

MicroMoE: Fine-grained Load Balancing for Mixture-of-Experts with Token Scheduling

Chenqi Zhao

cqzhangyu@pku.edu.cn

Peking University

China

Linhai Song

songlinhai@ict.ac.cn

Institute of Computing Technology, Chinese Academy of Sciences
China

Wenfei Wu

wenfeiwu@pku.edu.cn

Peking University

China

Yuchen Xu

xu.yuchen@pku.edu.cn

Peking University

China

Abstract

Mixture-of-Experts (MoE) has emerged as a promising approach to scale up deep learning models due to its significant reduction in computational resources. However, the dynamic nature of MoE leads to load imbalance among experts, severely impacting training efficiency. While previous research has attempted to address the load balancing challenge, existing solutions either compromise model accuracy or introduce additional system overhead. As a result, they fail to achieve fine-grained load balancing, which is crucial to optimizing training efficiency.

We propose MicroEP, a novel parallelization strategy to achieve fine-grained load balancing in MoE systems. MicroEP is capable of achieving optimal load balancing in every micro-batch through efficient token scheduling across GPUs. Furthermore, we propose MicroMoE, an efficient distributed MoE training system with MicroEP’s load balancing capabilities. Our experimental results demonstrate that MicroMoE improves the end-to-end training throughput by up to 47.6% compared with the state-of-the-art system, and almost consistently achieves optimal load balance among GPUs.

Keywords: distributed training, mixture-of-experts, load balance

1 Introduction

In recent years, transformer-based large models have fueled significant advancements across various domains, including natural language processing [7, 44, 54], computer vision [32, 48], and multimodality [4, 33]. These achievements underscore that scaling up model size is a straightforward yet fundamental approach to enhancing model performance. Nevertheless, deploying large models necessitates substantial computational resources. To tackle this, Mixture-of-Experts (MoE) proposes a sparsely-activated architecture [50], enabling the scaling up of model capacity with only a sub-linear increase in computational cost. Nowadays, MoE has gained widespread adoption in state-of-the-art

models, such as GPT-4 [38], Llama 4 [34], Gemini 2.5 [14], Grok-4 [60], Phi-3.5-MoE [1], Claude 4 [2], Qwen3 [62], and DeepSeek-V3 [9].

However, the sparse nature of MoE can indeed introduce inefficiencies during training. MoE divides a model layer into multiple *experts* and dynamically routes each input token to its *top-K* most suitable experts at runtime. This dynamic routing can result in significant load imbalances among experts [18, 26]. Furthermore, recent distributed training frameworks commonly employ *expert parallelism* (EP) to distribute experts across multiple GPUs [11]. While EP reduces memory consumption per GPU, it introduces a straggler problem: All GPUs must wait for the most heavily loaded GPU in the EP group to complete computation before proceeding, resulting in wasted computational resources and reduced system throughput. Therefore, the load imbalance issue poses a significant challenge in efficiently training MoE models.

To address the load imbalance issue of MoE, many works propose *algorithmic solutions* by modifying the routing algorithms. These approaches incorporate load-balancing loss [10, 11, 55], route tokens to less suitable experts [29, 68], or simply drop excess tokens to alleviate the burden on heavily loaded experts [11, 26, 29]. However, these modifications to model logic can degrade model accuracy and potentially impact model convergence [15, 56, 61]. As a result, they introduce a trade-off between model accuracy and system efficiency [43, 53].

Other works propose *systematic solutions* for load balancing in MoE without compromising the model accuracy. A series of works balance the GPU loads by adjusting the expert-to-GPU mapping [37, 42, 52, 56, 57, 64]. For instance, FlexMoE [37] adaptively replicates experts based on their popularity, distributing loads of more popular experts to more GPUs. However, they inevitably introduce substantial system overheads when adjusting expert placements. As a result, they introduce a trade-off between expert adjustment overhead and load balancing performance.

The aforementioned trade-offs prevent existing solutions from effectively achieving ideal load balancing, which should satisfy two requirements:

R1: Maximizing system efficiency requires fine-grained load balance among all GPUs. Giant model sizes demand numerous GPUs for training [9, 53], making even minor load imbalances costly in terms of wasted GPU time. Some algorithmic solutions can achieve perfect load balance by restricting the number of tokens to each expert [26]. However, their accuracy-balance trade-off necessitates significant compromise of model accuracy to equalize all GPU loads [13, 68]. Many systematic solutions schedule the expert layout to optimize load balancing. Typically, they treat each expert replica as a scheduling unit for communication efficiency [37, 42, 56, 64]. However, scheduling at the expert granularity results in a limited, discrete scheduling space, making it challenging to achieve optimal load balance across a vast, near-continuous space of possible load distributions.

R2: Dynamic expert load distributions require micro-batch-level adaptation at runtime. Data heterogeneity across micro-batches causes severe variations in expert loads, necessitating dynamic load balancing to optimize GPU utilization consistently. Algorithmic solutions using load-balancing loss rely on per-iteration parameter updates for adaptive adjustments [11, 55], which are too slow to handle per-microbatch load imbalances. Existing systematic solutions involve migrating expert parameters to align with the changing load distributions [37, 56, 64]. However, the substantial size of these parameters presents a significant challenge to making frequent adjustments seamlessly.

To meet these requirements, we adopt a novel systematic approach by **scheduling input tokens** to perform load balancing for MoE systems, which has two advantages: First, token scheduling allows a precise control over computational workloads. Unlike expert scheduling solutions, where all replicas of an expert have even loads [37, 56], token scheduling may allocate uneven loads to each expert replica, thus achieving more fine-grained load balance among GPUs (**R1**). Second, token scheduling leverages the pathways of existing EP communication. This minimal communication overhead allows efficient load balancing within each micro-batch (**R2**).

However, performing token scheduling in existing EP frameworks faces two challenges. **C1:** Current frameworks dispatch tokens within an EP group, which contains exactly one replica of each expert [47, 51]. As a result, the GPU loads are fixed by expert loads, which are determined by the routing algorithm. To tackle this, we schedule tokens among expert replicas in *multiple EP groups*. **C2:** Identical expert placement across EP groups severely limits the scheduling space (see Figure 3b). Since a popular expert has identical EP ranks in all EP groups, naïve token scheduling cannot balance loads across different EP ranks, resulting in only local load balance. To tackle this, we tailor the *expert placement* to expand the scheduling space and achieve global load balance.

Furthermore, we theoretically analyze the influence of expert placement on token scheduling and derive the properties of optimal expert placement with *graph theory* (§6.1).

Formally, we propose MicroEP, a novel EP strategy that leverages token scheduling for fine-grained load balancing across GPUs. MicroEP optimizes load balancing from two perspectives: (1) From a short-term perspective, MicroEP schedules tokens within every micro-batch using *linear programming* to minimize the maximum GPU loads (§5.1). To optimize system efficiency, MicroEP reduces the communication volume through *locality-aware routing* (§5.2) and minimizes the scheduling latency through efficient *overlapping* (§5.4). (2) From a long-term perspective, MicroEP tailors *expert placement* to optimize the token scheduling performance and mitigate coarse-grained imbalances. Based on our theoretical analysis, MicroEP adopts multiple expert placement strategies (§6.2-6.3) and an *adaptive replacement* mechanism (§6.4). These optimizations enable MicroEP to maintain complete load balance under dynamic and highly skewed expert loads.

Based on MicroEP, we design MicroMoE, an efficient distributed MoE training system. We implement MicroMoE upon Megatron-LM [51] and evaluate its load balancing performance. Experimental results show that MicroMoE achieves up to 47.6% improvement on end-to-end throughput compared with Megatron-LM and maintains complete load balance even under highly imbalanced workloads.

Briefly, we make the following contributions:

- We propose MicroEP to achieve dynamic and fine-grained load balancing among GPUs with token scheduling (§4).
- We formulate the load balancing problem into a linear programming problem and solve it efficiently (§5).
- We propose several expert placement strategies and an adaptive replacement mechanism to optimize MicroEP’s load balancing capacity (§6).
- We implement MicroMoE, an efficient distributed MoE training system based on MicroEP. We conduct comprehensive experiments to demonstrate that MicroMoE exhibits superior system efficiency and load balancing capability (§7).

2 Background

2.1 Transformer and MoE

Transformer is the state-of-the-art architecture for deep learning models [54]. A transformer model consists of multiple sequential blocks, each comprising an attention layer and a feed-forward network (FFN). Figure 1a illustrates the architecture of a standard (dense) transformer block.

Figure 1b illustrates an MoE transformer block [11]. MoE replaces the FFN with multiple *expert* FFNs, coordinated by a *gate* network. After the attention computation, the gate network assigns each token to its *top-K* most suitable experts.

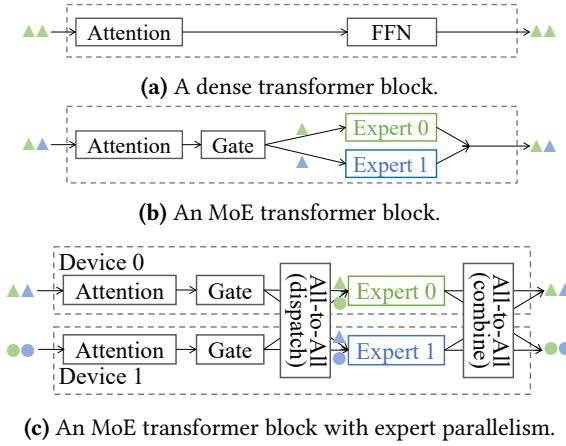


Figure 1. An example of transformer, MoE, and expert parallelism.

Then, these selected experts process the token and aggregate their results to produce the final result.

2.2 Expert Parallelism and Expert Data Parallelism

The increasing scale of deep learning models has necessitated the development of various parallelization strategies for distributed training across multiple computational devices. We introduce some of them in this section.

Data parallelism (DP) is a fundamental parallelization strategy that partitions input data across devices to enhance training throughput. In the forward pass, each DP rank (e.g., a device) maintains a complete copy of the model parameters while processing distinct *micro-batches* of data independently. In the backward pass, all devices in a DP group aggregate their gradients for parameter synchronization. Recent advances, such as ZeRO [46], have further optimized DP’s memory efficiency by eliminating memory redundancies across devices.

Tensor parallelism (TP) addresses per-device memory constraints by partitioning model parameter tensors across devices. Each TP rank processes complete inputs using its allocated slice of model parameters, with results aggregated within the TP group to complete the computation.

Pipeline parallelism (PP) addresses memory constraints by partitioning model layers [20, 35]. While PP introduces lower communication overhead compared with TP, it wastes some computational resources due to pipeline bubbles [20, 41].

Expert parallelism (EP) is a widely employed technique in distributed MoE systems [11]. EP is a specialized hybrid parallelization strategy combining elements of DP and TP. By exploiting the inherent sparsity of MoE, EP benefits from both DP’s training efficiency and TP’s memory efficiency. As shown in Figure 1c, EP partitions the FFN layer by distributing distinct experts across multiple devices. At runtime, the attention layer and the gate network utilize conventional DP

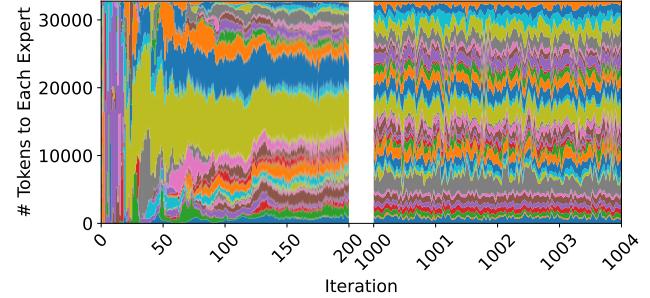


Figure 2. Expert load distribution of GPT 32×1.3B layer 20 in some training iterations.

to process different tokens on separate devices. Next, devices within the EP group perform an all-to-all (A2A) communication operation to dispatch tokens towards their designated expert devices. Then, each device processes its received tokens with its local experts. Finally, the devices perform another all-to-all operation to return tokens to their original devices.

Expert data parallelism (EDP) is a specialized form of DP emerging from EP. In configurations where the DP degree (number of DP ranks) exceeds the EP degree, systems must employ EDP across multiple EP groups within a DP group. Each EDP group consists of devices sharing the same EP rank across different EP groups. Devices within an EDP group maintain replicated instances of identical experts while processing distinct tokens from different EP groups. The synchronization of parameters and gradients among these expert replicas follows conventional DP mechanisms, ensuring consistency across the distributed system.

2.3 Load Imbalances in MoE

The dynamic token-to-expert assignment in MoE systems inherently results in significant load imbalances across experts, presenting a critical performance bottleneck for distributed training. We conduct an experiment to trace the expert load distribution in a training process with model configurations detailed in Tabel 2. As illustrated in the left part of Figure 2, **the expert load distribution exhibits substantial variability and skewness**, particularly during the initial training iterations.

The imbalance issue severely impacts resource utilization and training throughput in MoE systems with EP. In current MoE implementations, the FFN computation time of a GPU is approximately proportional to the total number of tokens assigned to the experts on this GPU [13]. Since GPUs within an EP group must synchronize through all-to-all communication both before and after FFN computations, all GPUs must wait until the slowest GPU completes computation before proceeding. Consequently, the system’s overall throughput is bottlenecked by the most heavily loaded device, namely

the *straggler*. This strict synchronization requirement emphasizes the necessity of effective load balancing strategies to optimize MoE training efficiency.

Moreover, the straggler effect continuously degrades training efficiency in *every micro-batch*. As shown in the right part of Figure 2, **expert load distribution fluctuates significantly between consecutive micro-batches**. This dynamic nature of load imbalances necessitates fine-grained load balancing at the micro-batch level to maintain optimal training efficiency.

3 Analysis of Token Scheduling

3.1 Benefits of Token Scheduling

Our goal is to achieve fine-grained load balancing for MoE systems within every micro-batch. We aim to find a systematic approach without compromising model accuracy. In MoE, FFN computation requires two key elements: expert parameters and input tokens. Therefore, we can perform load balancing through scheduling either of them.

We adopt *token scheduling* rather than expert parameter scheduling for two key reasons: (1) Token scheduling provides fine-grained control of GPU loads, which is necessary for optimal load balance (**R1**). In contrast, expert scheduling only provides coarse-grained load balancing due to the significantly smaller number of experts compared with tokens. (2) Token scheduling incurs minimal overhead, enabling efficient scheduling within every micro-batch (**R2**). Since EP already involves token dispatching, we can piggyback token arrangements on existing all-to-all operations **without extra communication overhead**. In contrast, expert scheduling requires costly parameter migration across devices, making it inefficient to execute in each micro-batch. These advantages make tokens, rather than expert parameters, a more practical target for scheduling optimization.

3.2 Challenges in Token Scheduling

We adopt a *token scheduling* approach to balance loads across GPUs. However, this approach poses two fundamental challenges:

Challenge 1: *Vanilla EP restricts token dispatching within each EP group, providing no space for scheduling.*

We aim to balance GPU loads by scheduling input tokens. To achieve this, we must determine the *scheduling space*. Specifically, when a token is assigned to an expert, we must be able to compute it on one of *multiple* GPUs.

Unfortunately, the scheduling space is unattainable in the current EP paradigm. As shown in Figure 3a, vanilla EP restricts token dispatching within the scope of an EP group, where each EP group contains exactly one replica of each expert’s parameters. When a token is assigned to an expert, it must be computed on the GPU hosting that expert’s replica in its EP group. Consequently, the token-to-GPU mapping is

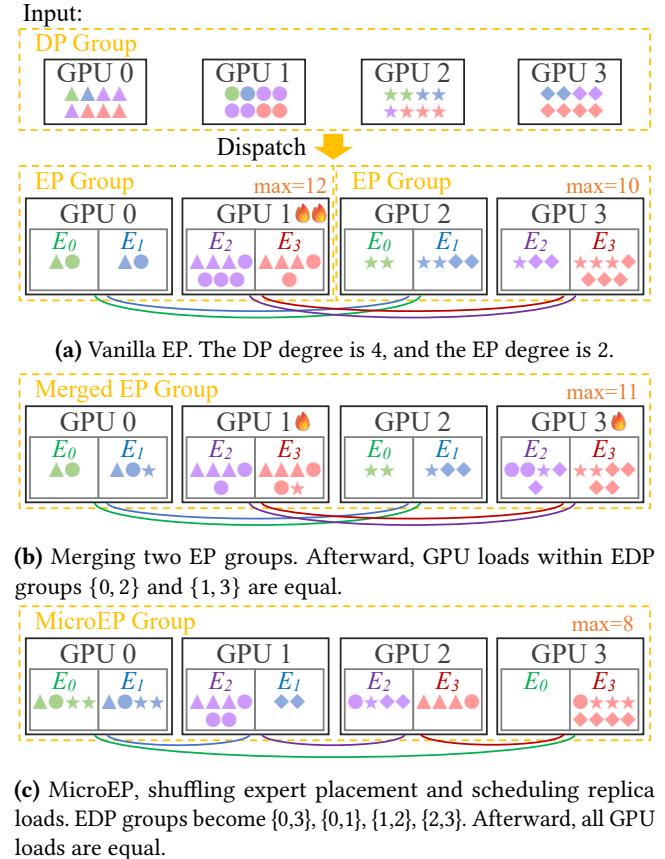


Figure 3. Converting EP to MicroEP. The shape and color of a symbol indicate the source GPU and the assigned expert of a token. The bottom curves indicate EDP groups.

fixed by the token-to-expert mapping, eliminating any space for computation scheduling.

Solution 1: Our key observation is that we can find scheduling space through *expert data parallelism*. We call the set of GPUs that host replicas of expert e as the EDP group of expert e . Since all replicas of an expert maintain identical parameters, a token can be computed equivalently with any replica in its designated expert’s EDP group. This observation inspires us to merge multiple EP groups and schedule tokens across their experts’ EDP groups, as shown in Figure 3b.

Challenge 2: Identical expert placement across EP groups results in constrained scheduling space.

While this approach improves token distribution, it falls short of optimal load balancing. The fundamental limitation lies in the restricted scheduling space—load balancing occurs only within individual EDP groups. In the current EP paradigm, each EP group maintains identical expert placement. Consequently, the EDP groups of different experts are either completely disjoint or identical (e.g., in Figure 3b, the EDP groups of expert 0,1 are both GPU $\{0, 2\}$, while the EDP groups of expert 2,3 are both GPU $\{1, 3\}$). This constraint indicates

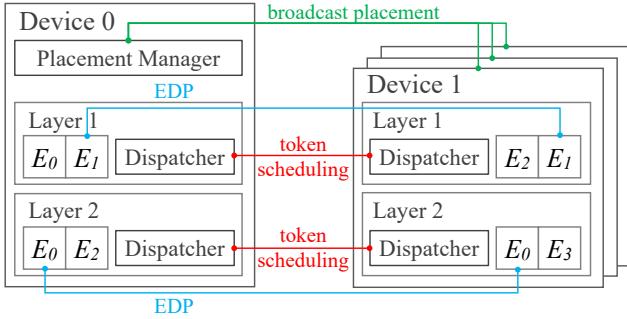


Figure 4. MicroMoE architecture.

that we can at most equalize workloads within each EDP group, while significant load imbalances may persist across different EDP groups.

Solution 2: To expand the scheduling space, we **shuffle the expert placement** within the merged EP groups. This operation creates **intersecting EDP groups** across different experts, substantially enlarging the scheduling space. As illustrated in Figure 3c, this approach can potentially achieve perfect load balance across all GPUs.

4 MicroMoE Overview

We introduce a novel expert parallelism strategy, namely MicroEP. MicroEP optimizes load balancing across GPUs within *every micro-batch* through *token scheduling* across *EDP groups*.

Based on MicroEP, we propose MicroMoE, an efficient distributed MoE training system. Figure 4 illustrates the architecture of MicroMoE. Following PP, each device hosts multiple model layers. Each MoE layer consists of a *token dispatcher* and multiple local expert replicas. Expert replica placement is coordinated by a global *placement manager* residing on device 0. We briefly illustrate the workflow of MicroMoE as follows.

Prerequisites. MicroEP obtains scheduling space from intersecting EDP groups, which require two conditions: the DP degree must exceed the EP degree, and each GPU must host multiple expert replicas. We find that these requirements are satisfied in many open-sourced MoE models [9, 53]. For example, DeepSeek-V3’s pre-training configuration employs 64-way EP and 128-way DP, with each GPU hosting 8 experts [9].

Configuration. MicroEP introduces an integer parameter d , where $1 < d \leq \frac{\text{DP_degree}}{\text{EP_degree}}$. MicroEP merges every d EP groups into a MicroEP group. In the following sections, we consider the simplest case where $d = \frac{\text{DP_degree}}{\text{EP_degree}}$, i.e., the entire DP group forms a single MicroEP group.

Initialization. During model initialization, the placement manager generates expert placements within all MicroEP groups and broadcasts the placements to all devices. The placement strategies are detailed in §6.

Runtime. During model training, MicroEP balances GPU workloads by scheduling tokens within the MicroEP group. Specifically, after the gate networks assign tokens to specific experts, the dispatchers route tokens to particular expert replicas within their respective EDP groups. The scheduling algorithm is detailed in §5.

After scheduling, the dispatchers initiate an all-to-all (dispatch) operation within the MicroEP group, sending tokens to their routed expert replicas. Then, each GPU executes FFN computation to process received tokens using its local expert replicas. Afterward, a second all-to-all (combine) operation returns processed tokens to their source GPUs. These operations follow the same patterns as vanilla EP, but occur in a communication group d times larger.

An example. Surprisingly, MicroEP can almost always achieve perfect load balance among GPUs, unless the expert load distribution is extremely skewed (as shown in §7.3). We can intuitively learn how MicroEP achieves such good performance from the example in Figure 3c. Since expert 3 is heavily loaded, it places 8 tokens on GPU 3, preventing expert 0 from placing any token on the same GPU. Fortunately, expert 0 can place all its 4 tokens on GPU 0. Similarly, expert 2 also “shifts” some load from GPU 2 to GPU 1, while expert 1 shifts some load from GPU 1 to GPU 0. Consequently, the computational burden of heavily loaded experts is effectively distributed across the entire MicroEP group.

5 Token Scheduling

Token scheduling is the core of MicroEP for fine-grained load balancing. The scheduling algorithm runs in two steps: The first step is to distribute expert loads among their replicas, balancing GPU loads (§5.1). The second step is to route tokens to specific expert replicas, enforcing the calculated replica loads (§5.2). Then, we discuss where and when to launch the scheduling in §5.3 and §5.4. Finally, we discuss some additional scheduling optimizations in Appendix A.

5.1 Determining Replica Loads

We aim to balance GPU loads through optimal distribution of expert loads among their replicas. We formulate the load balancing problem as an optimization problem.

- **Notations.** Let E represent the set of all experts, G_{MicroEP} represent the set of GPUs in the concerned MicroEP group, and G_{EDP}^e represent the set of GPUs in the EDP group of expert e . Since we currently focus on a single MicroEP group, we assume that $G_{\text{EDP}}^e \subseteq G_{\text{MicroEP}}, \forall e \in E$.
- **Variables.** The variables are $\{x_e^g : e \in E, g \in G_{\text{EDP}}^e\}$, where x_e^g is the replica load of expert e on GPU g .
- **Constraints.** Let $\text{load}_e(e \in E)$ denote the total load (number of tokens) of expert e in the concerned MicroEP group. A valid distribution must ensure that

Table 1. List of Symbols and Notations.

Symbol	Description
E	set of all experts
$G_{MicroEP}$	set of GPUs in the concerned MicroEP group
d	a parameter = $\frac{ G_{MicroEP} }{EP_degree}$
G_{EDP}^e	EDP group of expert e
$load_e$	total load of expert e
x_e^g	replica load of expert e on GPU g
$input_e^g$	input load of expert e from GPU g
m	optimal objective value of LPP 1

each expert distributes its total load across its replicas.

- *Objective.* Due to synchronization requirements for all-to-all communication before and after expert computation, the GPU with the highest load becomes the performance bottleneck. Therefore, we should minimize the maximum GPU load across the MicroEP group.

We formulate the optimization problem as follows.

$$\begin{aligned} & \text{minimize}_{g \in G_{MicroEP}} \left\{ \sum_{e \in E: g \in G_{EDP}^e} x_e^g \right\}, \\ & \text{subject to } \sum_{g \in G_{EDP}^e} x_e^g = load_e, \quad \forall e \in E, \\ & \quad x_e^g \geq 0, \quad \forall e \in E, g \in G_{EDP}^e. \end{aligned} \quad (1)$$

Problem 1 is a *linear programming problem* (LPP), which can be efficiently solved in polynomial time relative to the number of GPUs and experts. The number of variables is $O(|E|d)$, and the number of constraints is $O(|E| + |G_{MicroEP}|)$. Given its modest scale, we solve this LPP using a single CPU thread with the HiGHs solver [21]. GPU acceleration or multi-threading would not yield significant performance benefits for this scale of optimization.

Notably, across different micro-batches, while the constraint matrix (determined by expert placement G_{EDP}^e) remains the same, only the constraint bounds ($load_e$) vary. This property enables the *warm-start* of the LPP solving by reusing the immediate states of the previous solution, significantly reducing optimization overhead.

5.2 Routing Tokens to Replicas

Our next step is routing tokens to specific expert replicas, enforcing the calculated replica loads (x_e^g). Specifically, we determine this *token-to-replica* mapping using a sequential routing strategy: First, we arrange tokens assigned to expert e from all GPUs in sequence, along with an ordered list of expert e 's replicas. Then, we iterate through these tokens, routing each to the first replica that has not yet reached its allocated load x_e^g .

Algorithm 1: Routing tokens to expert replicas.

```

Input: { $input_e^g$ }, { $x_e^g$ }
1  $\{remain\_input_e^g\} = \{input_e^g\}$ 
2  $\{remain\_x_e^g\} = \{x_e^g\}$ 
3 for  $e \in E$  do
4   // First, route local tokens to local replicas.
5   for  $g \in G_{EDP}^e$  do
6      $y = \min(remain\_input_e^g, remain\_x_e^g)$ 
7     route the next  $y$  tokens of expert  $e$  from GPU
       $g$  to the replica on GPU  $g$ 
8      $remain\_input_e^g \leftarrow remain\_input_e^g - y$ 
9      $remain\_x_e^g \leftarrow remain\_x_e^g - y$ 
10  // Then, route global tokens to global replicas.
11  for  $g \in G_{MicroEP}$  do
12    for  $g' \in G_{EDP}^e$  do
13       $y = \min(remain\_input_e^g, remain\_x_e^g)$ 
14      route the next  $y$  tokens of expert  $e$  from
        GPU  $g$  to the replica on GPU  $g'$ 
15       $remain\_input_e^g \leftarrow remain\_input_e^g - y$ 
16       $remain\_x_e^g \leftarrow remain\_x_e^g - y$ 

```

To enhance efficiency, we can manipulate token ranges rather than individual tokens. Let $input_e^g$ denote the input load of expert e from GPU g , i.e., the number of tokens on GPU g assigned to expert e ($\sum_{g \in G_{MicroEP}} input_e^g = load_e$). The pseudo code of token routing is shown in Algorithm 1, Lines 10-16.

Locality-aware routing. Previous research has underscored that all-to-all communication introduces significant overhead to MoE systems [18, 22, 28]. To reduce the all-to-all communication volume, we can leverage *data locality* during token routing. Specifically, when GPU g holds a replica of expert e , we prioritize routing tokens from GPU g to its local expert replica before considering remote replicas. The pseudo code is shown in Algorithm 1, Lines 4-9. Furthermore, we can consider communication overhead during optimization, which is detailed in Appendix A.1.

5.3 Distributed Scheduling across Devices

As described in §5.1 and §5.2, the scheduling algorithm requires the global load information ($input_e^g$) across the entire MicroEP group as input. This raises the question of where to place the scheduler.

We consider two candidate locations for the scheduler: *centralized* on one device or *distributed* across all devices. A centralized scheduler needs to gather load information from all devices, perform the scheduling, and scatter the results. Alternatively, distributed schedulers require all devices to perform an all-gather operation to collect global load information and execute the scheduling algorithm independently.

We choose the distributed approach for better scalability because it requires only one communication operation compared to two operations in the centralized approach. Since the load information is small, the primary performance factor is latency rather than throughput, making fewer communication operations advantageous.

Notably, the distributed scheduling maintains consistency because MicroEP’s scheduling algorithm is deterministic—all devices use identical inputs, follow identical algorithms, and produce identical results.

5.4 Overlapping Scheduling to Hide Latency

Since we need to execute token scheduling in every microbatch, minimizing the scheduling overhead is significant for system efficiency. Therefore, we reduce the scheduling overhead by *overlapping* scheduling with other operations.

As described in §4, the scheduling executes immediately after the gate network and before the all-to-all communication. We observe that existing distributed training frameworks usually perform some operations during this period. For instance, Megatron-LM [51] executes a token permutation operation to replicate tokens by top-K times and sort tokens by expert indices before all-to-all dispatching. Therefore, we can overlap the scheduling on CPUs with the permutation on GPUs. For frameworks without suitable overlapping operations, we propose a pipeline mechanism to hide scheduling latency, which is detailed in Appendix A.2.

6 Expert Placement

Expert placement is fundamental to the load balancing performance of token scheduling in MicroEP. As shown in LPP 1, the expert placement determines the EDP groups (G_{EDP}^e), which are the key components of the constraint matrix and thus significantly affect the optimization result. Although our experiments demonstrate that some simple placements, such as random shuffling, usually provide satisfactory results (§7.3), identifying optimal placement strategies is still crucial for maximizing system performance. In the following sections, we first analyze “*What is an optimal expert placement?*” in §6.1 and then discuss “*How to construct an optimal expert placement?*” in §6.2 and §6.3. Finally, we propose an *adaptive replacement* mechanism in §6.4.

6.1 Analysis of Optimal Expert Placement

Analysis of the LPP solution. Since the optimal expert placement should minimize the objective value of LPP 1, we begin by analyzing the solution of this problem. Let m denote the optimal objective value of LPP 1, which represents the minimized maximum load across all GPUs.

Lemma. There exists an optimal solution to LPP 1 that satisfies: Let G_{max} denote the set of GPUs with load m in this solution. An expert distributes no load within G_{max} unless its EDP group is entirely in G_{max} .

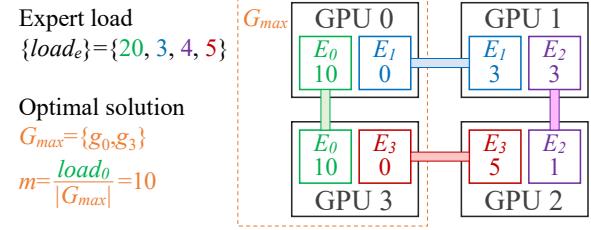


Figure 5. The graph abstraction of an example expert placement. Color bars indicate edges (experts) between vertices (GPUs). G_{max} contains GPU 0, 3. Expert 0 is entirely in G_{max} . Experts 1,3 partially intersect with G_{max} and cannot distribute any load within G_{max} .

Proof. Among all optimal solutions, we choose the solution with the minimal $|G_{max}|$ (number of GPUs with load m). Consider experts whose EDP groups partially intersect with G_{max} . Suppose that the EDP group of expert e consists of two GPUs $g_0 \in G_{max}$ and $g_1 \notin G_{max}$. If expert e did distribute any load on GPU g_0 , it could move a tiny amount of load from g_0 to g_1 . Afterward, the loads of both g_0 and g_1 would be less than m , yielding a new optimal solution with smaller $|G_{max}|$. The new solution would contradict our choice of the solution with the minimal $|G_{max}|$. Therefore, expert e cannot distribute any load on GPU g_0 . \square

Figure 5 shows an example of the above lemma. From this lemma, we can derive an equation by calculating the total loads within G_{max} in two ways:

$$m \cdot |G_{max}| = \sum_{e \in E: G_{EDP}^e \subseteq G_{max}} load_e \quad (2)$$

From Equation 2, we can derive an equation to calculate m by enumerating every possible G_{max} ¹:

$$m = \max_{G_{max} \subseteq G_{MicroEP}} \left\{ \frac{1}{|G_{max}|} \sum_{e \in E: G_{EDP}^e \subseteq G_{max}} load_e \right\} \quad (3)$$

Equation 3 provides a mathematical approach to determine the optimal objective value of LPP 1 without actually executing the LPP solving.

Graph abstraction of expert placement. Furthermore, we formulate expert placement with graph theory. Let each GPU $g \in G_{MicroEP}$ represent a vertex, and each expert $e \in E$ represent a hyperedge connecting all GPUs in G_{EDP}^e . Hence, an expert placement can be represented by an undirected hypergraph², denoted by $\mathcal{G}(G_{MicroEP}, \{G_{EDP}^e : e \in E\})$. Figure 5 shows an example of the graph and the optimal solution corresponding to an expert placement.

Now, we can explain Equation 3 from the perspective of graph theory: We assign the weight of the edge (expert) e

¹We omit the derivation from Equation 2 to Equation 3.

²When $d = 2$, the hypergraph is a conventional graph. For simplicity, we omit the prefix “hyper” in the rest of this paper.

as $load_e$. We define the *density* of a weighted graph as the sum of all edge weights divided by the number of vertices³. Consider the subgraph in \mathcal{G} induced by G_{max} . The edges in the induced subgraph represent the experts whose EDP groups are entirely in G_{max} ($\{e \in E : G_{EDP}^e \subseteq G_{max}\}$). According to Equation 3, the optimal objective value m is the maximum density across all induced subgraphs of graph \mathcal{G} . Since the goal of expert placement is to minimize m , we can characterize the optimal expert placement as follows:

Property of the optimal expert placement. *The optimal expert placement is the graph whose maximum induced subgraph density is minimal.*

Having understood the property of the optimal expert placement, our next challenge is how to find such a placement. We distinguish between two scenarios based on our knowledge of the expert load $load_e$ in §6.2 and §6.3.

6.2 Symmetric Placement without Real Expert Loads

If we have no prior knowledge of the real expert load distribution, we can construct *symmetric placements*, treating all experts equally. In such scenarios, we can assume that all expert loads follow an independent and identically distributed (i.i.d.) pattern. Thereby, the problem becomes: Given the number of vertices and edges, and the identical distribution of edge weights, how to construct a graph that minimizes the expectation of the maximum induced subgraph density?

We recognize the challenge of the above problem due to the vast space of possible graphs and distributions. Nevertheless, we propose a near-optimal symmetric placement strategy for many practical configurations using *Cayley graphs* [58]. Our intuition is that the inherent symmetry of Cayley graphs ensures a balanced distribution of edges across vertices, preventing some induced subgraphs from having significantly larger density than others. Since Cayley graphs involve complex group theory, we illustrate our construction method in Appendix B. We demonstrate the effectiveness of our constructions through both example studies (Appendix B.2) and experimental evaluations (§7.3).

6.3 Asymmetric Placement with Real Expert Loads

If we know real expert load distributions in advance, we can construct *asymmetric placements* based on specific distributions. Unlike symmetric scenarios, we can vary both replica counts and replica locations across different experts for better load balancing, similar to previous works [37, 56, 64].

We adopt an empirical strategy to construct a near-optimal asymmetric expert placement in two steps: (1) First, we determine the number of replicas for each expert with a *greedy* algorithm: We maintain a heap of experts sorted by load-per-replica, and iteratively allocate remaining replicas to the expert with the maximum load-per-replica. (2) Second, we

³This definition may differ from conventional graph density definitions in the literature.

determine the placement of expert replicas across GPUs with *Monte Carlo sampling*: We randomly generate many placement graphs, and choose the one whose maximum induced subgraph density is minimal.

6.4 Adaptive Replacement

Based on the symmetric and asymmetric placements, we further propose an *adaptive replacement* (AR) mechanism for MicroMoE to optimize performance under dynamic expert loads.

Relationship between token scheduling and adaptive replacement. The adaptive replacement mechanism complements the token scheduling in §5 by addressing different levels of load balancing. Token scheduling performs *transient, fine-grained* load balancing through per-micro-batch token arrangement, while adaptive replacement handles *long-term, coarse-grained* load imbalances through periodic expert arrangement.

Specifically, for moderately imbalanced workloads, token scheduling effectively maintains optimal balance with static, symmetric placements (as shown in §7.3). For highly skewed workloads, MicroMoE adopts asymmetric placements to mitigate coarse-grained imbalances before using token scheduling for fine-grained optimization. Since asymmetric placements require real-time expert loads, MicroMoE adopts adaptive replacement to monitor expert load distributions and adjust placements when significant distributional shifts are detected.

Implementation of adaptive replacement. We implement the adaptive replacement mechanism in MicroMoE using the *placement manager* (according to Figure 4). (1) During model initialization, the placement manager initializes the model states of all devices using the symmetric placement strategy, providing conservative and general load balancing capabilities. (2) During training, the placement manager monitors expert load information within each micro-batch in the background. (3) For every few iterations, the placement manager predicts future load distributions using historical data with time series analysis techniques, such as moving averages [8]. Then, it evaluates the performance of current placements on future distributions using Equation 3. If the future performance drops below a specific threshold, the placement manager generates new optimal asymmetric placements and re-initializes global model states accordingly.

Difference between MicroMoE’s adaptive replacement and expert scheduling solutions. The system implementation of MicroMoE’s adaptive replacement is similar to existing expert scheduling solutions, such as FlexMoE [37] and SmartMoE [64]. Nevertheless, their design goals and algorithms are fundamentally different:

Design goals: In FlexMoE, changing expert placement is the only means for load balancing. However, in MicroMoE, the primary weapon is token scheduling, while adaptive replacement is a further optimization to token scheduling.

Even with static placement, MicroMoE can still achieve good load balancing performance at the micro-batch granularity.

Algorithms: The different design goals of FlexMoE and MicroMoE lead to different placement algorithms. In FlexMoE, all replicas of an expert have identical loads, i.e., the replica load is simply calculated as the expert load divided by the replica count. In contrast, the replica loads in MicroMoE are determined by the LPP 1. Therefore, MicroMoE requires the graph theory in §6.1 to guide the placement strategy.

7 Evaluation

7.1 Experimental Setup

Testbed. Our testbed consists of 4 nodes, each equipped with 8 NVIDIA H100 80GB SXM GPUs connected via 900 GBps NVLink. Nodes are interconnected using two 400 Gbps Infiniband NICs per node.

Models. We use GPT [3] and Mixtral [23] models for evaluation. We use GPT 32×1.3B to represent an MoE model converted from a 1.3B dense GPT model with 32 experts. We pretrain these models with the Wikipedia dataset [12]. A detailed list of model hyperparameters is provided in Appendix C.1.

Baselines. We compare MicroMoE with four baseline systems.

- Megatron-LM [51] is a popular distributed training framework for large language models (LLMs). It supports various parallelism strategies, as well as state-of-the-art optimizations [6].
- SmartMoE [64] balances GPU loads by adjusting the expert placement within EP groups.
- FlexMoE [37] achieves load balancing by dynamically adjusting replica counts based on expert loads. FlexMoE places expert replicas across the entire DP group, similar to MicroMoE’s asymmetric placement (§6.3).
- DeepSpeed [47] is a high-performance distributed framework for both LLM training and inference. We enable ZeRO-1 [46] optimization in DeepSpeed (currently, ZeRO-2 does not support PP, and Tutel [22] does not support top-K>1).

We compare two variants of MicroMoE: “MicroMoE (w/o AR)” uses static, symmetric placement (§6.2), while “MicroMoE” uses adaptive, asymmetric placement (§6.3-6.4).

Implementation. We implement MicroMoE upon Megatron-LM [51]. MicroMoE provides a model wrapper similar to Pytorch’s Distributed Data Parallel (DDP) [30], enabling users to benefit from MicroEP’s fine-grained load balancing capabilities within their training jobs. We modify Megatron-LM, including its MoELayer and DDP with Python, and implement the token scheduling algorithm in MicroEP with C++.

For fair comparison, we also implement SmartMoE and FlexMoE in Megatron-LM, as SmartMoE’s repository is outdated (last commit in 2023) [64], and FlexMoE is not open-sourced [37].

Parallelization configurations. Due to the limited inter-node network bandwidth in our testbed, we only employ PP for inter-node parallelism. Specifically, we set the PP degree to the number of nodes used, and the DP degree to 8. We set the EP degree to 4, resulting in 2 EP groups per DP group. We set the parameter d in MicroEP to 2, resulting in a single MicroEP group per DP group. We disable TP due to its high communication overhead.

Other configurations. We use a small auxiliary loss (listed in the appendix) to prevent extreme load imbalance from degrading model accuracy. We enable the distributed optimizers in Megatron-LM, which resembles DeepSpeed’s ZeRO-1 [46]. We disable the token dropping mechanism introduced by GShard [26]. We use BF16 precision.

7.2 End-to-end Performance

Figure 6 shows the end-to-end performance of all systems, varying models and number of GPUs. DeepSpeed exhibits poor performance with 16 or 32 experts. This is because DeepSpeed always adopts a padding mechanism, padding the load of each expert to the maximum expert load [26]. This mechanism wastes significant time and memory when expert loads are highly imbalanced. With as few as 8 experts, the inefficiency of padding becomes less significant, allowing DeepSpeed to outperform Megatron-LM due to its system-level optimizations.

Comparing SmartMoE and MicroMoE (w/o AR), where experts have uniform replica counts, MicroMoE (w/o AR) exhibits superior performance, thanks to MicroEP’s fine-grained token scheduling. While SmartMoE attempts to optimize performance by changing expert placement for load balancing among GPUs, it sometimes performs worse than vanilla Megatron-LM. This is because SmartMoE optimizes expert placement based on long-term expert load distributions. However, expert loads are highly dynamic during training, and SmartMoE’s long-term optimal placement may be sub-optimal for individual micro-batches.

Comparing FlexMoE and MicroMoE (with AR), where experts have varied replica counts, MicroMoE exhibits superior performance, thanks to MicroEP. **Compared with Megatron-LM, MicroMoE improves the end-to-end throughput by up to 47.6%, with an average improvement of 36.9%.** The average performance improvement of MicroMoE surpasses FlexMoE, the second-best system, by 13.9%.

7.3 Load Balancing Capability

We evaluate the load balancing capabilities of SmartMoE, FlexMoE, and MicroMoE with skewed expert loads. We generate expert loads following a Zipfian distribution with skewness s , where the probability of a token being assigned to the i -th most loaded expert is proportional to i^{-s} .

Figure 7 shows the load balancing performance of different systems across varied skewness s . SmartMoE’s maximum

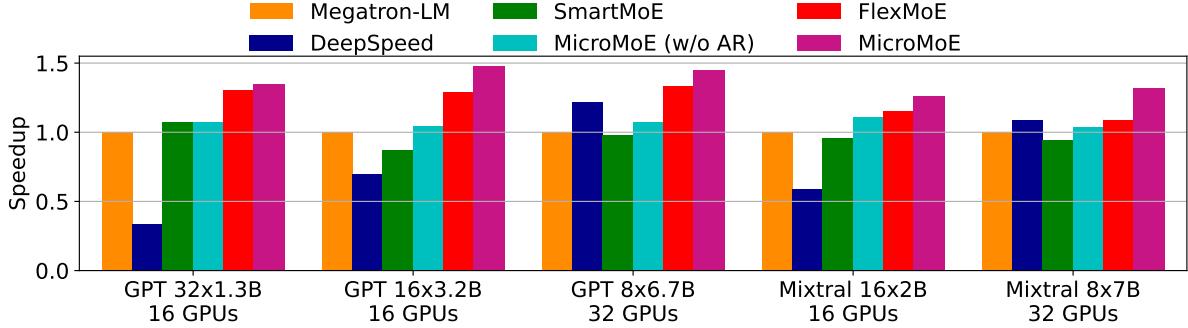


Figure 6. End-to-end speedup of different systems compared with Megatron-LM.

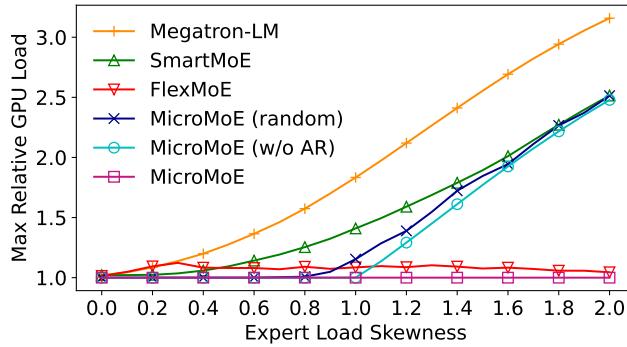


Figure 7. Max GPU load normalized by average GPU load, varying skewness of expert loads. DP_degree=8, num_experts=32.

GPU load increases as load skewness increases. While FlexMoE maintains relatively balanced GPU loads by adjusting expert replica counts, it falls short of achieving optimal load balance due to its lack of fine-grained dynamicity. MicroMoE (random) represents MicroMoE with pure random placement, which performs slightly worse than MicroMoE (w/o AR) with symmetric placement. MicroMoE (w/o AR) achieves perfect load balance when $s < 1$ thanks to MicroEP’s fine-grained token scheduling. Nonetheless, its performance degrades when $s > 1$ as uniform replica counts are insufficient for severe imbalances. MicroMoE with asymmetric placements can always achieve perfect load balance, due to the combination of both coarse-grained expert replacement and fine-grained token scheduling. Overall, **MicroMoE exhibits the best load balancing capability among all systems and consistently achieves optimal load balance.**

7.4 Execution Time Breakdown

Figure 8 shows the execution time breakdown of an MoE layer across different systems. We omit DeepSpeed due to its poor performance. For all remaining systems, the primary bottleneck is expert computation time. MicroMoE achieves the shortest computation time by maintaining perfect load

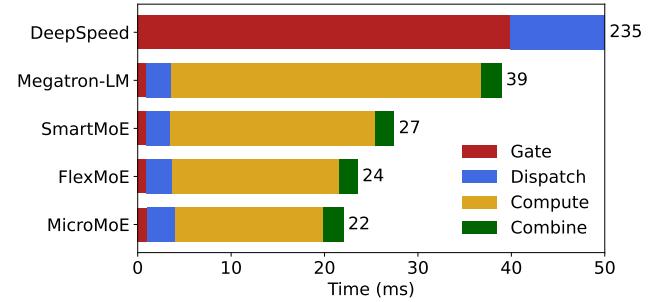


Figure 8. Execution time breakdown of an MoE layer. DP_degree=8, num_experts=32, micro_batch_size=8, sequence_length=2048, topK=2, hidden_size=4096, skewness s=1.

balance (with either symmetric or asymmetric placement, as shown in Figure 7).

Specifically, the dispatch time consists of two primary components: (1) Preparation time, which includes the all-gather operation of expert load information and the scheduling of MicroEP. While MicroMoE introduces additional overhead in dispatch time due to token scheduling operations, we effectively minimize this impact through overlapping with other operations in Megatron-LM. (2) All-to-all communication time. Each all-to-all communication in dispatch and combine requires approximately 1.3 ms in Megatron-LM. We further evaluate dispatch time with DeepEP [9], a high-performance all-to-all backend. The results are presented in Appendix C.2.

7.5 Overhead Analysis

Scheduling Overhead. We evaluate the scheduling overhead of MicroEP, including the LPP solving time and token routing time. Our evaluation reveals that the LPP solving time is the dominant factor, which scales with the number of experts and GPUs. As shown in Figure 9, the scheduling overhead remains remarkably low, with a minimum time of approximately 100 us. Even with 64 GPUs and 256 experts, the scheduling time remains under 1 ms. This minimal

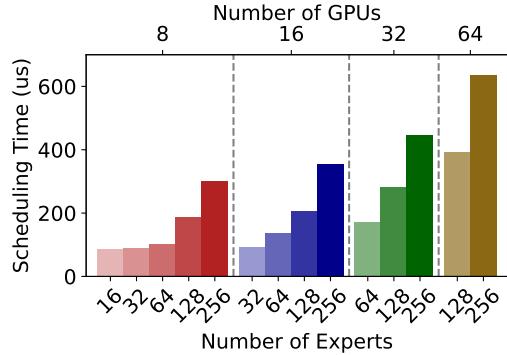


Figure 9. Scheduling time for MicroEP, varying number of experts and GPUs.

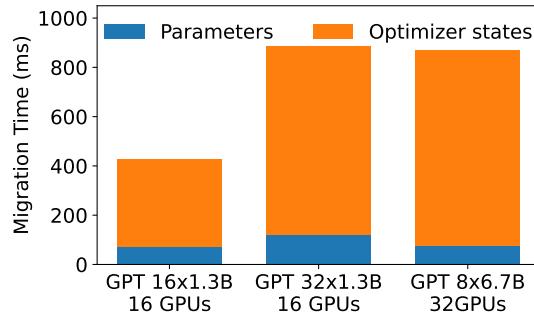


Figure 10. Migration time for adaptive replacement of MicroMoE.

overhead per micro-batch enables MicroEP to maintain high training throughput while providing load balancing benefits. **Replacement overhead.** The adaptive replacement strategy in MicroMoE necessitates model re-initialization to transition to new configurations. Although replacement is beneficial for load balancing, model re-initialization causes temporary suspension of training. Our evaluation focuses on two key components in the replacement overhead: the migration time of expert parameters and their optimizer states. As shown in Figure 10, the total migration time typically spans hundreds of milliseconds across different model configurations. These results emphasize the importance of carefully selecting the expert replacement frequency to optimize the trade-off between per-micro-batch training efficiency and overall replacement overhead.

7.6 Ablation Study

We conducted an ablation study to evaluate three optimizations in MicroMoE: (1) warm solving in §5.1, (2) locality-aware routing in §5.2, and (3) overlapping in §5.4. All these optimizations aim to reduce the dispatch time. As shown in Figure 11, both warm solving and overlapping reduce scheduling time, while locality-aware routing reduces all-to-all communication time. With the combination of optimizations, MicroMoE introduces only 0.4 ms additional dispatch time

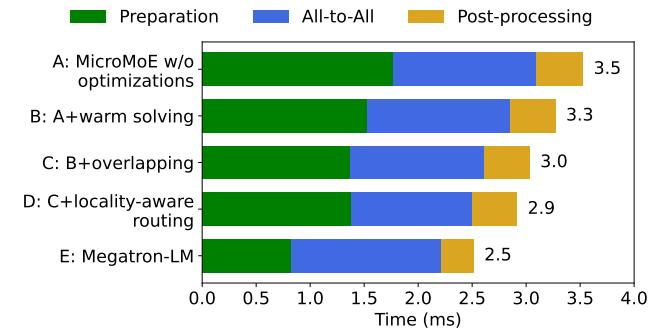


Figure 11. Ablation study of dispatch time, with the same setting as Figure 8.

compared to vanilla Megatron-LM. This modest overhead is substantially outweighed by the reduced computation time, ensuring the overall system efficiency of MicroMoE.

8 Related Work

MoE training systems. Switch Transformer [11] pioneers expert parallelism for transformer-based models. FairSeq [39] and FastMoE [17] provide PyTorch [40] plugins to optimize MoE training. DeepSpeed-MoE [45] combines EP with various optimizations, such as ZeRO [46] for large-scale MoE training. Tutel [22] dynamically adapts parallelism strategies and pipeline degrees to handle dynamic workloads of MoE.

MoE load balancing. Many studies propose algorithmic solutions to address the load imbalance issue of MoE. Shazeer et al. [50] introduce a load-balancing loss, which is further refined by subsequent works [5, 11, 16, 29, 55, 59]. GShard [26] introduces the concept of expert capacity to limit the number of tokens assigned to each expert, which is adopted by later works [11, 22, 29]. Some studies propose alternative gating algorithms in place of the original top-K gating strategy [27, 36, 49, 68]. As a systematic approach, MicroMoE is compatible with all these algorithmic solutions.

Many studies offer systematic solutions for load balancing in MoE. FasterMoE [18] models the time consumption of MoE and enables broadcasting the most popular expert to all devices. MoESys [63] dynamically adjusts the number of training nodes to ensure load balance in multi-task scenarios. SmartMoE [64] focuses on optimizing expert placement. FlexMoE [37] adapts the number of replicas of each expert based on their popularity. Several other works [42, 52, 56, 57] embrace similar approaches to FlexMoE, and further reduce the adjustment overhead by integrating expert migration with ZeRO’s communication operations [46]. While these solutions typically balance loads at a per-iteration, expert-level granularity, MicroMoE is capable of achieving optimal load balance at a per-micro-batch granularity.

Janus [31] proposes a data-centric EP paradigm that avoids the inherent load imbalance issue in conventional expert-centric paradigms. However, Janus requires migrating all

expert parameters within every micro-batch, which incurs substantial communication overhead, especially when dealing with large batch numbers and model scales.

9 Discussion

Inference. The load imbalance issue also exists in distributed inference systems with EP [25, 47, 67]. Since inference jobs are more latency-sensitive than training, additional optimizations may be necessary to minimize the scheduling overhead of MicroMoE. For instance, we can replace the linear programming optimization with randomized algorithms or heuristic algorithms for reduced computational complexity, or predict the expert load distributions ahead of the gate network. We plan to employ MicroMoE for inference scenarios in future work.

Scalability. Scaling MicroMoE to large clusters and increased expert counts introduces new challenges to the system. When scaling up, MicroMoE benefits from an expanded scheduling space, which enhances the load balancing capability of MicroEP. However, this larger scheduling space increases computational overhead in scheduling, potentially degrading the system efficiency. To compromise between load balancing capability and system efficiency, we can organize GPUs and experts into groups and perform scheduling at the group level, similar to prior works [9, 26, 64].

FSDP. We currently implement MicroMoE based on Megatron-LM’s DDP [51]. Megatron-LM also supports Fully Sharded Data Parallel (FSDP) [66], which resembles DeepSpeed’s ZeRO-3 [46], sharding model parameters and gradients for memory saving. Moreover, FSDP allows users to selectively enable sharding for model modules, providing a fine-grained control of ZeRO’s optimizations. We plan to integrate MicroMoE with FSDP in future work.

10 Conclusion

We propose MicroEP, a novel expert parallelism strategy to achieve fine-grained load balancing in MoE. MicroEP dynamically balances GPU loads within every micro-batch through token scheduling. We primarily make two optimizations in MicroEP: First, we formulate the token scheduling process as a linear programming problem, which can be solved efficiently. Second, we theoretically analyze the relationship between expert placement and load balancing capacity and develop two placement strategies for different training scenarios. Finally, we propose MicroMoE, an efficient MoE training system based on MicroEP. Our experimental evaluation demonstrates that MicroMoE achieves significant performance improvements, with up to 47.6% end-to-end speedup compared to Megatron-LM, and almost consistently maintains optimal load balance across GPUs.

This work does not raise any ethical issues.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Alie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojgan Javaheripi, Xin Jin, Nikos Karampatzakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambuddha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyra Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL]
- [2] Anthropic. 2025. Introducing Claude 4. <https://www.anthropic.com/news/clause-4>
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS ’20)*. <https://dl.acm.org/doi/abs/10.5555/3495724.3495883>
- [4] Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. 2023. Multi-Modal Gated Mixture of Local-to-Global Experts for Dynamic Image Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV ’23)*. https://openaccess.thecvf.com/content/ICCV2023/html/Cao_Multi-Modal_Gated_Mixture_of_Local-to-Global_Experts_for_Dynamic_Image_Fusion_ICCV_2023_paper.html
- [5] Chang Chen, Min Li, Zhihua Wu, Dianhai Yu, and Chao Yang. 2022. TA-MoE: Topology-Aware Large Scale Mixture-of-Expert Training. In *Advances in Neural Information Processing Systems (NIPS ’22)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 22173–22186. https://proceedings.neurips.cc/paper_files/paper/2022/file/8b465dd58ac50e1b0b22894fd581f62f-Paper-Conference.pdf
- [6] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training Deep Nets with Sublinear Memory Cost. arXiv:1604.06174 [cs.LG]
- [7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi,

- Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: scaling language modeling with pathways. *J. Mach. Learn. Res.* (2023). <https://dl.acm.org/doi/10.5555/3648699.3648939>
- [8] Peizhuang Cong, Aomufei Yuan, Shimao Chen, Yuxuan Tian, Bowen Ye, and Tong Yang. 2024. Prediction Is All MoE Needs: Expert Load Distribution Goes from Fluctuating to Stabilizing. arXiv:2404.16914 [cs.LG]
- [9] DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL]
- [10] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. In *Proceedings of the 39th International Conference on Machine Learning (ICML '22)*. 5547–5569. <https://proceedings.mlr.press/v162/du22c.html>
- [11] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* (2022). <https://dl.acm.org/doi/10.5555/3586589.3586709>
- [12] Wikimedia Foundation. 2025. Wikimedia Downloads. <https://dumps.wikimedia.org/>
- [13] Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. 2023. Megablocks: Efficient sparse training with mixture-of-experts. *Proceedings of Machine Learning and Systems (MLSys '23)* 5 (2023), 288–304.
- [14] Google. 2025. Gemini 2.5: Our most intelligent AI model. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025>
- [15] Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Sicong Leng, Qimei Cui, and Xudong Jiang. 2025. Advancing Expert Specialization for Better MoE. arXiv:2505.22323 [cs.CL]
- [16] Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. 2021. DSelect-k: Differentiable Selection in the Mixture of Experts with Applications to Multi-Task Learning. In *Advances in Neural Information Processing Systems (NIPS '21)*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 29335–29347. https://proceedings.neurips.cc/paper_files/paper/2021/file/f5ac21cd0ef1b88e9848571aeb53551a-Paper.pdf
- [17] Jiaao He, Jiezhang Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. 2021. FastMoE: A Fast Mixture-of-Expert Training System. arXiv:2103.13262 [cs.LG]
- [18] Jiaao He, Jidong Zhai, Tiago Antunes, Haojie Wang, Fuwen Luo, Shangfeng Shi, and Qin Li. 2022. FasterMoE: modeling and optimizing training of large-scale dynamic pre-trained models. In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '22)*. 120–134. doi:10.1145/3503221.3508418
- [19] Zhiyi Hu, Siyuan Shen, Tommaso Bonato, Sylvain Jeaugey, Cedell Alexander, Eric Spada, James Dinan, Jeff Hammond, and Torsten Hoeffer. 2025. Demystifying NCCL: An In-depth Analysis of GPU Communication Protocols and Algorithms. arXiv:2507.04786 [cs.DC]
- [20] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, and zhifeng Chen. 2019. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/093f65e080a295f8076b1c5722a46aa2-Paper.pdf
- [21] Q. Huangfu and J. A. J. Hall. 2018. Parallelizing the dual revised simplex method. *Mathematical Programming Computation* 10, 1 (2018), 119–142. doi:10.1007/s12532-017-0130-5
- [22] Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, HoYuen Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. 2023. Tutel: Adaptive Mixture-of-Experts at Scale. In *Proceedings of Machine Learning and Systems (MLSys '23)*, D. Song, M. Carbin, and T. Chen (Eds.), Vol. 5. Curran, 269–287. https://proceedings.mlsys.org/paper_files/paper/2023/file/5616d34cf8f73942cf5aa922842556-Paper-mlsys2023.pdf
- [23] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lampe, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG]
- [24] Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Reducing activation recomputation in large transformer models. In *Proceedings of Machine Learning and Systems (MLSys '23)*, Vol. 5. 341–353. https://proceedings.mlsys.org/paper_files/paper/2023/hash/80083951326cf5b35e5100260d64ed81-Abstract-mlsys2023.html
- [25] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles* (Koblenz, Germany) (SOSP '23). Association for Computing Machinery, New York, NY, USA, 611–626. doi:10.1145/3600006.3613165
- [26] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. arXiv:2006.16668 [cs.CL]
- [27] Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. 2021. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning (ICML '21)*. PMLR, 6265–6274. <https://proceedings.mlr.press/v139/lewis21a>
- [28] Jiamin Li, Yimin Jiang, Yibo Zhu, Cong Wang, and Hong Xu. 2023. Accelerating Distributed MoE Training and Inference with Lina. In *2023 USENIX Annual Technical Conference (USENIX ATC '23)*. USENIX Association, Boston, MA, 945–959. <https://www.usenix.org/conference/atc23/presentation/li-jiamin>
- [29] Jing Li, Zhiping Sun, Xuan He, Li Zeng, Yi Lin, Entong Li, Binfan Zheng, Rongqian Zhao, and Xin Chen. 2024. LocMoE: A Low-Overhead MoE for Large Language Model Training. arXiv:2401.13920 [cs.LG]
- [30] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. 2020. PyTorch distributed: experiences on

- accelerating data parallel training. *Proc. VLDB Endow.* 13, 12 (2020), 3005–3018. doi:10.14778/3415478.3415530
- [31] Juncai Liu, Jessie Hui Wang, and Yimin Jiang. 2023. Janus: A Unified Distributed Training Framework for Sparse Mixture-of-Experts Models. In *Proceedings of the ACM SIGCOMM 2023 Conference (ACM SIGCOMM '23)*. Association for Computing Machinery, New York, NY, USA, 486–498. doi:10.1145/3603269.3604869
- [32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '21)*. https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html
- [33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.
- [34] Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>
- [35] Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Se-shadri, Nikhil R. Devanur, Gregory R. Ganger, Phillip B. Gibbons, and Matei Zaharia. 2019. PipeDream: generalized pipeline parallelism for DNN training. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles* (Huntsville, Ontario, Canada) (SOSP '19). Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3341301.3359646
- [36] Xiaonan Nie, Qibin Liu, Fangcheng Fu, Shenhan Zhu, Xupeng Miao, Xiaoyang Li, Yang Zhang, Shouda Liu, and Bin Cui. 2024. LSH-MoE: Communication-efficient MoE Training via Locality-Sensitive Hashing. In *Advances in Neural Information Processing Systems (NIPS '24)*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 54161–54182. https://proceedings.neurips.cc/paper_files/paper/2024/file/61674667d642ae52f6bb281bea90ee29-Paper-Conference.pdf
- [37] Xiaonan Nie, Xupeng Miao, Zilong Wang, Zichao Yang, Jilong Xue, Lingxiao Ma, Gang Cao, and Bin Cui. 2023. FlexMoE: Scaling Large-scale Sparse Pre-trained Model Training via Dynamic Device Placement. *Proc. ACM Manag. Data* (2023), 19 pages. doi:10.1145/3588964
- [38] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutscher, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffee, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschle, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makantu, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reichenro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emry Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Witchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotstet, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Jun-tang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [39] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*. <https://arxiv.org/abs/1904.01038>
- [40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. <https://dl.acm.org/doi/10.5555/3454287.3455008>
- [41] Penghui Qi, Xinyi Wan, Guangxing Huang, and Min Lin. 2023. Zero Bubble Pipeline Parallelism. arXiv:2401.10241 [cs.DC]
- [42] Yuhao Qing, Guichao Zhu, Fanxin Li, Lintian Lei, Zekai Sun, Xiuxian Guan, Shixiong Zhao, Xusheng Chen, Dong Huang, Sen Wang, and Heming Cui. 2025. Hecate: Unlocking Efficient Sparse Model Training via Fully Sharded Sparse Data Parallelism. arXiv:2502.02581 [cs.DC]
- [43] Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Demons in the Detail: On Implementing Load Balancing Loss for Training Specialized Mixture-of-Expert Models. arXiv:2501.11873 [cs.LG]

- [44] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multi-task Learners. (2019). <https://openai.com/research/better-language-models/>
- [45] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. 2022. DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale. In *Proceedings of the 39th International Conference on Machine Learning (ICML '22)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.), Vol. 162. PMLR, 18332–18346. <https://proceedings.mlr.press/v162/rajbhandari22a.html>
- [46] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory optimizations Toward Training Trillion Parameter Models. In *International Conference for High Performance Computing, Networking, Storage and Analysis (SC '20)*. 1–16. doi:10.1109/SC41405.2020.00024
- [47] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) (KDD '20). Association for Computing Machinery, New York, NY, USA, 3505–3506. doi:10.1145/3394486.3406703
- [48] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. <https://dl.acm.org/doi/10.5555/3540261.3540918>
- [49] Stephen Roller, Sainbayar Sukhbaatar, arthur szlam, and Jason Weston. 2021. Hash Layers For Large Sparse Models. In *Advances in Neural Information Processing Systems (NIPS '21)*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 17555–17566. https://proceedings.neurips.cc/paper_files/paper/2021/file/92bf5e6240737e0326ea59846a83e076-Paper.pdf
- [50] Noah Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. [arXiv:1701.06538 \[cs.CL\]](https://arxiv.org/abs/1701.06538)
- [51] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. [arXiv:1909.08053 \[cs.CL\]](https://arxiv.org/abs/1909.08053)
- [52] Athinagoras Skiadopoulos, Mark Zhao, Swapnil Gandhi, Thomas Norrie, Shrijeet Mukherjee, and Christos Kozyrakis. 2025. Accelerating Mixture-of-Experts Training with Adaptive Expert Replication. [arXiv:2504.19925 \[cs.DC\]](https://arxiv.org/abs/2504.19925)
- [53] Yehui Tang, Xiaosong Li, Fangcheng Liu, Wei Guo, Hang Zhou, Yaoyuan Wang, Kai Han, Xianzhi Yu, Jinpeng Li, Hui Zang, Fei Mi, Xiaojun Meng, Zhicheng Liu, Hanting Chen, Binfan Zheng, Can Chen, Youliang Yan, Ruiming Tang, Peifeng Qin, Xinghao Chen, Dacheng Tao, and Yunhe Wang. 2025. Pangu Pro MoE: Mixture of Grouped Experts for Efficient Sparsity. [arXiv:2505.21411 \[cs.CL\]](https://arxiv.org/abs/2505.21411)
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. 6000–6010. <https://dl.acm.org/doi/10.5555/3295222.3295349>
- [55] Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. 2024. Auxiliary-Loss-Free Load Balancing Strategy for Mixture-of-Experts. [arXiv:2408.15664 \[cs.LG\]](https://arxiv.org/abs/2408.15664)
- [56] Wei Wang, Zhiqian Lai, Shengwei Li, Weijie Liu, Keshi Ge, Yujie Liu, Ao Shen, and Dongsheng Li. 2023. Prophet: Fine-grained Load Balancing for Parallel Training of Large-scale MoE Models. In *2023 IEEE International Conference on Cluster Computing (CLUSTER '23)*. 82–94. doi:10.1109/CLUSTER52292.2023.00015
- [57] Wei Wang, Zhiqian Lai, Shengwei Li, Weijie Liu, Keshi Ge, Ao Shen, Huayou Su, and Dongsheng Li. 2024. Pro-Prophet: A Systematic Load Balancing Method for Efficient Parallel Training of Large-scale MoE Models. [arXiv:2411.10003 \[cs.DC\]](https://arxiv.org/abs/2411.10003)
- [58] Wikipedia. 2025. Cayley graph. https://en.wikipedia.org/wiki/Cayley_graph
- [59] Shaohua Wu, Jiangang Luo, Xi Chen, Lingjun Li, Xudong Zhao, Tong Yu, Chao Wang, Yue Wang, Fei Wang, Weixu Qiao, Houbo He, Zeru Zhang, Zeyu Sun, Junxiong Mao, and Chong Shen. 2024. Yuan 2.0-M32: Mixture of Experts with Attention Router. [arXiv:2405.17976 \[cs.AI\]](https://arxiv.org/abs/2405.17976)
- [60] xAI. 2025. Grok 4. <https://x.ai/news/grok-4>
- [61] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. 2024. OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models. [arXiv:2402.01739 \[cs.CL\]](https://arxiv.org/abs/2402.01739)
- [62] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. [arXiv:2505.09388 \[cs.CL\]](https://arxiv.org/abs/2505.09388)
- [63] Dianhai Yu, Liang Shen, Hongxiang Hao, Weibao Gong, Huachao Wu, Jiang Bian, Liron Dai, and Haoyi Xiong. 2024. MoESys: A Distributed and Efficient Mixture-of-Experts Training and Inference System for Internet Services. *IEEE Transactions on Services Computing* 17, 5 (2024), 2626–2639. doi:10.1109/TSC.2024.3399654
- [64] Mingshu Zhai, Jiaao He, Zixuan Ma, Zan Zong, Runqing Zhang, and Jidong Zhai. 2023. SmartMoE: Efficiently Training Sparsely-Activated Models through Combining Offline and Online Parallelization. In *2023 USENIX Annual Technical Conference (USENIX ATC '23)*. 961–975. <https://www.usenix.org/conference/atc23/presentation/zhai>
- [65] Shulai Zhang, Ningxin Zheng, Haibin Lin, Ziheng Jiang, Wenlei Bao, Chengquan Jiang, Qi Hou, Weihao Cui, Size Zheng, Li-Wen Chang, Quan Chen, and Xin Liu. 2025. COMET: Fine-grained Computation-communication Overlapping for Mixture-of-Experts. In *Eighth Conference on Machine Learning and Systems (MLSys '25)*. [https://arxiv.org/pdf/2502.19811](https://arxiv.org/pdf/2502.19811.pdf)
- [66] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balooglu, Pritam Damanis, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel. *Proc. VLDB Endow.* 16, 12 (2023), 3848–3860. doi:10.14778/3611540.3611569
- [67] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2025. SGLang: efficient execution of structured language model programs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) (NIPS '24). Curran Associates Inc., Article 2000, 27 pages. <https://dl.acm.org/doi/10.5555/3737916.3739916>
- [68] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, zhifeng Chen, Quoc V Le, and James Laudon. 2022. Mixture-of-Experts with Expert Choice Routing (NIPS '22). In *Advances in Neural Information Processing Systems (NIPS '22)*. 7103–7114. https://proceedings.neurips.cc/paper_files/paper/2022/file/2f00ecd787b432c1d36f3de9800728eb-Paper-Conference.pdf

A Optimizations to MicroEP

A.1 Communication-Aware Scheduling

Many studies demonstrate that the all-to-all communication also takes a significant amount of time in MoE layers [18, 22, 28]. However, the optimization problem 1 only considers the computation time. Therefore, we can further consider the communication time during scheduling.

In many existing frameworks [47, 51], the time of an MoE layer is mainly determined by the communication time + the computation time (unless using some communication-computation overlapping techniques, such as Comet [65] and DualPipe [9]). Therefore, we introduce an additional objective to minimize the maximum communication volume on a GPU. According to Algorithm 1, Line 6, the local data volume of GPU g can be formulated as $local_g = \sum_{e \in E: g \in G_e^e} \min(x_e^g, input_e^g)$. For each GPU, we consider its all-

to-all communication volume as the greater value between its send volume and receive volume. The send/receive volume of GPU g can be calculated from x_e^g , $input_e^g$, and $local_g$. Putting them together, we derive the new optimization problem:

$$\text{minimize } comp + \alpha \cdot comm,$$

$$\begin{aligned} \text{subject to } & comp = \max_{g \in G_{\text{MicroEP}}} \left\{ \sum_{e \in E: g \in G_e^e} x_e^g \right\}, \\ & comm = \max_{g \in G_{\text{MicroEP}}} \{ \max(send_g, recv_g) \}, \\ & send_g = \left(\sum_{e \in E: g \in G_e^e} input_e^g \right) - local_g, \\ & recv_g = \left(\sum_{e \in E: g \in G_e^e} x_e^g \right) - local_g, \\ & local_g = \sum_{e \in E: g \in G_e^e} \min(x_e^g, input_e^g), \\ & \sum_{g \in G_e^e} x_e^g = load_e, \quad \forall e \in E, \\ & x_e^g \geq 0. \end{aligned} \tag{4}$$

α is a constant parameter reflecting the weight of the communication time in the optimization problem. In practice, we can set α to the ratio of expert computation throughput to all-to-all communication throughput. The optimization problem 4 is still an LPP. We omit the derivation of converting problem 4 to a standard LPP format.

Topology-aware scheduling. We can further consider the impact of network topology on communication. In a typical network topology, intra-node communication (e.g., through NVLink) is many times faster than inter-node communication (e.g., through Infiniband). Therefore, we can consider

this difference during scheduling for better communication efficiency.

In the LPP 4, we can split the communication time into intra-node and inter-node. Intra-node communication has a lower weight (α_1) than inter-node communication (α_2). Additionally, we modify the routing mechanism as follows: First, route local tokens to local replicas within the same GPUs. Second, route tokens to replicas on other GPUs in the same nodes. Third, route tokens to other global GPUs.

Overhead. While communication-aware scheduling minimizes communication volume, it increases computational complexity due to additional parameters and constraints in the LPP (comparing LPP 4 and 1). Therefore, we should carefully consider when to enable the communication-aware scheduling, performing a trade-off between the LPP solving time and the communication time.

A.2 Pipelining MicroEP

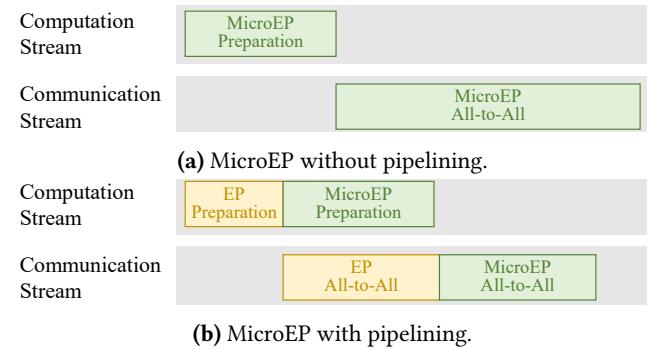


Figure 12. Pipelining MicroEP.

Pipelining offers an additional approach to hide the scheduling overhead of MicroEP, different from the overlapping technique in §5.4. Pipelining is particularly useful in scenarios where it is impossible to fully hide the scheduling time through overlapping. For example, some high-performance frameworks, such as DeepEP [9], have no intermediate operations to overlap between the gate network and all-to-all communication; when dealing with large-scale deployments with numerous GPUs and experts, the scheduling can incur substantial overhead (as illustrated in §7.5). With pipelining, we can overlap the scheduling of some tokens with the all-to-all communication of other tokens. This approach provides more flexibility in latency hiding.

We find many design choices to perform pipelining for MicroEP and illustrate one of them as follows. We observe that MicroEP can achieve perfect load balance in most circumstances (as shown in §7.3). Our intuition is that even if the expert load becomes more imbalanced, MicroEP can still achieve perfect balance. Inspired by this observation, we can split the tokens into two parts: We apply EP to the

former part⁴, and apply MicroEP to the latter part. Moreover, the optimization problem 1 should consider the computational workloads in both EP and MicroEP. In this way, we can overlap the scheduling of the latter part with the all-to-all communication of the former part, as shown in Figure 12.

Overhead. Pipelining MicroEP introduces some additional system overhead. For example, splitting one all-to-all communication operation into two increases synchronization time and GPU kernel launching time. Therefore, we recommend using MicroEP with pipelining in scenarios with substantial scheduling time but minimal system overhead.

B Expert Placement using Cayley Graphs

B.1 Cayley Graphs

In many practical applications, we can construct near-optimal symmetric expert placements in MicroEP using Cayley graphs. The inherent symmetry of Cayley graphs makes them well-suited for constructing optimal expert placements.

A Cayley graph is constructed from a *group A* and its *generating set S*. Each element in the group $a \in A$ is assigned as a vertex. For every $a \in A$ and $s \in S$, there is an edge from the vertex a to the vertex as .

We assume the MicroEP parameter $d = 2$, so the hypergraph is a conventional graph. We observe that in practical applications, the quantities of GPUs and experts are usually powers of two. Let the number of GPUs be 2^p , and the number of experts per GPU be 2^q . Consequently, the number of vertices is 2^p , the degree of each vertex is 2^q , and the number of edges is 2^{p+q-1} . In practice, we heuristically construct many Cayley graphs for different (p, q) s.

B.2 Examples of Cayley Graphs

We illustrate some example constructions as follows.

Example 1: 8 vertices, 8 edges.

We have $p = 3, q = 1$. The group is $(\mathbb{Z}_8, +)$, and the generating set is $\{1, -1\}$. The constructed graph is a cycle.

Example 2: 16 vertices, 32 edges.

We have $p = 4, q = 2$. The group is $(\mathbb{Z}_4 \times \mathbb{Z}_4, +)$, and the generating set is $\{(0, 1), (0, -1), (1, 0), (-1, 0)\}$. This graph is a 4×4 toroidal grid graph, as shown in Figure 13a.

Example 3: 8 vertices, 16 edges.

We have $p = 3, q = 2$. The group is $(\mathbb{Z}_2 \times \mathbb{Z}_4, +)$, and the generating set is $\{(0, 1), (0, -1), (1, 1), (1, -1)\}$. The constructed graph is shown in Figure 13b, which is isomorphic to the complete bipartite graph $K_{4,4}$.

This construction satisfies a good property: $\forall i \in 1, \dots, 8$, the maximum edge counts among all induced subgraphs with exactly i vertices is minimal.

Example 4: 8 vertices, 32 edges.

Note that a complete graph with 8 vertices has $C_8^2 = 28$ edges. Since complete graphs are certainly optimal, we can

⁴Since we have already changed expert placement, this EP is somehow different from typical EP and more like FlexMoE.

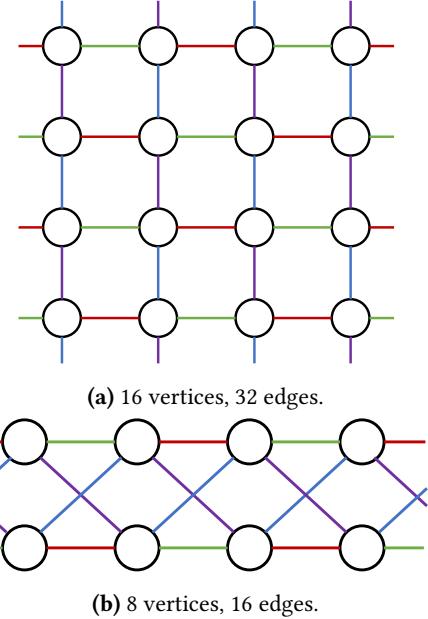


Figure 13. Examples of Cayley graphs.

first generate a complete graph and then add the remaining $32 - 28 = 4$ edges. For the remaining 4 edges, we can simply create an edge between every vertex pair $(0, 1), (2, 3), (4, 5), (6, 7)$ without using Cayley theory.

This method is generalizable to scenarios with more edges: We can first generate multiple complete graphs and then allocate the remaining edges. Note that the number of edges is a power of 2, the number of vertices is 2^p , and the number of edges in a complete graph is $\frac{2^p(2^p-1)}{2}$. Consequently, the number of remaining edges must still be a power of 2, ranging from 2^{p-1} to 2^{2p-2} .

B.3 Synchronization Consistency

Different EDP groups across experts can lead to a consistency issue during parameter and gradient synchronization. Specifically, the synchronization for different experts occurs in different EDP groups, which may incur deadlocks. To prevent deadlocks, we add a consistency restriction in expert placement: All replicas of an expert must have identical local expert indices. For example, in Figure 3c, the replicas of expert 2 are the first local replicas of both GPU 1 and GPU 2; the colors of edges in Figure 13 also indicate the local expert indices. Since DDP executes parameter synchronization following the order of local parameters (and gradient synchronization following the reverse order) [30], deadlocks are effectively avoided.

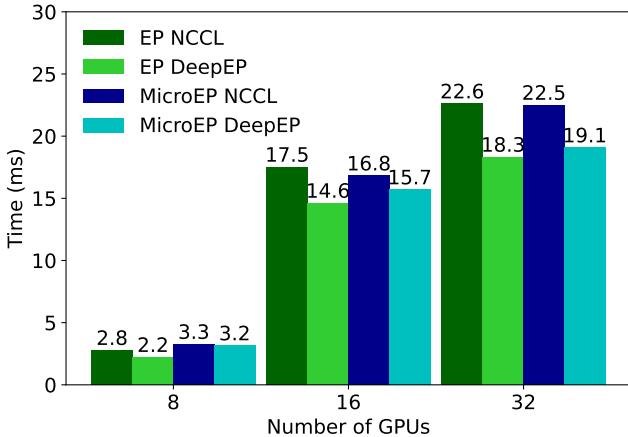


Figure 14. Dispatch time of MicroEP and EP with DeepEP and NCCL, varying number of GPUs.

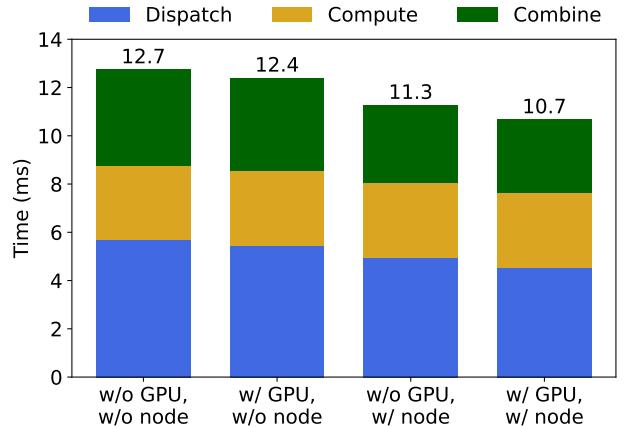


Figure 15. Execution time breakdown of an MoE layer with MicroEP, varying the levels of communication-aware scheduling.

C Supplementary Experiments

C.1 Detailed Experimental Settings

This section provides detailed configurations for our experiments in §7. Table 2 lists the detailed hyperparameters for models used in §7.2.

Activation recomputation is a technique to reduce memory footprint by avoiding recording activations in the forward pass and recomputing them in the backward pass [6]. Furthermore, *selective activation recomputation* enables recomputing a subset of model modules to perform a fine-grained trade-off between computation efficiency and memory [24]. We enable selective recomputation in Megatron-LM to recompute only the MoE FFN, avoiding the out-of-memory (OOM) issue while maintaining relatively high throughput. Since DeepSpeed currently does not support selective recomputation, we recompute the whole layer in DeepSpeed. Furthermore, we find that we can adjust the granularity of selective recomputation at runtime. When the expert loads are highly imbalanced, we can recompute the whole MoE layer for better memory efficiency, avoiding OOM. Otherwise, we can recompute only the MoE FFN for better computation efficiency. For fair comparison, we do not adjust the recomputation granularity during evaluation.

C.2 Evaluation of DeepEP

We evaluate the dispatch time of MicroEP and vanilla EP with DeepEP [9], a high-performance all-to-all communication backend. Megatron-LM [51] currently supports both NCCL [19] (by default) and DeepEP for all-to-all communication. We additionally implement MicroEP with DeepEP.

Before analyzing the experimental results, two important experimental considerations should be noted: (1) Our testbed consists of 8 GPUs and only 2 NICs per node, resulting in limited inter-node network bandwidth. Therefore, we avoid

employing EP or MicroEP across multiple nodes in §7. However, in this section, we focus on the performance of different communication backends, so we expand the communication group into multiple nodes. Consequently, the all-to-all time for inter-node communication is significantly longer than the intra-node communication. (2) In §7, we focus on the system performance, so we compare MicroEP using 8 GPUs per group with EP using 4 GPUs per group ($d = 2$). However, in this section, we focus on the communication performance, so we compare MicroEP and EP using the same group size.

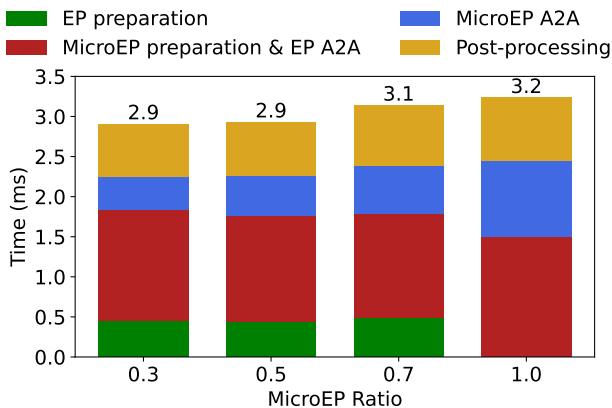
Figure 14 shows the dispatch time comparison between MicroEP and EP using both DeepEP and NCCL, varying number of GPUs. We use the same setting as in §7.4, except for the all-to-all group size. DeepEP exhibits better performance than NCCL due to its high-performance all-to-all implementation. When using NCCL, MicroEP requires less time than EP thanks to the locality-aware routing in §5.2. However, when using DeepEP, MicroEP requires more time than EP due to the data format incompatibilities between DeepEP and Megatron-LM. Consequently, Megatron-LM needs to pre-process the data for DeepEP, while MicroEP incurs a higher pre-processing overhead than EP.

C.3 Evaluation of Communication-Aware Scheduling

We evaluate the performance of the communication-aware scheduling in Appendix A.1. The communication-aware scheduling considers two levels of locality in token dispatching: GPU-level (intra-node) locality and node-level (inter-node) locality. We set $\alpha_1 = 0.1$, $\alpha_2 = 1.0$ as the weights of intra-node and inter-node communication in Problem 4. We use DeepEP as the communication backend due to its superior performance and reduced system overhead compared to NCCL. We believe that the dispatch time of DeepEP provides a more accurate reflection of the communication volume. For other

Table 2. List of Model Hyperparameters

Model	GPT 32×1.3B	GPT 16×3.2B	GPT 8×6.7B	Mixtral 16×2B	Mixtral 8×7B
# layers	24	16	32	32	32
# attention heads	16	32	32	32	32
hidden size	2048	4096	4096	2048	4096
FFN hidden size	8192	16384	16384	8192	14336
sequence length	2048	2048	2048	4096	4096
# experts	32	16	8	16	8
top-K	2	2	2	2	2
micro batch size	4	2	2	2	1
global batch size	512	512	512	256	256
learning rate	1e-5	2e-6	1e-6	1e-5	1e-6
load-balancing loss coeff.	1e-4	1e-4	1e-4	1e-4	5e-4
# GPUs	16	16	32	16	32
PP degree	2	2	4	2	4
EP degree	4	4	4	4	4

**Figure 16.** Dispatch time breakdown with pipelining, vary the ratios of data in MicroEP (1.0 indicates no pipelining).

parameters, we use 16 GPUs, 32 experts, hidden_size=2048, sequence_length=4, micro_batch_size=4. We use randomly generated tokens as input.

We compare the execution time of an MoE layer while enabling/disabling the GPU-level/node-level locality in the communication-aware scheduling. As shown in Figure 15, the overall execution time decreases as we consider more levels of locality during scheduling.

C.4 Evaluation of MicroEP with Pipelining

We evaluate the performance of MicroEP with pipelining in Appendix A.2. We enable the communication-aware scheduling and DeepEP. We use 8 GPUs and 128 experts. Other parameters are the same as Appendix C.3.

We compare the dispatch time with varying ratios of data in MicroEP. Figure 16 demonstrates that pipelining can reduce the dispatch time by overlapping MicroEP preparation

with EP all-to-all communication. However, the dispatch time increases as the MicroEP ratio increases. This is because the EP all-to-all time decreases and becomes insufficient to fully hide the MicroEP scheduling time.