-level decisions, while discarding less promising candidates early. SHA enjoys theoretical guarantees on the probability of selecting an optimal (or near-optimal) configuration within a given budget (Jamieson & Talwalkar, 2016; Li et al., 2017).

EA for low-level plan generation. Given the task grouping from Level 1 and the coarse-grained GPU assignment from Level 2, we generate fine-grained GPU assignments through Level 3 to 5. By treating the devices assigned to tasklets within the same pipeline stage as a graph partition, and those assigned to the same task group as a coarsened graph partition, the procedure can be viewed as a graph partitioning problem with a complex objective on the device topology graph. Following the line of research that uses evolutionary algorithm (EA) (Bui & Moon, 1996; Soper et al., 2004; Yuan et al., 2022) for graph partitioning, we develop a EA for low-level plan generation.

Concretely, we randomly initialize medium-grained GPU assignments at Level 3, enumerate all feasible intra-model parallelization strategies at Level 4, and then randomly initialize fine-grained GPU assignments at Level 5. These fine-grained GPU assignments serve as the initial population of the EA, which produces the next generation as follows. It first generates a new "offspring" o via mutation from the population. We than conduct swapping on o to find a new individual o^* that leads to lower cost. Finally, we add o^* to the population and remove the worst individual in the population if o^* achieves a lower cost.

As described in Section 3.2, the multi-level search framework consists of five levels. While omitting Level 3 does not affect the completeness of the search space, it is introduced to provide a group-level structure that enables GPU mutation and swapping across task/GPU groups, thereby improving the efficiency of the search process. We customize the mutation operator and incorporate a swap-based local search. The mutation operator, with a certain probability, replaces a GPU in a training-task group with a higher-TFLOPS one selected from GPUs not assigned to any training-task group. The local search greedily improves GPU-group locality while keeping group sizes fixed. For each candidate, we repeatedly evaluate cross-group GPU swaps and apply the one with the largest positive gain in a locality score that captures, but is not limited to, machine-, zone-, and region-level affinities. The process terminates when no improving swap exists, and the improvements obtained by the phenotype are not mapped back to the genotype or incorporated into the population. This design allows faster evolution while preserves population diversity and prevents premature convergence (Hinton & Nowlan, 1987; Baldwin, 1896).

3.5 ILP-Based Scheduling Algorithm

While the hybrid scheduling algorithm can efficiently identify near-optimal solutions in large-scale deployments, we further provide an ILP-based scheduling algorithm to obtain optimal solutions when sufficient search budgets are available. Our problem formulation (Section 3.1) can be

straightforwardly converted into a corresponding ILP formulation by converting the discrete scheduling choices into binary decision variables. The ILP formulation preserves the same search space as our hybrid scheduling algorithm, but replaces heuristic exploration with exact optimization.

Specifically, for each RL task, we enumerate all feasible parallelization strategies over candidate data, pipeline, and tensor parallelism, and associate each strategy with a binary decision variable. Based on the selected strategy, we further introduce binary variables to determine whether each tasklet is instantiated and to assign each instantiated tasklet to a specific GPU. On top of these decision variables, we use the analytical cost model to parameterize the execution cost of each task under heterogeneous devices. In particular, the formulation captures the memory footprint and execution time of each tasklet. These tasklet-level costs are then aggregated into task-level durations according to the selected partitioning and assignment strategies. To model end-to-end execution, we additionally introduce time variables for each task, including start time, duration, and completion time, and minimize the overall workflow makespan. The formulation not only enforces the constraints defined in Section 3.1 but also incorporates the task dependencies as constraints.

4 HetRL System

Centering on our proposed scheduling algorithm, we built HetRL, a distributed system for efficient RL training in heterogeneous environments. Figure 2 illustrates the system overview of HetRL, which comprises four key components: a scheduler, a profiler, a load balancer, and a distributed execution engine. Next, we describe an overview of the system's end-to-end execution and the role of each component.

4.1 Execution Overview

An RL training job submitted to HetRL contains an RL algorithm, a dataset, models for different tasks, an optimizer, numerical precision, global batch size, sequence lengths of prompts and responses, and other optional configurations. After receiving a job request, the HetRL profiler first collects hardware information about the computing environment, including the computation power (TFLOPs), memory capacity (GBs), and HBM bandwidth (GB/s) of available GPUs, intra-machine bandwidth (GB/s), and network delay (ms) and bandwidth (Gbps) between them. Then the HetRL scheduler searches for a near-optimal execution plan according to the workflow information and hardware information. As described in Section 3, the scheduler leverages the multi-level search framework and our proposed scheduling algorithms for generation and selection of candidate plans, and uses its cost model to evaluate various candidate plans in terms of training throughput and memory footprint. Finally, the HetRL execution engine deploys the RL training job in the heterogeneous environment according to the selected execution plan. The HetRL execution engine is implemented on top of verl (Sheng et al., 2025) with extended support for fine-grained resource assignment and load balancing, and it uses Megatron-LM (Shoeybi et al., 2019) and vLLM (Kwon et al., 2023) as training and serving engines.

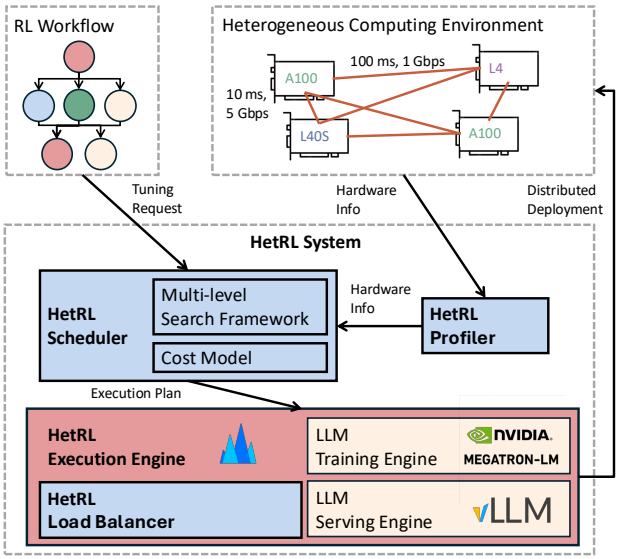

Figure 2. HetRL system overview.

4.2 Load Balancing

The HetRL load balancer incorporates several load balancing strategies to better accommodate RL workflows to heterogeneous computing environments. Load balancing can be roughly divided into two categories: data-level and layer-level. To achieve data-level load balancing, it adjusts the local batch sizes across GPUs within a DP group during the actor rollout task based on estimates from the cost model. For the other tasks, where sequence lengths are known beforehand, it assign samples with longer sequence length to more powerful GPUs. To achieve layer-level load balancing, it adjusts the layer distribution across pipeline stages based on estimates from the cost model. These three load balancing strategies enlarge the search space of our scheduling algorithms and can be implemented without invasive modifications to mainstream frameworks (e.g., verl, Megatron-LM and vLLM). We leave the integration of more advanced load balancing strategies (Um et al., 2024) as future work.

5 Evaluation

We first compare the end-to-end performance of HetRL against SoTA RL training systems under various workloads and computing environments. Subsequently, we evaluate

Table 1. GPU specifications.

Model	Arch	Size (GB)	FP16 Perf. (TFLOPS)	HBM (GB/s)	Intra (GB/s)
A100	Ampere	40	312	2039	600
L40S	Ada	48	366	864	64
L4	Ada	24	121	300	64

the search efficiency and the effectiveness of load balancing.

5.1 Experimental Setup

Hardware. We conduct experiments on a testbed with a total of 64 GPUs, of which 24 are A100s, 24 are L40Ss, and 16 are L4s. The detail GPU specifications are listed in Table 1. To simulate the heterogeneous network environments, we measure the latencies and bandwidths between 10 different regions (Virginia, Ohio, Paris, Stockholm, London, Ireland, Spain, Zurich, Frankfurt, and Milan) and apply them to our testbed, as shown in Figure 3 (a) and (b). Our evaluation covers four different network environments:

- *Scenario 1 (Single-Region).* This is a standard setting provided by cloud service providers. We do not enforce latency or bandwidth controls here.
- *Scenario 2 (Multi-Region-Hybrid).* We consider individual GPUs across Ohio and Virginia, but subset of Virginia GPUs are at the edge and only have direct connections to GPUs within Virginia. The inter-region connections between Ohio and Virginia have 10 ms of delay and 5 Gbps of bandwidth, while connections involving edge GPUs have a bandwidth of 1 Gbps.
- *Scenario 3 (Multi-Country).* We consider individual GPUs cross eight different regions in Europe. For inter-region connections, the delay is $5\sim30$ ms and the bandwidth is $1.9\sim5.0$ Gbps.
- *Scenario 3 (Multi-Continent).* We consider individual GPUs cross eight different regions acroos Europe and US. For inter-region connections, the delay is $5\sim60$ ms and the bandwidth is $0.9\sim5.0$ Gbps.

Models and RL algorithms. We choose representative open-source LLMs, the Qwen series, and consider three different sizes of Qwen models, 4B, 8B, and 14B. For the RL algorithm, we evaluate two popular methods, PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024), considering both their synchronous and asynchronous versions. In the evaluation, we use the same size of LLM for all models and tasks within each RL workflow. However, HetRL can flexibly accommodate RL workflows that use models of different sizes to perform different tasks.

Datasets and hyperparameters. We conduct experiments on GSM8k dataset from OpenAI, an open-source, gold-standard benchmark for mathematical reasoning. We set the maximum length of both input prompts and output responses to 1024, the global batch size to 384, and the number of responses generated per prompt to 8. In the RL workflow, training uses mixed precision with the Adam optimizer, and inference and generation use BF16 precision. The above settings are consistent with previous studies on RL training systems (Sheng et al., 2025).

Baselines. We compare HetRL against two baselines:

- verl (Sheng et al., 2025) is a widely-used, open-source, SoTA RL training system in homogeneous setting. It proposes a hierarchical hybrid programming model to flexibly support various RL workflows and configurations.
- StreamRL (Zhong et al., 2025a) is the SoTA RL training system in heterogeneous and asynchronous setting. It supports the deployment of actor generation and the rest tasks in two separate data centers.

Since StreamRL is not open-source, we implemented its asynchronous version on top of verl. In addition, we chose vLLM as the inference engine and Megatron-LM as the training engine for all variants to ensure a fair comparison. We recognize other well-known RL training systems; however, we do not include them in the evaluation since their features and functionalities largely overlap with the baselines or they focus on RL algorithmic innovations.

Unless otherwise specified, all experiments are conducted under the experimental setup described above. For search efficiency comparisons, we use different hardware settings (i.e., CPU-only machines), while for training quality evaluations, we use alternative models and datasets, as detailed in the corresponding sections.

5.2 End-to-End Comparison

Our first experiment compares end to end performance of HetRL against verl and StreamRL in four different scenarios, corresponding to four rows of Figure 3. Figure 3(c - e) show the throughput of training Qwen-4B, 8B, and 14B with PPO and GRPO, respectively. In Single-Region scenario, HetRL outperforms verl by $1.51\sim2.05\times$ in synchronous RL training and outperforms StreamRL and verl by $1.1\sim1.31\times$ in asynchronous RL training. In Multi-Region-Hybrid scenario, HetRL outperforms verl by $3.01\sim4.99\times$ in synchronous RL training and outperforms StreamRL and verl by $1.11\sim1.27\times$ and $4.07\sim9.17\times$ in asynchronous RL training. In Multi-Country scenario, HetRL outperforms verl by $1.4\sim3.07\times$ in synchronous RL training and outperforms StreamRL and verl by $1.19\sim1.5\times$ and $1.71\sim4.0\times$ in asynchronous RL training. In Multi-Continent scenario, HetRL outperforms verl by $2.24\sim5.46\times$ in synchronous RL training and outperforms StreamRL and verl by $2.25\sim3.72\times$ and $4.38\sim10.76\times$ in asynchronous RL training.

HetRL achieves higher training throughputs than baselines

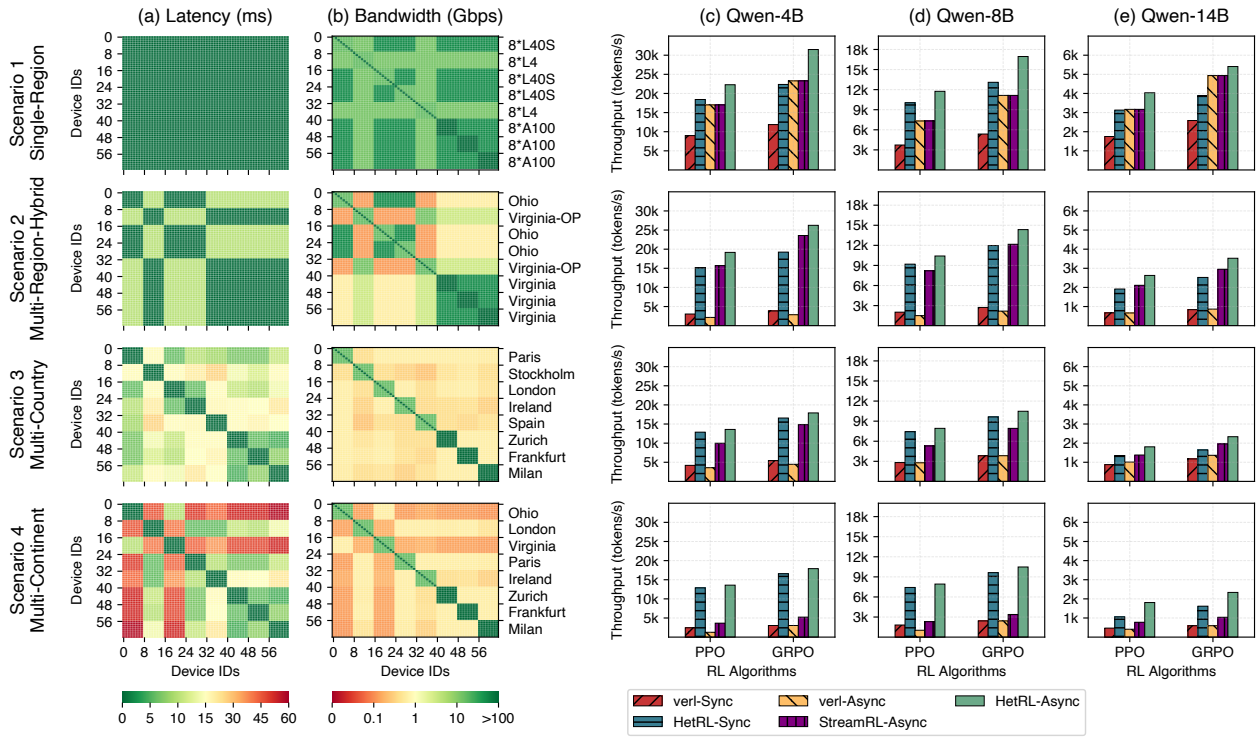

Figure 3. End to end compassion of HetRL with verl and StreamRL in four different scenarios. Column (a) and (b) visualize the delay and bandwidth of four scenario respectively; Column (c), (d), and (e) illustrate the PPO and GRPO throughput comparison respectively.

in all scenarios by fully considering the characteristics of heterogeneous GPUs and networks in its scheduling algorithm. The performance gaps in the scenarios 2-4 are larger due to the larger differences in network latencies and bandwidths between devices. While HetRL-Async is always faster than HetRL-Sync, we observe that verl-Async is sometimes slower than verl-Sync due to its unoptimized scheduling in heterogeneous environments. Except for the first scenario, StreamRL-Async achieves higher training throughput than verl-Async because its scheduling algorithm allows dividing GPU resources across data centers into two groups and assigning them to actor generation and the other tasks accordingly. The performance gaps between HetRL and the baselines for PPO and GRPO are different, mainly because GRPO does not have a critic model and critic inference and training tasks.

5.3 Effectiveness of Load Balancing

This section conducts an ablation study on the effect of load balancing. We evaluate the synchronous RL training of Qwen 4B, 8B, and 14B using PPO and GRPO. As illustrated in Figure 4, load balancing improves training throughput by up to 12% in the Single-Region scenario and up to 18% in the Cross-Region scenario, which confirms the effectiveness of our load balancing strategies. Our performance gains from load balancing is slightly smaller than related work such as Metis (Um et al., 2024), which is 19∼22%. The gap can be minimized by integrating more proposed load balancing strategies into HetRL. In asynchronous RL training, further improvements are not significant because

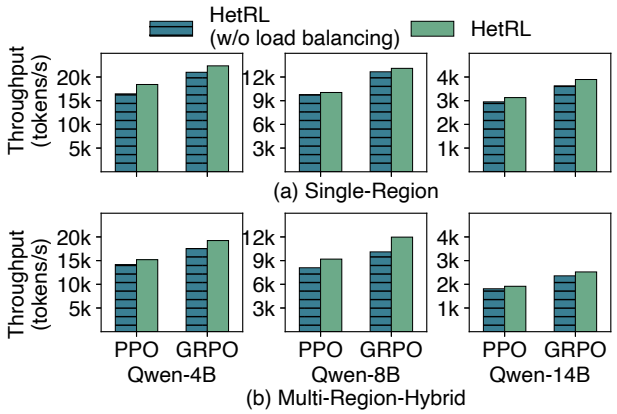

Figure 4. Effects of load balancing on synchronous RL training across model sizes under Single- and Multi-Region scenarios.

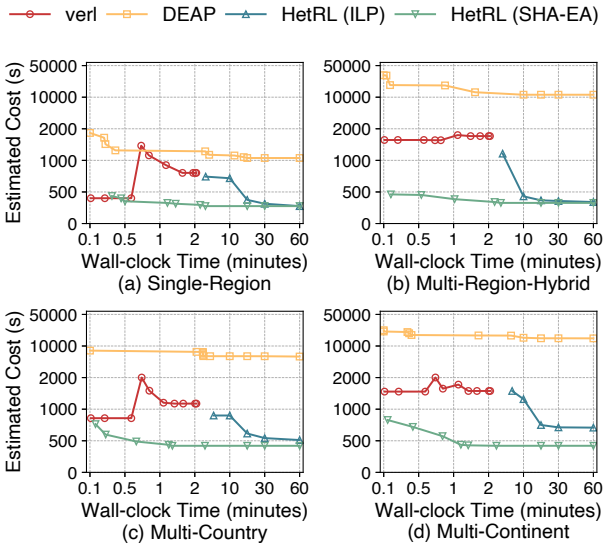

Figure 5. Search efficiency comparison across scheduling algorithms for training Qwen-8B with synchronous PPO.

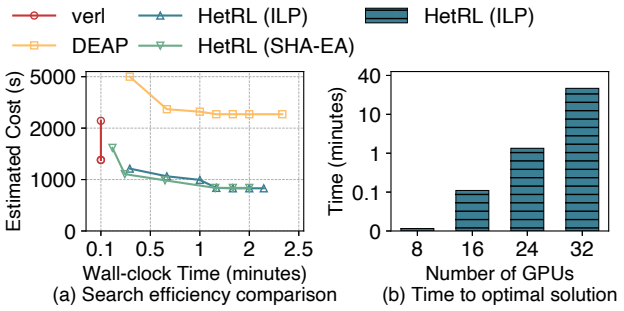

Figure 6. Small-scale settings: (a) search efficiency comparison across scheduling algorithms for infrastructures with 24 GPUs; (b) time to optimal solution of HetRL (ILP) across different numbers of GPUs.

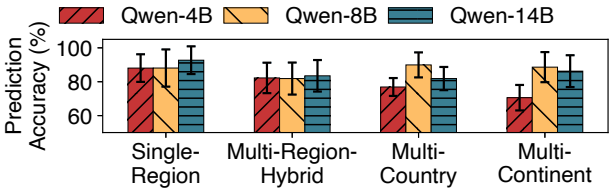

Figure 7. Prediction accuracy of the HetRL cost model for iteration time across different Qwen model sizes under multiple scenarios. Error bars denote standard deviation.

the resource scheduling between generation and training determines the performance.

5.4 Effectiveness of the Scheduling Algorithm

To evaluate the effectiveness of our proposed scheduling algorithms, we compare their search efficiency against verl's scheduling algorithm and a standard EA implemented using DEAP (Fortin et al., 2012) for this problem. The experiments in this section are conducted on a dual-socket server equipped with two 16-core AMD EPYC 9135 CPUs clocked at 3.65GHz, where either multi-threading or multi-processing is employed depending on the experimental variant. Figure 5 shows that our proposed scheduling algorithms significantly outperform verl's scheduling algorithm and DEAP. When HetRL (SHA-EA) and HetRL (ILP) are given sufficient search budget (i.e., wall-clock time), both approaches outperform verl due to their heterogeneity-aware cost models, and outperform DEAP due to our proposed design improvements. However, when the search budget is limited, the plans obtained by HetRL (ILP) are worse than those generated by verl, as the ILP-based scheduling algorithm relies on exact solvers that can obtain optimal solutions but incur high computational overhead. In comparison, HetRL (SHA-EA) can efficiently identify near-optimal plans that outperform other variants under different search budgets and achieve comparable quality to those obtained by HetRL (ILP) with one hour of search. For small-scale settings (e.g., no more than 24 GPUs), HetRL (ILP) can identify optimal solutions in less than three minutes, as shown in Figure 6. Meanwhile, the performance gaps between the solutions obtained by HetRL (SHA-EA) and the optimal solutions obtained by HetRL (ILP) are within 1%. These results highlight the complementary strengths of the two approaches, enabling practical deployment across diverse settings by leveraging HetRL (SHA-EA) for efficient search of near-optimal solutions at scale and HetRL (ILP) for exact optimal solutions.

5.5 Validation of the Cost Model

We now turn to the validation of our analytical cost model, illustrated in Figure 7. In the Single-Region scenario, the iteration time prediction error of our cost model for RL workloads is comparable to that of existing estimators for pre-training workloads. However, in Multi-Region, Multi-Country, and Multi-Continent scenarios, the prediction error is slightly higher due to the increased complexity of RL workloads. In addition, the current implementation of RL weight updates in HetRL, directly derived from verl, is not highly optimized and may be 100× slower (Licker et al., 2026) than the performance upper bound estimated by analytical cost models, further affecting prediction accuracy. Recent SoTA implementations (Licker et al., 2026) reduce this gap by leveraging point-to-point communication.

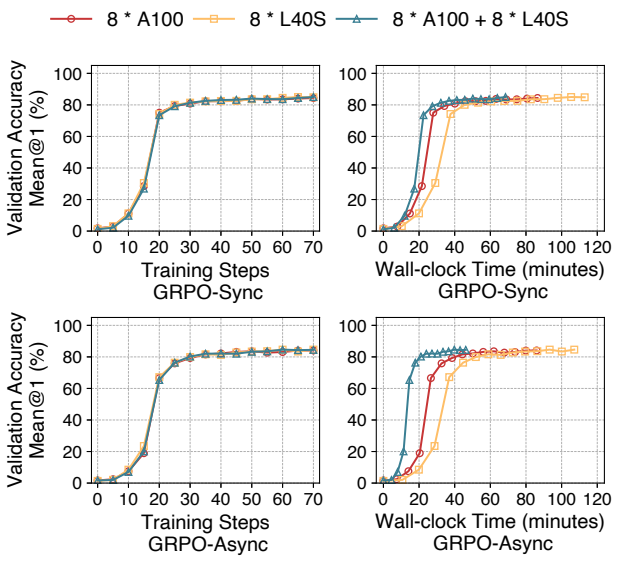

Figure 8. Training dynamics of Qwen3-1.7B-Base on GSM8K with GRPO.

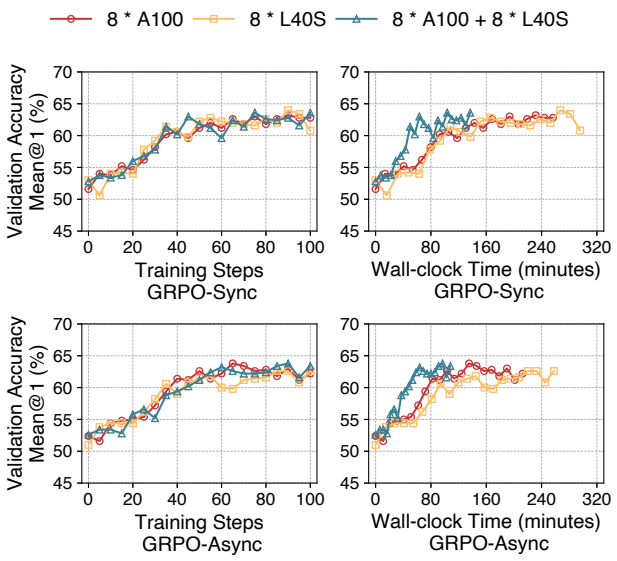

Figure 9. Training dynamics of Qwen3-1.7B-Base on MATH-500 with GRPO.

5.6 Impact of GPU Heterogeneity

Training quality. Now we examine examine the impact of GPU heterogeneity on training quality. We additionally use Qwen3-1.7B-Base and the MATH-500 dataset, and the rest experimental setup remains consistent with those described in Section 5.1. Although potential precision issues may arise when exchanging data across heterogeneous GPUs, we observe that their impact is negligible for both GRPO-Sync and GRPO-Async across different datasets, as shown in Figures 8 and 9. For GRPO-Sync, data exchange across heterogeneous GPUs occurs primarily within each stage, whereas for GRPO-Async, it primarily occurs between stages (i.e., actor training and actor rollout). When comparing by training steps, using heterogeneous GPUs achieves comparable final validation accuracy. When comparing by wall-clock time, aggregating the computational and memory capacities of heterogeneous GPUs enables significantly faster convergence.

Training throughput. Our final set of experiments evaluates how different systems perform under varying combinations of heterogeneous GPUs. HetRL outperforms verl by $1.72\sim4.33\times/1.77\sim3.42\times/1.57\sim3.33\times/1.57\sim3.03\times$ under PPO-Sync/GRPO-Sync/PPO-Async/GRPO-Async, as shown in Figure 10. When leveraging the larger memory capacity and more compute resources of heterogeneous GPUs (Figure 10(ALL GPUs)), HetRL outperforms its deployments on limited homogeneous GPUs (Figure 10(24*A100)) by $1.57\sim2.0\times$. Due to resource constraints, this set of experiments is limited to the Single-Region scenario. However, the analysis results obtained from our cost model suggest that the observed improvements can be extended to the rest three scenarios.

Figure 3(bottom three rows) and Figure 10(24×A100) show that, by leveraging heterogeneous compute resources across regions, HetRL achieves $1.09\sim1.77\times$ the throughput of using only the limited homogeneous GPUs in a single region. These results show that (1) HetRL effectively harnesses heterogeneous compute resources to accelerate RL training; and (2) when available homogeneous GPUs are limited, using heterogeneous resources from the same region or even across regions is a viable alternative.

6 Limitations and Discussion

GPU heterogeneity. Our evaluation only uses three different NVIDIA GPUs (Table 1) and relies on AWS OFI NCCL (aws, 2025b) and AWS EFA (aws, 2025a) for inter-node communication. HetRL has not yet been tested on other NVIDIA GPU generations, GPUs from other vendors, or other networking stacks. We adopt mainstream RL algorithms without modification and focus our evaluation on training throughput. For training quality, we use a representative RL algorithm (GRPO), an open-source model (Qwen3-1.7B-Base), and datasets (GSM8K and MATH-500) for case studies. A comprehensive investigation of the impact of potential precision issues arising from data exchange across heterogeneous GPUs on convergence is left for future work.

Cost-efficiency. This work studies accelerating RL training with heterogeneous GPUs and network. GPU resource pric-

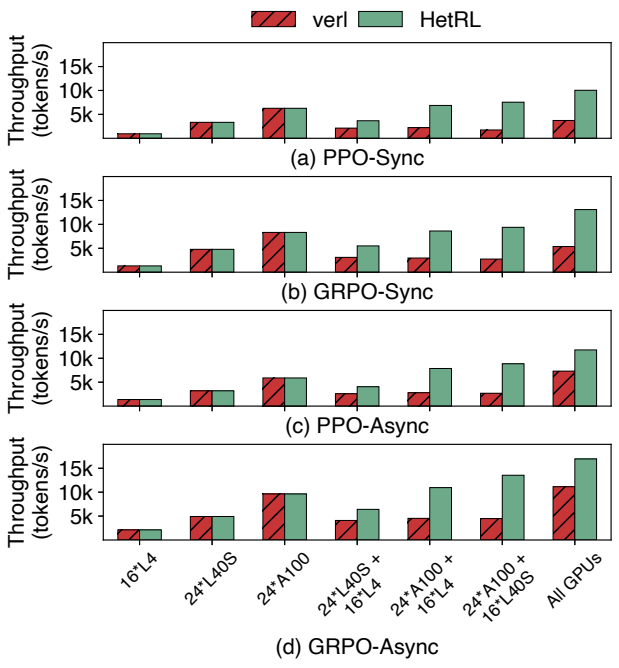

Figure 10. Throughput comparison of HetRL and verl for training Qwen-8B using different RL algorithms under varying combinations of GPUs.

ing varies by vendor, region, and purchasing plan, and is subject to adjustment and fluctuation over time. Therefore, this paper does not include a cost-efficiency comparison.

Simplicity of the scheduling heuristic. Compared to non-hybrid alternatives, the proposed hybrid scheduling algorithm introduces additional complexity into HetRL. Below, we discuss why these simpler alternatives cannot achieve comparable performance as shown in Section 5.4, their connections to our proposed hybrid scheduling algorithm, and additional design considerations that complement Section 3.

- Pure SHA. Pure SHA is insufficient for our setting, as it operates only on predefined arms (high-level decisions) and cannot effectively construct overall plans in a large combinatorial space. We therefore employ an EA to generate high-quality low-level plans under each high-level decision rather than naive enumeration or sampling.
- Pure EA (DEAP). SHA-EA can be viewed as an EA augmented with hierarchical statistical information from SHA, enabling effective early termination of candidates under unpromising high-level decisions. In contrast, pure EA relies only on information from the current population, without leveraging broader statistical signals.
- Pure ILP. An ILP formulation solved with standard solvers is inefficient in our setting, as our analytical cost model involves deeply nested min–max and max–max operations.
- SHA-ILP. While SHA can effectively prune unpromising high-level decisions, using standard solvers for low-level plan construction is impractical, as each arm requires GBs of DRAM and minutes to initialize variables and enforce constraints before solving the problem, especially when the number of GPUs is moderate to large.
- Other combinations. SHA and EA are among the simplest algorithms for search budget allocation and plan generation. Using other combinations in the same way does not reduce design or implementation complexity.

Online redeployment. The current design and implementation of HetRL do not support online redeployment with fine-grained adaptive mechanisms in highly dynamic environments. Nevertheless, HetRL can accommodate medium-granularity network variability by performing scheduling before the end of the current iteration and during model checkpointing, leveraging the checkpointing capabilities provided by verl. The updated scheduling plan can then be applied immediately after checkpointing.

Multi-tenancy. HetRL focuses on optimizing the deployment of a single RL training job. Efficiently scheduling multiple concurrent RL workloads, however, requires different scheduling algorithms and corresponding system designs, some of which are mentioned in Section 7.

7 Related Work

Systems for RL training. A series of recent work has focused on optimizing RL training. (Hu et al., 2025; Yao et al., 2023; Shen et al., 2024; Sheng et al., 2025) are pioneering and popular RL training systems that focused on the design of flexible and efficient programming models to support diverse RL workflows when they were proposed, but have continuously adopted system optimizations proposed by subsequent research. RLHFuse (Zhong et al., 2025b) introduces stage fusion strategies to improve GPU utilization for synchronous RL training. (Noukhovitch et al., 2025) explores trade-offs between training speed and quality in one-step off-policy, asynchronous RL training and AReaL (Fu et al.) proposes staleness-aware asynchronous RL training. (Wu et al., 2025a; Han et al., 2025; Zhong et al., 2025a) provides optimized system support for asynchronous RL training by minimizing bubbles caused by task dependencies. StreamRL (Zhong et al., 2025a) confirms the benefits of RL training on cross-data heterogeneous GPUs, but has strict limitations on the heterogeneity of computing resources. Distinct from existing systems, HetRL is the first system tailored for RL training in heterogeneous environments.

Heterogeneity-aware LLM training and serving. There are plenty of recent efforts have investigated on the deployment of LLM serving and training on heterogeneous GPUs

and networks. ThunderServe (Jiang et al., 2025) formulates heterogeneity-aware LLM serving as job shop scheduling problem and adopts tabu search to identify highly optimized deployment plan. HexGen (Jiang et al., 2024b) and Helix (Mei et al., 2025) formulates the problem as a maxflow problem and adopts EA and mixed integer linear programming algorithm, respectively. For heterogeneity-aware LLM training, DTFM (Yuan et al., 2022) formulates it as graph partition problem and also adopt EA. Metis (Um et al., 2024) considers the deployment on heterogeneous GPUs and homogeneous networks and also adopts depth-first search. Sailor (Strati et al., 2025) and ThunderServe (Jiang et al., 2025) further study fault tolerance and elasticity issues in geo-distributed LLM training and serving and propose heuristic lightweight re-scheduling algorithms. These systems are targeted for single-model or single-task deployments, but their scheduling optimizations can be integrated as subcomponents into our scheduling algorithm based on the multi-level search framework and successive halving. Our work shares a similar vision in improving the utilization of geo-distributed heterogeneous GPU resources, but this is the first attempt to tailor a system for RL workflows involving multiple models and tasks with complex computational and data dependencies.

Resource scheduling for DL jobs in multi-tenant clusters. There has been a lot of research on scheduling optimization for DL jobs in multi-tenant clusters. Gandiva (Xiao et al., 2018) explores opportunities for co-locating jobs on the same GPU(s) via temporal sharing, where AntMan (Xiao et al., 2020) and GSLICE (Dhakal et al., 2020) colocates traditional DL jobs via spatial sharing. Aegaeon (Xiang et al., 2025) and OpenTela (Yao et al., 2026) orchestrate cluster-level GPU resources for multi-tenant LLM serving. Extending HetRL with optimizations for multi-tenant cluster scheduling is a promising direction for future work.

8 Conclusion

In this paper, we explore the opportunity to deploy RL training in infrastructures with heterogeneous GPU models and network characteristics. Toward this end, we present HetRL, a distributed system tailored for such deployments, equipped with (1) a hybrid scheduling algorithm that partitions RL workflows into tasklets, assigns them to heterogeneous computing resources, and efficiently identifies near-optimal solutions, and (2) an ILP-based scheduling algorithm for obtaining optimal solutions. Our empirical studies suggest that HetRL achieves up to 9.17$\times$ higher throughput than SoTA RL training systems under various workloads and settings, demonstrating the effectiveness of our scheduling algorithms and system optimizations.

References

Aws elastic fabric adapter. https://aws.amazon.com/hpc/efa/, 2025a.

Aws ofi nccl. https://github.com/aws/aws-ofi-nccl, 2025b.

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.662. URL https://doi.org/10.18653/v1/2024.acl-long.662.

Baldwin, J. M. A new factor in evolution. *Adaptive individuals in evolving populations: Models and algorithms*, pp. 59–80, 1896.

Bui, T. N. and Moon, B. R. Genetic algorithm and graph partitioning. *IEEE Trans. Computers*, 45(7):841–855, 1996. doi: 10.1109/12.508322. URL https://doi.org/10.1109/12.508322.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Ségerie, C., Carroll, M., Peng, A., Christoffersen, P. J. K., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A. D., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. Safe RLHF: safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=TyFrPOKYXw.

DeepSeek-AI. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/ARXIV.2412.19437. URL https://doi.org/10.48550/arXiv.2412.19437.

Dhakal, A., Kulkarni, S. G., and Ramakrishnan, K. K. GSLICE: controlled spatial sharing of gpus for a scalable

inference platform. In Fonseca, R., Delimitrou, C., and Ooi, B. C. (eds.), *SoCC '20: ACM Symposium on Cloud Computing, Virtual Event, USA, October 19-21, 2020*, pp. 492–506. ACM, 2020. doi: 10.1145/3419111.3421284. URL https://doi.org/10.1145/3419111.3421284.

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., and Gagné, C. DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175, jul 2012.

Fu, W., Gao, J., Shen, X., Zhu, C., Mei, Z., He, C., Xu, S., Wei, G., Mei, J., JIASHU, W., et al. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Gao, Y., He, Y., Li, X., Zhao, B., Lin, H., Liang, Y., Zhong, J., Zhang, H., Wang, J., Zeng, Y., Gui, K., Tong, J., and Yang, M. An empirical study on low GPU utilization of deep learning jobs. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, ICSE 2024, Lisbon, Portugal, April 14-20, 2024*, pp. 96:1–96:13. ACM, 2024. doi: 10.1145/3597503.3639232. URL https://doi.org/10.1145/3597503.3639232.

Han, Z., You, A., Wang, H., Luo, K., Yang, G., Shi, W., Chen, M., Zhang, S., Lan, Z., Deng, C., Ji, H., Liu, W., Huang, Y., Zhang, Y., Pan, C., Wang, J., Huang, X., Li, C., and Wu, J. Asyncflow: An asynchronous streaming RL framework for efficient LLM post-training. *CoRR*, abs/2507.01663, 2025. doi: 10.48550/ARXIV.2507.01663. URL https://doi.org/10.48550/arXiv.2507.01663.

He, Y., Yang, H., Lu, Y., Klimovic, A., and Alonso, G. Resource multiplexing in tuning and serving large language models. In Altinbüken, D. and Stutsman, R. (eds.), *Proceedings of the 2025 USENIX Annual Technical Conference, USENIX ATC 2025, Boston, MA, USA, July 7-9, 2025*, pp. 1639–1655. USENIX Association, 2025. URL https://www.usenix.org/conference/atc25/presentation/he-yongjun.

Hinton, G. E. and Nowlan, S. J. How learning can guide evolution. *Complex Syst.*, 1(3), 1987. URL http://www.complex-systems.com/abstracts/v01_i03_a06.html.

Hu, J., Wu, X., Shen, W., Liu, J. K., Wang, W., Jiang, S., Wang, H., Chen, H., Chen, B., Fang, W., Xianyu, Cao, Y., Xu, H., and Liu, Y. Openrlhf: A ray-based easy-to-use, scalable and high-performance RLHF framework. In Habernal, I., Schulam, P., and Tiedemann, J. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - System Demonstrations, Suzhou, China, November 4-9, 2025*, pp. 656–666. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-DEMOS.48. URL https://doi.org/10.18653/v1/2025.emnlp-demos.48.

Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M. X., Lee, H., Ngiam, J., Le, Q. V., Wu, Y., and Chen, Z. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 103–112, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/093f65e080a295f8076b1c5722a46aa2-Abstract.html.

Jamieson, K. G. and Talwalkar, A. Non-stochastic best arm identification and hyperparameter optimization. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pp. 240–248. JMLR.org, 2016. URL http://proceedings.mlr.press/v51/jamieson16.html.

Jiang, Y., Fu, F., Yao, X., He, G., Miao, X., Klimovic, A., Cui, B., Yuan, B., and Yoneki, E. Demystifying cost-efficiency in llm serving over heterogeneous gpus. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, Canada, July 13-19, 2025*. OpenReview.net, 2024a.

Jiang, Y., Yan, R., Yao, X., Zhou, Y., Chen, B., and Yuan, B. Hexgen: Generative inference of large language model over heterogeneous environment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pp. 21946–21961. PMLR / OpenReview.net, 2024b. URL https://proceedings.mlr.press/v235/jiang24f.html.

Jiang, Y., Fu, F., Yao, X., Wang, T., Cui, B., Klimovic, A., and Yoneki, E. Thunderserve: High-performance and cost-efficient LLM serving in cloud environments. In Zaharia, M., Joshi, G., and Lin, Y. C. (eds.), *Proceedings of the Eighth Conference on Machine Learning and Systems, MLSys 2025, Santa Clara, CA, USA, May 12-15, 2025*. OpenReview.net/mlsys.org, 2025. URL https://openreview.net/forum?id=44PwmgOpAt.

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback.

Trans. Mach. Learn. Res., 2025, 2025. URL https://openreview.net/forum?id=f7OkIurx4b.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In Flinn, J., Seltzer, M. I., Druschel, P., Kaufmann, A., and Mace, J. (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Bras, R. L., Tafjord, O., Wilhelm, C., Soldaini, L., Smith, N. A., Wang, Y., Dasigi, P., and Hajishirzi, H. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124, 2024. doi: 10.48550/ARXIV.2411.15124. URL https://doi.org/10.48550/arXiv.2411.15124.

Li, L., Jamieson, K. G., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=ry18Ww5ee.

Li, Z., Zheng, L., Zhong, Y., Liu, V., Sheng, Y., Jin, X., Huang, Y., Chen, Z., Zhang, H., Gonzalez, J. E., and Stoica, I. Alpaserve: Statistical multiplexing with model parallelism for deep learning serving. In Geambasu, R. and Nightingale, E. (eds.), *17th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2023, Boston, MA, USA, July 10-12, 2023*, pp. 663–679. USENIX Association, 2023. URL https://www.usenix.org/conference/osdi23/presentation/li-zhouhan.

Licker, N., Hu, K., Zaytsev, V., and Chen, L. fabric-lib: Rdma point-to-point communication for llm systems. In *Proceedings of the Seventh Annual Conference on Machine Learning and Systems, MLSys 2026, Bellevue,, WA, USA, May 18-22, 2026*. mlsys.org, 2026.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Mei, Y., Zhuang, Y., Miao, X., Yang, J., Jia, Z., and Vinayak, R. Helix: Serving large language models over heterogeneous gpus and network via max-flow. In Eeckhout, L., Smaragdakis, G., Liang, K., Sampson, A., Kim, M. A., and Rossbach, C. J. (eds.), *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, ASPLOS 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pp. 586–602. ACM, 2025. doi: 10.1145/3669940.3707215. URL https://doi.org/10.1145/3669940.3707215.

Noukhovitch, M., Huang, S., Xhonneux, S., Hosseini, A., Agarwal, R., and Courville, A. C. Asynchronous RLHF: faster and more efficient off-policy RL for language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=FhTAG591Ve.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y. Zero: memory optimizations toward training trillion parameter models. In Cuicchi, C., Qualters, I., and Kramer, W. T. (eds.), *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, pp. 20. IEEE/ACM, 2020. doi: 10.1109/SC41405.2020.00024. URL https://doi.org/10.1109/SC41405.2020.00024.

Rota, G.-C. The number of partitions of a set. *The American Mathematical Monthly*, 71(5):498–504, 1964.

Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1506.02438.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Shen, G., Wang, Z., Delalleau, O., Zeng, J., Dong, Y., Egert, D., Sun, S., Zhang, J. J., Jain, S., Taghibakhshi, A., Ausin, M. S., Aithal, A., and Kuchaiev, O. Nemo-aligner: Scalable toolkit for efficient model alignment. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=yK2eGE8QVW.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053, 2019. URL http://arxiv.org/abs/1909.08053.

Soper, A. J., Walshaw, C., and Cross, M. A combined evolutionary search and multilevel optimisation approach to graph-partitioning. *J. Glob. Optim.*, 29(2):225–241, 2004. doi: 10.1023/B:JOGO.0000042115.44455.F3. URL https://doi.org/10.1023/B:JOGO.0000042115.44455.f3.

Strati, F., Elvinger, P., Kerimoglu, T., and Klimovic, A. ML training with cloud GPU shortages: Is cross-region the answer? In *Proceedings of the 4th Workshop on Machine Learning and Systems, EuroMLSys 2024, Athens, Greece, 22 April 2024*, pp. 107–116. ACM, 2024. doi: 10.1145/3642970.3655843. URL https://doi.org/10.1145/3642970.3655843.

Strati, F., Zhang, Z., Manos, G., Périz, I. S., Hu, Q., Chen, T., Buzcu, B., Han, S., Delgado, P., and Klimovic, A. Sailor: Automating distributed training over dynamic, heterogeneous, and geo-distributed clusters. In Won, Y., Kwon, Y., Yuan, D., and Isaacs, R. (eds.), *Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles, SOSP 2025, Lotte Hotel World, Seoul, Republic of Korea, October 13-16, 2025*, pp. 204–220. ACM, 2025. doi: 10.1145/3731569.3764839. URL https://doi.org/10.1145/3731569.3764839.

Um, T., Oh, B., Kang, M., Lee, W., Kim, G., Kim, D., Kim, Y., Muzzammil, M., and Jeon, M. Metis: Fast automatic distributed training on heterogeneous gpus. In Bagchi, S. and Zhang, Y. (eds.), *Proceedings of the 2024 USENIX Annual Technical Conference, USENIX ATC 2024, Santa Clara, CA, USA, July 10-12, 2024*, pp. 563–578. USENIX Association, 2024. URL https://www.usenix.org/conference/atc24/presentation/um.

Wu, B., Wang, S., Tang, Y., Ding, J., Helenowski, E., Tan, L., Xu, T., Gowda, T., Chen, Z., Zhu, C., Tang, X., Qian, Y., Zhu, B., and Hou, R. Llamarl: A distributed asynchronous reinforcement learning framework for efficient large-scale LLM training. *CoRR*, abs/2505.24034, 2025a. doi: 10.48550/ARXIV.2505.24034. URL https://doi.org/10.48550/arXiv.2505.24034.

Wu, Y., Liu, X., Jin, S., Xu, C., Qian, F., Mao, Z. M., Lentz, M., Zhuo, D., and Stoica, I. Hetermoe: Efficient training of mixture-of-experts models on heterogeneous gpus. *CoRR*, abs/2504.03871, 2025b. doi: 10.48550/ARXIV.2504.03871. URL https://doi.org/10.48550/arXiv.2504.03871.

Xiang, Y., Li, X., Qian, K., Yang, Y., Zhu, D., Yu, W., Zhai, E., Liu, X., Jin, X., and Zhou, J. Aegaeon: Effective GPU pooling for concurrent LLM serving on the market. In Won, Y., Kwon, Y., Yuan, D., and Isaacs, R. (eds.), *Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles, SOSP 2025, Lotte Hotel World, Seoul, Republic of Korea, October 13-16, 2025*, pp. 1030–1045. ACM, 2025. doi: 10.1145/3731569.3764815. URL https://doi.org/10.1145/3731569.3764815.

Xiao, W., Bhardwaj, R., Ramjee, R., Sivathanu, M., Kwatra, N., Han, Z., Patel, P., Peng, X., Zhao, H., Zhang, Q., Yang, F., and Zhou, L. Gandiva: Introspective cluster scheduling for deep learning. In Arpaci-Dusseau, A. C. and Voelker, G. (eds.), *13th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2018, Carlsbad, CA, USA, October 8-10, 2018*, pp. 595–610. USENIX Association, 2018. URL https://www.usenix.org/conference/osdi18/presentation/xiao.

Xiao, W., Ren, S., Li, Y., Zhang, Y., Hou, P., Li, Z., Feng, Y., Lin, W., and Jia, Y. Antman: Dynamic scaling on GPU clusters for deep learning. In *14th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2020, Virtual Event, November 4-6, 2020*, pp. 533–548. USENIX Association, 2020. URL https://www.usenix.org/conference/osdi20/presentation/xiao.

Yao, X., Jiang, Y., Badanin, I., Hu, Q., Yuan, B., Yoneki, E., Schlag, I., and Klimovic, A. Opentela: Unifying decentralized hpc clusters for heterogeneous llm serving. In *20th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2026, Seattle, WA, USA, July 13–15, 2026*. USENIX Association, 2026.

Yao, Z., Aminabadi, R. Y., Ruwase, O., Rajbhandari, S., Wu, X., Awan, A. A., Rasley, J., Zhang, M., Li, C., Holmes, C., Zhou, Z., Wyatt, M., Smith, M., Kurilenko, L., Qin, H., Tanaka, M., Che, S., Song, S. L., and He, Y. Deepspeed-chat: Easy, fast and affordable RLHF training of chatgpt-like models at all scales. *CoRR*, abs/2308.01320, 2023. doi: 10.48550/ARXIV.2308.01320. URL https://doi.org/10.48550/arXiv.2308.01320.

Yuan, B., He, Y., Davis, J., Zhang, T., Dao, T., Chen, B., Liang, P., Ré, C., and Zhang, C. Decentralized training of foundation models in heterogeneous environments. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a37d615b61f999a5fa276adb14643476-Abstract-Conference.html.

Zhang, Z. Understanding gpu architecture implications on llm serving workloads. Master's thesis, ETH Zurich, 2024.

Zheng, L., Li, Z., Zhang, H., Zhuang, Y., Chen, Z., Huang, Y., Wang, Y., Xu, Y., Zhuo, D., Xing, E. P., Gonzalez, J. E., and Stoica, I. Alpa: Automating inter- and intra-operator parallelism for distributed deep learning. In Aguilera, M. K. and Weatherspoon, H. (eds.), *16th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2022, Carlsbad, CA, USA, July 11-13, 2022*, pp. 559–578. USENIX Association, 2022. URL https://www.usenix.org/conference/osdi22/presentation/zheng-lianmin.

Zhong, Y., Zhang, Z., Song, X., Hu, H., Jin, C., Wu, B., Chen, N., Chen, Y., Zhou, Y., Wan, C., Zhou, H., Jiang, Y., Zhu, Y., and Jiang, D. Streamrl: Scalable, heterogeneous, and elastic RL for llms with disaggregated stream generation. *CoRR*, abs/2504.15930, 2025a. doi: 10.48550/ARXIV.2504.15930. URL https://doi.org/10.48550/arXiv.2504.15930.

Zhong, Y., Zhang, Z., Wu, B., Liu, S., Chen, Y., Wan, C., Hu, H., Xia, L., Ming, R., Zhu, Y., and Jin, X. Optimizing RLHF training for large language models with stage fusion. In Benson, T. A. and Mysore, R. N. (eds.), *22nd USENIX Symposium on Networked Systems Design and Implementation, NSDI 2025, Philadelphia, PA, USA, April 28-30, 2025*, pp. 489–503. USENIX Association, 2025b. URL https://www.usenix.org/conference/nsdi25/presentation/zhong.

A Proof of Proposition 1

We follow the notation in Section 3.1.

Proof. We prove NP-hardness for the abstract formulation in Section 3.1 via a theoretical reduction from the *balanced graph partitioning* problem, by constructing a specific instantiation of the objective function C. This construction of the objective function is used solely for complexity analysis. In contrast, the cost model introduced in Appendix B is designed to capture execution behavior in practical systems.

Given an instance of balanced graph partitioning on a weighted undirected graph $H = (U, F)$ where $|U| = k \cdot s$, the goal is to partition U into k disjoint subsets of equal size s while minimizing the cut weight. The balanced graph partitioning problem is NP-hard.

We construct a restricted instance of the abstract formulation in Section 3.1. Consider a single RL task with a fixed partitioning strategy $(TP, DP, PP) = (t, d, p)$ such that $t \cdot d \cdot p = |U|$. This yields $|V_L| = |U|$ tasklets, leaving only the assignment $\sigma : V_L \to V_D$ as the decision variable.

We create $|V_D| = |U|$ devices and partition them into d disjoint groups (islands), each containing $t \cdot p$ devices. Let $k := d$ and $s := t \cdot p$, so that $k \cdot s = |U|$. We set $M_{\text{working}}(l) = 0$, $M_{\text{model}}(l) = 1$, and $M_{\text{gpu}}(d) = 1$ for all tasklets and devices. Then, by constraints (C2) and (C3), each device hosts exactly one tasklet.

We establish a bijection between vertices $u \in V_D$ and tasklets $l_u \in V_L$. For each edge $(u, v) \in F$ with weight w_{uv}, we define the communication cost as 0 if l_u and l_v are assigned to the same island, and w_{uv} otherwise. The objective of the constructed instance is defined as the total communication cost over all edges in F.

Under this construction, any assignment σ induces a balanced partition of U into k subsets of size s, and its objective value equals the cut weight of the partition. Conversely, any balanced partition yields a feasible assignment with the same objective value.

Therefore, the reduction is polynomial-time, and the constructed objective matches the cut weight of the induced partition. Hence, the restricted scheduling problem is NP-hard. Since this restricted problem is a special case of the abstract scheduling problem in Section 3.1, the general problem is also NP-hard. □

B Cost Modeling

In this section, we first model computational and communication costs of the main components involved in training and serving transformer models. Then we model the cost of the main tasks involved in RL training. Finally, we model the end-to-end cost for different RL training algorithms.

Since the modeling of memory cost is independent of device and network attributes, we simply follow prior work (Sheng et al., 2025; Zheng et al., 2022) and ensure that the tasklets' memory footprint fits on every assigned device.

B.1 Notation

We extend the notation in Section 3 for device topology graph. Let $\mathbf{G_D} = (V_D, E_D, \mathbf{comp}, \mathbf{mem}, \mathbf{hbm}, \mathbf{A}, \mathbf{B})$ denote the device topology graph of heterogeneous environments, where $V_D = \{d_1, \ldots, d_N\}$ are a set of N devices and $E_D = V_D \times V_D$ are communication channels between devices. For each device d, the computation capability, memory capacity, and HBM bandwidth are $\text{comp}_d, \text{mem}_d, \text{hbm}_d \in \mathbb{R}_+$. Collect them into vectors $\mathbf{comp}, \mathbf{mem}, \mathbf{hbm} \in \mathbb{R}_+^N$. Each edge $e_{d,d'}$ is labeled with latency $\alpha_{d,d'}$ and bandwidth $\beta_{d,d'}$, where $\alpha_{d,d'}, \beta_{d,d'} \in \mathbb{R}_+$. Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}_+^{N \times N}$ be the corresponding matrices.

We extend the notation in Section 3 for tasklet graph. Let $\mathbf{G_L}_{i,j}^t = (V_{L_{i,j}}^t, E_{L_{i,j}}^t)$ denote the subgraph constructed by the set of tasklets of the j-th pipeline stage of the i-th replica in the t-th task. Let $\mathbf{G_D}_{i,j}^t = (V_{D_{i,j}}^t, E_{D_{i,j}}^t)$ be the subgraph constructed by the set of devices assigned to these tasklets and we have $\sigma(V_{L_{i,j}}^t) = V_{D_{i,j}}^t$. Similarly, let $\mathbf{G_L}_{j,k}^t = (V_{L_{j,k}}^t, E_{L_{j,k}}^t)$ denote the subgraph constructed by the set of tasklets of the k-th tensor shards of the j-th pipeline stage in the t-th task and $\mathbf{G_D}_{j,k}^t = (V_{D_{j,k}}^t, E_{D_{j,k}}^t)$ denote the assigned devices.

Now we introduce other required notation. Let h_1^t, h_2^t, nl^t be the hidden size, the intermediate size, and number of layers of the LLM of the t-th task. Let nl_j^t, TP_j^t be number of layers and TP degree in j-th pipeline stage of the t-th task. Let seq_{in} and seq_{out} denote the maximum sequence lengths of the input prompts and output responses. Let mbs, nm denote micro-batch size and number of micro-batches. Please note that we have preprocessed nm based on the number of responses generated per prompt, the data parallelism degree. Let $B_{\text{BF16}}, B_{\text{FP32}}$ be the data sizes of the BF16 and FP32 data types. Let ring denote a function that returns the set of all feasible ring graphs of given devices. For brevity, we have omitted the vocabulary and token embeddings in the cost model, but they are included in our actual implementation.

B.2 Modeling component-level computation and communication costs.

Tensor parallelism communication. The communication volume transferred between a pair of neighboring GPUs in

$\mathbf{G_D}^t_{i,j}$ per all-reduce during TP can be estimated by:

$$\mathrm{cv}_{\mathrm{tp}}(t,i,j) = B_{\mathrm{BF16}} \cdot mbs \cdot (\mathrm{seq}_{\mathrm{in}} + \mathrm{seq}_{\mathrm{out}}) \cdot h_1^t \cdot \frac{2 \cdot (\mathrm{TP}_j^t - 1)}{\mathrm{TP}_j^t}$$

The TP communication cost of forward passes for $\mathbf{G_L}^t_{i,j}$ can be estimated by:

$$C_{\mathrm{tp}}(t,i,j) = 2 \cdot nm \cdot nl_j^t \cdot \min_{r \in \mathrm{ring}(\mathbf{G_D}^t_{i,j})} \max_{e_{d,d'} \in r} (\alpha_{d,d'} + \frac{\mathrm{cv}_{\mathrm{tp}}(t,i,j)}{\beta_{d,d'}})$$

If recomputation is enabled, the TP communication cost of both forward and backward passes for $\mathbf{G_L}^t_{i,j}$ can be estimated by:

$$C_{\mathrm{tp}}(t,i,j) = 6 \cdot nm \cdot nl_j^t \cdot \min_{r \in \mathrm{ring}(\mathbf{G_D}^t_{i,j})} \max_{e_{d,d'} \in r} (\alpha_{d,d'} + \frac{\mathrm{cv}_{tp}(t,i,j)}{\beta_{d,d'}})$$

The overall TP communication cost can be estimated by:

$$C_{\mathrm{tp}}^t = \max_{i,j} C_{\mathrm{tp}}(t,i,j)$$

Pipeline parallelism communication. The communication volume transferred between the j-th and $j+1$-th pipeline stages per micro-batch during PP can be estimated by:

$$\mathrm{cv}_{\mathrm{pp}}(t,i,j) = B_{\mathrm{BF16}} \cdot mbs \cdot (\mathrm{seq}_{\mathrm{in}} + \mathrm{seq}_{\mathrm{out}}) \cdot h_1^t$$

The PP communication cost of forward passes between $\mathbf{G_L}^t_{i,j}$ and $\mathbf{G_L}^t_{i,j+1}$ can be by:

$$C_{\mathrm{pp}}(t,i,j) = nm \cdot \min_{d \in V_D{}^t_{i,j}, d' \in V_D{}^t_{i,j+1}} (\alpha_{d,d'} + \frac{\mathrm{cv}_{\mathrm{pp}}(t,i,j)}{\beta_{d,d'}})$$

The PP communication cost of both forward and backward passes between $\mathbf{G_L}^t_{i,j}$ and $\mathbf{G_L}^t_{i,j+1}$ can be estimated by:

$$C_{\mathrm{pp}}(t,i,j) = 2 \cdot nm \cdot \min_{d \in V_D{}^t_{i,j}, d' \in V_D{}^t_{i,j+1}} (\alpha_{d,d'} + \frac{\mathrm{cv}_{\mathrm{pp}}(t,i,j)}{\beta_{d,d'}})$$

The overall PP communication cost can be estimated by:

$$C_{\mathrm{pp}}^t = \max_{i,j} C_{\mathrm{pp}}(t,i,j)$$

Data parallelism communication. The communication volume transferred between a pair of neighboring GPUs in $\mathbf{G_D}^t_{j,k}$ per all-reduce during DP can be estimated by:

$$\mathrm{cv}_{\mathrm{dp}}(t,j,k) = B_{\mathrm{BF16}} \cdot nl_j^t \cdot (4 \cdot {h_1^t}^2 + 3 \cdot h_1^t \cdot h_2^t) \cdot \frac{2 \cdot (|V_L{}^t_{j,k}| - 1)}{|V_L{}^t_{j,k}| \cdot \mathrm{TP}_j^t}$$

The DP communication cost for $\mathbf{G_D}^t_{j,k}$ per iteration can be estimated by:

$$C_{\mathrm{dp}}(t,j,k) = \min_{r \in \mathrm{ring}(\mathbf{G_D}^t_{j,k})} \max_{e_{d,d'} \in r} (\alpha_{d,d'} + \frac{\mathrm{cv}_{\mathrm{dp}}(t,j,k)}{\beta_{d,d'}})$$

The overall DP communication cost can be estimated by:

$$C_{\mathrm{dp}}^t = \max_{j,k} C_{\mathrm{dp}}(t,j,k)$$

Computation. The computational cost of a transformer layer per forward pass per sample can be estimated by:

$$C_{\mathrm{comp}}^{\mathrm{qkvo_proj}}(t,i,j,k) = 2 \cdot 4 \cdot (\mathrm{seq}_{\mathrm{in}} + \mathrm{seq}_{\mathrm{out}}) \cdot {h_1^t}^2$$
$$C_{\mathrm{comp}}^{\mathrm{attn}}(t,i,j,k) = 2 \cdot 2 \cdot (\mathrm{seq}_{\mathrm{in}} + \mathrm{seq}_{\mathrm{out}})^2 \cdot h_1^t$$
$$C_{\mathrm{comp}}^{\mathrm{mlp}}(t,i,j,k) = 2 \cdot 3 \cdot (\mathrm{seq}_{\mathrm{in}} + \mathrm{seq}_{\mathrm{out}}) \cdot h_1^t \cdot h_2^t$$
$$C_{\mathrm{comp}}^{\mathrm{layer}}(t,i,j,k) = C_{\mathrm{comp}}^{\mathrm{qkvo_proj}} + C_{\mathrm{comp}}^{\mathrm{attn}} + C_{\mathrm{comp}}^{\mathrm{mlp}}$$

The computation cost of forward passes for $l_{i,j,k}^t$ on device $d = \sigma(l_{i,j,k}^t)$ can be estimated by:

$$C_{\mathrm{comp}}(t,i,j,k) = nm \cdot mbs \cdot nl_j^t \cdot \frac{C_{\mathrm{comp}}^{\mathrm{layer}}}{\mathrm{comp}_d \cdot \mathrm{TP}_j^t}$$

The computation cost of both forward and backward passes for $l_{i,j,k}^t$ on device $d = \sigma(l_{i,j,k}^t)$ can be estimated by:

$$C_{\mathrm{comp}}(t,i,j,k) = 3 \cdot nm \cdot mbs \cdot nl_j^t \cdot \frac{C_{\mathrm{comp}}^{\mathrm{layer}}}{\mathrm{comp}_d \cdot \mathrm{TP}_j^t}$$

The computation cost for $\mathbf{G_L}^t_{i,j}$ can be estimated by:

$$C_{\mathrm{comp}}(t,i,j) = \max_k C_{\mathrm{comp}}(t,i,j,k)$$

The overall computation cost can be estimated by:

$$C_{\mathrm{comp}}^t = \max_{i,j} C_{\mathrm{comp}}(t,i,j)$$

Please note that here we set $\mathrm{seq}_{\mathrm{out}}$ to 0 in the estimation for the actor generation task.

Pipeline bubbles. The cost of pipeline bubbles for the i-th replica can be estimated by:

$$C_{\mathrm{bubble}}(t,i) = \sum_{j \neq 0} \frac{C_{\mathrm{comp}}(t,i,j) + C_{\mathrm{tp}}(t,i,j) + C_{\mathrm{pp}}(t,i,j)}{nm}$$

The overall cost of pipeline bubbles can be estimated by:

$$C_{\mathrm{bubble}}^t = \max_i C_{\mathrm{bubble}}(t,i)$$

Decoding (HBM-bandwidth bound). The hbm cost for $l^t_{i,j,k}$ on device $d = \sigma(l^t_{i,j,k})$ can be estimated by:

$$C_{\text{hbm}}(t,i,j,k) = \text{seq}_{\text{out}} \cdot nm \cdot mbs \cdot \frac{B_{\text{BF16}} \cdot nl^t_j \cdot (4 \cdot {h^t_1}^2 + 3 \cdot h^t_1 \cdot h^t_2)}{dbs_d \cdot \text{hbm}_d \cdot \text{TP}^t_j},$$

where dbs_d is the decoding batch size of the LLM serving engine deployed on device d. The hbm cost for the i-th replica can be estimated by:

$$C_{\text{hbm}}(t,i) = \max_{j,k} C_{\text{hbm}}(t,i,j,k)$$

The overall hbm cost can be estimated by:

$$C^t_{\text{hbm}} = \max_i C_{\text{hbm}}(t,i)$$

Resharding. The communication volume for all-gather during model resharding between a pair of neighboring GPUs in the i-th model replica group of the actor training task t can be estimated by:

$$\text{cv}_{\text{all-gather}}(t,i) = B_{\text{BF16}} \cdot nl^t \cdot (4 \cdot {h^t_1}^2 + 3 \cdot h^t_1 \cdot h^t_2) \cdot \frac{|\mathbf{V}_{\mathbf{L}^t_i}| - 1}{|\mathbf{V}_{\mathbf{L}^t_i}|}$$

The all-gather cost for $\mathbf{G}_{\mathbf{D}^t_i}$ per iteration can be estimated by:

$$C_{\text{all-gather}}(t,i) = \min_{r \in \text{ring}(\mathbf{V}_{\mathbf{D}^t_i})} \max_{e_{d,d'} \in r} (\alpha_{d,d'} + \frac{\text{cv}_{\text{all-gather}}(t,i)}{\beta_{d,d'}})$$

The naive overall model resharding cost required for synchronous RL training can be estimated by:

$$C_{\text{reshard}} = \max_i C_{\text{all-gather}}(t,i)$$

Synchronization. The communication volume for all-gather during weight synchronization between a pair of neighboring GPUs in the i-th model replica group of actor training task t can be estimated by:

$$\text{cv}_{\text{all-gather}}(t,i) = B_{\text{BF16}} \cdot nl^t \cdot (4 \cdot {h^t_1}^2 + 3 \cdot h^t_1 \cdot h^t_2) \cdot \frac{|\mathbf{V}_{\mathbf{L}^t_i}| - 1}{|\mathbf{V}_{\mathbf{L}^t_i}|}$$

The all-gather cost for $\mathbf{G}_{\mathbf{D}^t_i}$ per iteration can be estimated by:

$$C_{\text{all-gather}}(t,i) = \min_{r \in \text{ring}(\mathbf{V}_{\mathbf{D}^t_i})} \max_{e_{d,d'} \in r} (\alpha_{d,d'} + \frac{\text{cv}_{\text{all-gather}}(t,i)}{\beta_{d,d'}})$$

The communication volume for broadcast during weight synchronization between a pair of neighboring GPUs in the i'-th model replica group of actor generation task t' can be estimated by:

$$\text{cv}_{\text{broadcast}}(t',i') = B_{\text{BF16}} \cdot nl^{t'} \cdot (4 \cdot {h^{t'}_1}^2 + 3 \cdot h^{t'}_1 \cdot h^{t'}_2) \cdot \frac{|\mathbf{V}_{\mathbf{L}^t_i}| - 1}{|\mathbf{V}_{\mathbf{L}^t_i}|}$$

The broadcast cost for $\mathbf{G}_{\mathbf{D}^{t'}_{i'}}$ per iteration can be estimated by:

$$C_{\text{broadcast}}(t',i') = \min_{r \in \text{ring}(\mathbf{V}_{\mathbf{D}^{t'}_{i'}})} \max_{e_{d,d'} \in r} (\alpha_{d,d'} + \frac{\text{cv}_{\text{all-gather}}(t',i')}{\beta_{d,d'}})$$

The communication volume for naive point-to-point communication during weight synchronization between a GPU assigned to actor training task t and another GPU assigned to actor generation task t' can be estimated by:

$$\text{cv}_{\text{p2p}}(t,t') = B_{\text{BF16}} \cdot nl^t \cdot (4 \cdot {h^t_1}^2 + 3 \cdot h^t_1 \cdot h^t_2)$$

The point-to-point communication cost for $\mathbf{G}_{\mathbf{D}^{t'}_{i'}}$ per iteration can be estimated by:

$$C_{\text{p2p}}(t,t') = \min_{d \in |V^t_D|, d' \in |V^{t'}_D|} (\alpha_{d,d'} + \frac{\text{cv}_{\text{p2p}}(t',i')}{\beta_{d,d'}})$$

The naive overall weight synchronization cost required for synchronous RL training can be estimated by:

$$C_{\text{sync}} = \min_i C_{\text{all-gather}}(t,i) + \max_{i'} C_{\text{broadcast}}(t',i') + C_{\text{p2p}}(t,t')$$

B.3 Modeling task-level costs.

One can simply add up the overall costs provided by the above equations for estimation. However, for more fine-grained estimates, the cost models are as follows:

Generation. The cost of generation task can be estimated by:

$$\Psi^{\text{gen}}(C^t_{\text{comp}}, C^t_{\text{tp}}, C^t_{\text{pp}}, C^t_{\text{hbm}}) = \max_i(\max_j(C_{\text{comp}}(t,i,j) + C_{\text{tp}}(t,i,j) + C_{\text{pp}}(t,i,j) + C_{\text{hbm}}(t,i,j)))$$

Inference. The cost of inference tasks can be estimated by:

$$\Psi^{\text{inf}}(C^t_{\text{comp}}, C^t_{\text{tp}}, C^t_{\text{pp}}) = \max_i(\max_j(C_{\text{comp}}(t,i,j) + C_{\text{tp}}(t,i,j) + C_{\text{pp}}(t,i,j)))$$

Training. The cost of training tasks can be estimated by:

$$\begin{aligned}\Psi^{\text{train}}&(C^t_{\text{comp}}, C^t_{\text{tp}}, C^t_{\text{pp}}, C^t_{\text{dp}}, C^t_{\text{bubble}})\\ &= \max_i(\max_j(C_{\text{comp}}(t,i,j) + C_{\text{tp}}(t,i,j)\\ &\quad + C_{\text{pp}}(t,i,j)) + C_{\text{bubble}}(t,i)) + C^t_{\text{dp}}\end{aligned}$$

For load balancing strategies, one can tune $\text{seq}_{\text{in}}, mb, nl^t_j$ during cost modeling.

B.4 Modeling end-to-end costs.

Let $C^{1:6}(\cdot)$ denote the cost models of actor generation, reward inference, reference inference, critic inference, actor training, and critic training, respectively, and $\Phi(\cdot)$ denotes the cost of the tasks without dependencies which can be defined as

$$\begin{aligned}\Phi(\{C^t\}) &= \max_t C^t + (1-\eta)\left(\sum_t C^t - \max_t C^t\right),\\ &\eta \in [0,1],\end{aligned}$$

where the coefficient η parameterizes the level of task parallelism (0: sequential; 1: fully parallel; else: partial).

The cost mode of t-th task is given by:

$$C^t = \begin{cases}\Psi^{\text{gen}}(C^t_{\text{comp}}, C^t_{\text{hbm}}, C^t_{\text{tp}}, C^t_{\text{pp}}), & t = 1,\\ \Psi^{\text{inf}}(C^t_{\text{comp}}, C^t_{\text{tp}}, C^t_{\text{pp}}), & t \in \{2,3,4\},\\ \Psi^{\text{train}}(C^t_{\text{comp}}, C^t_{\text{tp}}, C^t_{\text{pp}}, C^t_{\text{dp}}, & \\ \quad C^t_{\text{bubble}}), & t \in \{5,6\},\end{cases}$$

Finally, the cost of different RL training algorithms can be estimated as below:

Synchronous PPO.

$$\begin{aligned}C_{\text{SyncPPO}} &= C^1 + \Phi(\{C^2, C^3, C^4\}) + \Phi(\{C^5, C^6\})\\ &\quad + C_{\text{reshard}}\end{aligned}$$

Asynchronous PPO.

$$\begin{aligned}C_{\text{AsyncPPO}} &= \max(C^1, \Phi(\{C^2, C^3, C^4\}) +\\ &\quad \Phi(\{C^5, C^6\})) + C_{\text{sync}}\end{aligned}$$

Synchronous GRPO.

$$C_{\text{SyncGRPO}} = C^1 + \Phi(\{C^2, C^3\}) + C^6 + +C_{\text{reshard}}$$

Asynchronous GRPO.

$$C_{\text{AsyncGRPO}} = \max(C^1, \Phi(\{C^2, C^3\}) + C^6) + C_{\text{sync}}$$